SATNAC 2014

# Southern Africa Telecommunication Networks and Applications Conference (SATNAC) 2014

**31 August - 3 September 2014**
**Boardwalk Conference Centre, Nelson Mandela Bay, Eastern Cape, South Africa**

# 2014 Proceedings

Hosted by Telkom


Ubiquitous Broadband – *An enabler to transform lives*

Venue: Boardwalk Conference Centre, Nelson Mandela Bay, Eastern Cape, South Africa

Date: 31 August to 3 September 2014

**Publication Information**

| | |
|---|---|
| Title: | Southern Africa Telecommunication Networks and Applications Conference (SATNAC) 2014 |
| Proceedings Format: | Printed |
| ISBN: | 978-0-620-61965-3 |
| Format: | CD-ROM |
| Editor: | Roy Volkwyn (Telkom) |
| Date of print: | August 2014 |
| Version: | First Edition |


SATNAC is the flagship of the Telkom Centre of Excellence (CoE) Programme.

Hosted by Telkom

# Table of Contents

# SATNAC 2014 Conference Sponsors

The SATNAC 2014 Committee would like to recognize the following sponsors:

## Diamond Sponsors

## Platinum Sponsors

## Gold Sponsors

## Silver Sponsors

# SATNAC 2014 Review Process

A formal 'Call for Papers' was issued, inviting anyone interested to submit a paper within categories specified by the Organizing Committee. Authors uploaded their papers via web interface onto a database. Papers were assigned to the review panel in the field to judge on the possible acceptance of the submission, based on the scope and depth of the subject matter.

The review process is based on the international de facto standard for blind paper reviews. The review process was undertaken by at least three experienced and well respected individuals. In the blind peer-review process, papers were scrutinized by a panel of South African reviewers, consisting of mainly respected academics, as well as several international experts. The reviewers were asked to provide specific feedback, both positive and negative. This was the only information from the review process disclosed to the authors; all other information was kept confidential.

Reviewers used a 4 point scale to rate the following criteria:

- Originality
- References
- Technical Quality
- Presentation Style

Reviewers gave an overall rating. This was followed by the reviewer comments, which assists the authors in improving and correcting their papers. Reviewers were asked to be as comprehensive as possible.

The reviewers submitted their scoring and comments via web interface onto the database. The Technical Programme Committee drew reports and aggregated the individual scores. The papers were ranked on their average weighted score. The programme dictated the number of papers that could be accepted. Papers were submitted to an online plagiarism database, before being accepted.

The reviewers' comments were forwarded to the author's, with a request to submit a final revised version. Only those papers of high enough quality as recommended by the respective reviewers are included in the SATNAC 2014 Proceedings as Full Reviewed Papers.

Two page Work-In-Progress papers were also invited but were not reviewed as rigorously. Several were accepted for oral presentations, while others for poster presentations. The poster session papers do not form part of the official conference proceedings.

Roy Volkwyn
Chairperson
Technical Programme Committee
SATNAC 2014

# Organising Committee

Alphonzo Samuels (Chairperson)
Gys Booysen
Marti Beukes
Graeme Allan

# Technical Programme Committee

Mr. Roy Volkwyn (TPC Chairperson), Telkom SA
Mr. Gys Booysen, Telkom SA

Mr. Chris Chavaranis, Alcatel-Lucent
Dr. Johann Kellerman, Alcatel-Lucent
Ms. Merryl Ford, CSIR
Mr. Emmanuel Adigun, Deloitte
Dr. I-Sheng Liu, Ericsson
Mr. Nishkar Govender, Ericsson
Ms. Venita Engelbrecht, FibreCo
Prof. Dr. Thomas Magedanz, Frauenhofer Institut
Dr. Anish Kurien, F'SATI/Tshwane University of Technology
Prof. K Djouani, F'SATI/Tshwane University of Technology
Prof TO Olwal, F'SATI/Tshwane University of Technology
Prof. Y Hamam, F'SATI/Tshwane University of Technology
Mr. Abri Rozendaal, IBM
Mr. Emile Swanson, IBM
Mr. Waheed Swales, Investment Solutions Limited
Ms. Phillippa Wilson, Jasco
Dr. Rolan Christian, KPMG
Dr. Paul Plantinga, Monash University
Prof. Andre Calitz, NMMU
Prof. Charmain Cilliers, NMMU
Mr Clayton Burger, NMMU
Prof. Janet Wesson, NMMU
Prof. Jean Greyling, NMMU
Dr. Lester Cowley, NMMU
Mr. Patrick Tchankue Sielinou, NMMU
Ms. Simone Beets, NMMU
Prof. Albert Helberg, North West University
Prof. Alwyn Hoffman, North West University
Mr. Arno de Coning, North West University
Dr. Charl van Heerden, North West University
Gerhard de Klerk, North West University
Prof. Hennie Kruger, North West University
Mr. Henri Marais, North West University
Ms. Leenta Grobler, North West University
Dr. Melvin Ferreira, North West University
Prof. Marelie Davel, North West University
Mr. Samuel van Loggerenberg, North West University
Prof. SE Terblanche, North West University
Dr. Tiny du Toit, North West University
Prof. Willie Venter, North West University
Dr. James Whitehead, Reutech Communications

Prof. Alfredo Terzoli, Rhodes University
Dr. Barry Irwin, Rhodes University
Prof. George Wells, Rhodes University
Prof. G. Foster, Rhodes University
Prof. Hannah Thinyane, Rhodes University
Ms. Ingrid Sieborger, Rhodes University
Mr. James Connan, Rhodes University
Dr. Karen Bradshaw, Rhodes University
Mr. Kevin Duff, Rhodes University
Prof. Mici Halse, Rhodes University
Dr. Mosiuoa Tsietsi, Rhodes University
Prof. Peter Wentworth, Rhodes University
Prof. Philip Machanick, Rhodes University
Prof. Richard Foss, Rhodes University
Mr. Yusuf Motara, Rhodes University
Dr. Richard Good, Smile Communications
Dr. Thomas Niesler, Stellenbosch University
Prof. Tony Krzesinski, Stellenbosch University
Prof. Johan du Preez, Stellenbosch University
Ms. Charna John, Telkom
Mr. David van der Merwe, Telkom
Mr. Eddie Mamahlodi, Telkom
Mr. Edward Lebese, Telkom
Mr. Eric Cartwright, Telkom
Mr. Grant Evert, Telkom
Mr. Ian Durston, Telkom
Dr. Imran Achmed, Telkom
Mr. Jaco Venter, Telkom
Ms. Mariska de Lange, Telkom
Ms. Melanie Delport, Telkom
Ms. Meredith van Rooyen, Telkom
Ms. Mphakiseng Masuabi, Telkom
Mr. Nigel Naidoo, Telkom
Mr. Per Klotzsch, Telkom
Mr. Siphile Sibaya, Telkom
Mr. Sherwin Barlow, Telkom
Mr. Tebogo Modiba, Telkom
Mr. Tennyson Chimbo, Telkom
Mr. Theran Naidoo, Telkom
Mr Zaid Paruk, Telkom
Ms. Zamandlela Ndlela, Telkom
Dr. L. Magagula, Tshwane University of Technology
Mr. Abel Ajibesin, University of Cape Town
Mr. Akinyemi Lateef Adesola, University of Cape Town
Dr. Alexandru Murgu, University of Cape Town
Dr. Boyan Soubachov, University of Cape Town
Mr. Clifford Sibanda, University of Cape Town
Prof. Edwin Blake, University of Cape Town
Mr. Enoruwa Obayiuwana, University of Cape Town
Mr. Henry Ohize, University of Cape Town
Prof. Hussein Suleman, University of Cape Town
Ms. Joyce Mwangama, University of Cape Town

Prof. Mqhele Dlodlo, University of Cape Town
Mr. Neco Ventura, University of Cape Town
Mr. Nicholas Katanekwa, University of Cape Town
Dr. Olabisi Falowo, University of Cape Town
Mr. Periola Ayodele Abiola, University of Cape Town
Mr. Richard Spiers, University of Cape Town
Mr. Samson Orimolade, University of Cape Town
Prof Tania Douglas, University of Cape Town
Dr. Khulumani Sibanda, University of Fort Hare
Prof. Mamelo Thinyane, University of Fort Hare
Mr. Mfundo Shakes Scott, University of Fort Hare
Mr. Sikhumbuzo Ngwenya, University of Fort Hare
Mr. Zelalem Shibeshi, University of Fort Hare
Prof. Andre Nel, University of Johannesburg
Dr. Khmaies Ouahada, University of Johannesburg
Dr. Rodolfo Martinez, University of Johannesburg
Dr. Meera Joseph, University of Johannesburg
Mr. P Robinson, University of Johannesburg
Dr. Suvendi Chinnappen, University of Johannesburg
Mr. Y Roodt, University of Johannesburg
Mr Bashan Naidoo, University of KwaZulu-Natal
Prof. Hongjun Xu, University of KwaZulu-Natal
Prof. Jules Tapamo, University of KwaZulu-Natal
Dr. Narushan Pillay, University of KwaZulu-Natal
Dr. Olutayo O. Oyerinde, University of KwaZulu-Natal
Dr. P.A.Owolawi, University of KwaZulu-Natal
Mr S Rezenom, University of KwaZulu-Natal
Prof. S H Mneney, University of KwaZulu-Natal
Dr. Tahmid Quazi, University of KwaZulu-Natal
Prof. Thomas Afullo, University of KwaZulu-Natal
Prof. Viranjay M. Srivastava, University of KwaZulu-Natal
Mr. Jonas Manamela, University of Limpopo
Prof. Attahiru Sule Alfa, University of Manitoba
Mr Babatunde Awoyemi, University of Pretoria
Mr. Bruno de Carvalho e Silva, University of Pretoria
Mr. Colman Mbuya, University of Pretoria
Mr. Hans Grobler, University of Pretoria
Mr. Jacques van Wyk, University of Pretoria
Prof. Louis Linde, University of Pretoria
Mr. Malcolm Sande, University of Pretoria
Mr. Mike Asiyo, University of Pretoria
Mr. Pieter Jansen van Vuuren, University of Pretoria
Dr. Reza Malekian, University of Pretoria
Mr Roy Fisher, University of Pretoria
Mr. Simon Barnes, University of Pretoria
Ms. Shruti Lall, University of Pretoria
Mr Smart Lubobya, University of Pretoria
Dr. Vivek Dwivedi, University of Pretoria
Mr. Pheeha Machaka, University of South Africa
Mr. Adiel Ismail, University of Western Cape
Prof. Antoine Bagula, University of Western Cape
Mr. Ian Cloete, University of Western Cape

Prof. Isabella M. Venter, University of Western Cape
Mr. Mehrdad Ghaziasgar, University of Western Cape
Mr. Michael J. Norman, University of Western Cape
Mr. Reginald M. Dodds, University of Western Cape
Prof. William D. Tucker, University of Western Cape
Mr. Bethel Mutanga, University of Zululand
Mr. Olukayode A. Oki, University of Zululand
Mr. Paul Tarwireyi, University of Zululand
Mr. Pragasen Mudali, University of Zululand
Mr. Alaba Akingbesote, University of Zululand
Prof. Christo Pienaar, Vaal University of Technology
Mr. Familua Ayokunle Damilola, Witwatersrand University
Prof. Fambirai Takawira, Witwatersrand University
Dr. Jaco Versfeld, Witwatersrand University
Dr. Ling Cheng, Witwatersrand University
Mr. Muhammad Waqar Saeed, Witwatersrand University
Mr. Radu-Ionut Constantinescu, Witwatersrand University
Ms. Reevana Balmahoon, Witwatersrand University
Dr. Renier Dreyer, Witwatersrand University
Prof. Rex van Olst, Witwatersrand University
Mr Stephen Chabalala, Witwatersrand University

# SATNAC 2014 Technical Programme

## 1. Access Networks Technologies

## 2. Converged Services



## 3. Core Network Technologies

## 4. Internet Services & End User Applications

## 5. Limited Range Communications

## 6. Management



## 7. Standards, Regulatory & Environmental

## Work In Progress 1: Access Networks Technologies



## Work In Progress 2: Converged Services

## Work In Progress 3: Core Network Technologies



## Work In Progress 4: Data Centre and Cloud



## Work In Progress 5: Internet Services & End User Applications



## Work In Progress 6: Limited Range Communications



## Work In Progress 7: Management



## Work In Progress 8: Standards, Regulatory & Environmental

# ACCESS NETWORK TECHNOLOGIES

# Optimal Passive Optical Network Planning Under Demand Uncertainty

S.P. van Loggerenberg†, M. Ferreira†, M.J. Grobler†, S.E. Terblanche‡
TeleNet Research Group
†School of Electrical, Electronic and Computer Engineering
‡Centre for Business Mathematics and Informatics
North-West University, Potchefstroom Campus, South Africa
Email: {20289278, melvin.ferreira, leenta.grobler, fanie.terblanche}@nwu.ac.za

***Abstract*—As a result of ever-increasing demand for access level bandwidth, long deployment cycles and the popularisation of more economically viable Point-to-Multipoint (P2MP) networks, service providers are moving to extensively future-proof fibre technologies to connect consumers. Of these, the Passive Optical Network (PON) is the most prevalent. Though the optimal planning of these networks have been studied by a number of authors recently, the typical situation where consumer demand is uncertain has yet to be addressed. By including stochastic demand in an Integer Linear Program (ILP) model through the use of 2-stage stochastic programming, this can be accounted for. In this paper, a discrete approach is followed, optimising the model with the addition of consumer demand scenarios using real-world Geographic Information System (GIS) data. Results show a definitive decrease in deployment cost when any of the scenarios realise, especially when splitter capacity is restrictive.**

***Index Terms*—ILP, Network Modelling, Optimisation, PON, Stochastic Modelling**

## I. Introduction

ACCORDING to [1], international bandwidth demand grew by 39 % during 2012, with Africa's bandwidth demand projected to grow annually at a rate of 51 % [2] between 2012 and 2019. This increase in demand is mostly attributed to the rise in popularity of access-level streaming video and requires extensively future-proof technologies, due in part to inability of service providers to roll out new technologies every year. Though access network technologies exist to fill the gap between current demand and short-term projected demand, such as Very-high-bit-rate Digital Subscriber Line (VDSL) and Long Term Evolution (LTE), they do not have the required bandwidth potential to keep up with long-term consumer bandwidth requirements. Due to recent standard advancements in fibre technology, P2MP fibre networks such as PON have become an economically viable alternative to copper, providing high bandwidth, noise-immune access networks from street level (Fibre to the Curb (FTTC)), right up to customer premises (Fibre to the Home (FTTH)).

The two main standards for PON are ITU-T G.984 Gigabit Passive Optical Network (GPON) [3] and IEEE 802.3ah Ethernet Passive Optical Network (EPON) [4]. EPON builds on the existing Ethernet standard, allowing quick integration with existing network infrastructure while providing bidirectional bandwidth of 1 Gb/s. GPON is based on Asynchronous Transfer Mode (ATM), providing legacy support while increasing up- and downstream bandwidth rates to 1.244 Gb/s and 2.448 Gb/s, respectively. Recently, 10 Gb/s versions of both standards have also been ratified, with G.987 XG-PON [5] and 802.3av 10G-EPON [6].

Planning of PON networks remains a challenge, with sub-optimal manual plans leading to unnecessarily inflated deployment costs, potentially in the order of millions of rands. Even though a number of approaches have been followed to produce optimal or close-to-optimal plans for P2MP fibre access networks [7]–[10] (see [11] for a more in-depth summary), even including fibre duct sharing [12], there still exists a gap between these approaches and real-world deployments: demand uncertainty. When service providers plan greenfield fibre networks, i.e. trenching and laying fibre where none existed before, the demand of consumers in the area is unknown, introducing a large uncertainty into the eventual utilisation of the network and effectively the Return on Investment (ROI). In this paper we will incorporate demand uncertainty into the PON planning model itself to improve solution quality over a number of possible outcomes, which has, to the best of our knowledge, not been done before.

The rest of the paper is organised as follows: Section II defines the PON planning problem in greater detail with respect to the topology and assumptions made. Section III introduces the concept of demand uncertainty and how it is handled in the planning model with section IV defining the resulting mathematical model. In section V, the uncertainty model is compared to a standard planning model before concluding the paper in section VI with a summary and suggestions for future work.

## II. Problem Definition

### A. Topology

PON follows a hierarchical tree topology with the Central Office (CO) as the root node. Fibres run from an Optical Line Terminal (OLT) at a CO to a number of splitters, where the optic signal is passively split into a number of downstream signals. These passive splitters are then each connected by fibre to a number of Optical Network Units (ONUs) at customer premises. Figure 1 illustrates the PON topology. In

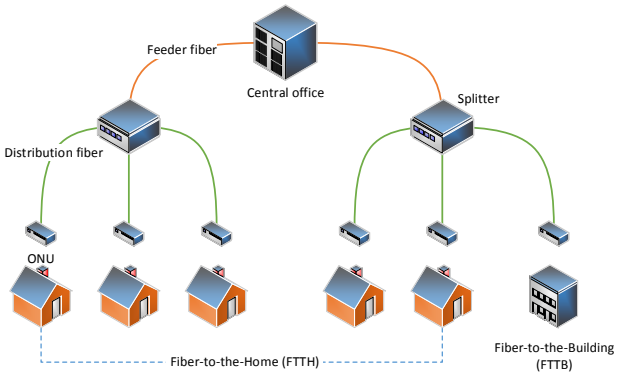



Fig. 1: Basic Passive Optical Network topology

this paper, the network of fibres between COs and splitters is called the *feeder* network, whereas the fibres between splitters and ONUs are collectively known as the *distribution* network.

The greatest advantage of PON is that a single feeder fibre from the CO serves a number of ONUs down the line, greatly reducing overall network cost. The GPON standard specifies the maximum number of ONUs that can be served from a single fibre, also known as the maximum *split ratio*, as 1:128, while EPON supports as many as the link budget allows, usually up to 1:64 [13]. These high split ratios require expensive high power optics to attain a feasible network reach, therefore split ratios of 1:16 and 1:32 are more common.

Given a number of ONU locations at customer premises, a CO location and a number of potential sites for splitters, the PON planning problem then becomes the search for a network that simultaneously solves the following [14]:

- Determining the optimal number of splitters,
- Allocating ONUs to splitters,
- Relocating splitters,
- Routing and aggregating fibre to minimize overall deployment cost.

Unfortunately, the structure of the problem makes solving this model extremely hard, known as Non-deterministic Polynomial time (NP)-hard in complexity theory. In other words, it has been mathematically proven that it is impossible to construct a deterministic algorithm that can solve this problem to optimality in polynomial time, or $O(n^k)$, with a fixed $k$ [14]. This behaviour ensures that anything but the smallest of datasets can not be solved in a feasible time.

### B. Model assumptions

Since the problem is so difficult to solve, a number of assumptions have been made to reduce complexity to a manageable level. Firstly, a few network constraints will be ignored for the model in question, including the placement of splicing boxes as well as the actual power budget constraints. Therefore, for this proof of concept model, it is assumed that individual lossless fibres can be placed in a duct without the need for splicing. The deployment area is also assumed to be greenfield, with no existing ducts available.

To simplify costing, a fixed cost per unit length of trenching and fibre is assumed, as well as a single splitter type. Furthermore, since we want to mainly test the distribution network side as this is where the uncertainty lies, we assume a single CO is available.

It should be noted that even though these factors are omitted in this paper, they can in fact be modelled (as was shown in our work in [15]).

### C. Input data

As input for PON planning model, most authors use randomly generated datasets with fixed distributions. This provides decent performance indication, but can hide potential weaknesses in the planning approach. To avoid this potential bias, we use real-world GIS data of a typical PON network topology. These datasets contain a set of geo-referenced points or *nodes*, defined by latitude and longitude, each connected to subset of neighbouring points with a set of *edges*, each with its own *weight* (in our case, the distance between the nodes).

To determine the distance between two non-connected nodes, a shortest-path algorithm like Dijkstra's algorithm can be used. In [16], the author provides an algorithm based on Dijkstra that outputs the $k$ shortest paths between two nodes on a graph. This algorithm can also give *all* the possible paths between two nodes if it terminates before reaching an arbitrarily large value for $k$, which is the configuration required for the provided models. Note that a heuristic solution can be computed by including only a subset of all the possible paths [12].

## III. Demand Uncertainty

As we mentioned before, for greenfield network deployment, it is quite rare that consumer demand is known at the planning stage. Even though the network could be designed based on the notion that every consumer will use the PON service, it is unrealistic and can lead to unnecessary expenses during deployment. Therefore, consumer demand has to be estimated or projected based on certain known metrics, e.g. demand for previous services or average income per household, to be able to plan a network accordingly.

To model demand uncertainty, we utilise a widely used technique known as 2-stage stochastic programming (see [17, p. 103]). This technique uses data available at the moment the decision is made (first stage) along with some recourse action that estimates the effect when the uncertainty is revealed (second stage). In this paper, we consider a discrete approach based on independent scenarios. In this approach, a scenario is a possible outcome of the consumer demand, containing a subset of the total demand with some given probability. Figure 2 shows two scenarios as an example.

A vital assumption for this approach is that only one scenario realises at the end, allowing the model to over utilise splitter capacity between scenarios for lower deployment cost. Therefore the probabilities of all scenarios realizing should sum to 100 %. For simplicity, in this paper we will assume each scenario has an equal probability to realise.

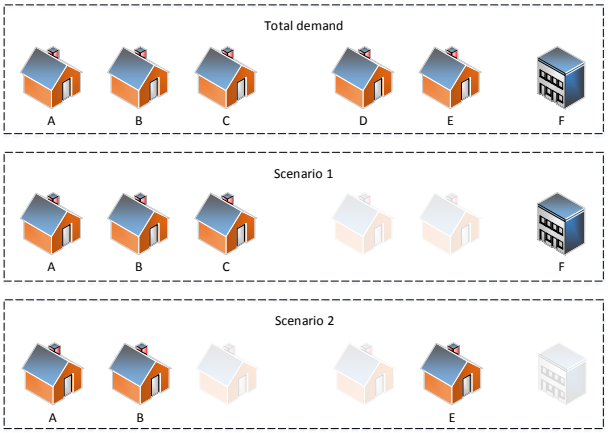



Fig. 2: Example of two scenarios

For the PON planning approach contained in this paper, we assume that the planning occurs in two *phases*. In the first phase, it is assumed that the service provider uses a plan to determine where to trench for fibre ducts. In the second phase, a demand scenario realises and the service provider proceeds to place fibre in the existing trenches where needed, effectively working with a brownfield network. Note that this second phase should not be confused with the recourse stage of the 2-stage stochastic programming approach, as they function externally and internally to the model respectively.

Unfortunately, this approach inevitably leads to increased complexity as we have to effectively enumerate a large number of possibilities for the uncertain demand to realise. This aspect of the problem, scenario generation, is not discussed in detail in this paper, although a number of techniques exist to generate independent scenarios, including using historical data or Monte Carlo methods [18], [19].

## IV. Mathematical Models

Three mathematical models are necessary to illustrate the workings and advantages of a 2-stage stochastic model. We will now discuss each of these below, by first introducing the required notation and then detailing the model formulation.

### A. Sets

- $\mathbf{U}$ - A set of all possible locations for ONUs in the form of Cartesian coordinates. The index *j* is used to indicate the coordinates of the *j*-th ONU.
- $\mathbf{D}$ - The set of all possible locations for splitters in the model. For this set, the index *i* is used to indicate the *i*-th splitter coordinates.
- $\mathbf{S}$ - The set of all scenarios. Index *s* is used for the *s*-th scenario.
- $\mathbf{K}$ - The set of all commodity pairs, i.e. all combinations of the CO and splitters and splitters and ONUs.
- $\mathbf{P}$ - The set of all possible paths between all commodity pairs.
- $\mathbf{E}$ - The set of all edges contained in the input graph.

### B. Subsets

- $\mathbf{K_F}(i) \subseteq \mathbf{K_F} \subset \mathbf{K}$ - Subset of all feeder commodity pairs between the CO and a splitter that contains the element $i$, i.e. all combinations of CO and splitters that contain either CO $i$ or splitter $i$.
- $\mathbf{K_D}(i) \subseteq \mathbf{K_D} \subset \mathbf{K}$ - Subset of all distribution commodity pairs between a splitter and an ONU that contains the element $i$, i.e. all combinations of splitters and ONUs that contain either splitter $i$ or ONU $i$. For example $\mathbf{K_D}(j), j \in \mathbf{U}$, contains all combinations of splitters and the $j$-th ONU.
- $\mathbf{P_D}(k) \subseteq \mathbf{P}$ - A subset containing all possible paths between the distribution commodity pair $k \in \mathbf{K_D}$.
- $\mathbf{P_F}(k) \subseteq \mathbf{P}$ - A subset containing all possible paths between the feeder commodity pair $k \in \mathbf{K_F}$.
- $\mathbf{P}(e) \subseteq \mathbf{P}$ - A subset consisting of all paths that contain the edge $e \in \mathbf{E}$.
- $\mathbf{P}(k, e) \subseteq \mathbf{P}(e)$ - A subset consisting of all paths between commodity pair $k \in \mathbf{K}$ that contain the edge $e \in \mathbf{E}$.

### C. Variables

- $y_{ps}^{\mathrm{D}}$ - Binary variable used to indicate usage of the $p$-th distribution path for scenario $s$, $p \in \mathbf{P}, s \in \mathbf{S}$. The variable takes on a value of 1 if the $p$-th path is used for scenario $s$ and 0 if it's unused.
- $y_p^{\mathrm{F}}$ - Binary variable used to indicate usage of the $p$-th feeder path, $p \in \mathbf{P}$. The variable takes on a value of 1 if the $p$-th path is used and 0 otherwise.
- $\hat{y}_p^{\mathrm{D}}$ - Binary variable used to indicate usage of the $p$-th distribution path across all scenarios, $p \in \mathbf{P}$. The variable takes on a value of 1 if the $p$-th path is used and 0 otherwise.
- $y_p$ - Binary variable used to indicate usage of the $p$-th path, $p \in \mathbf{P}$. The variable takes on a value of 1 if the $p$-th path is used and 0 otherwise.
- $x_e$ - Binary variable used to indicate usage of the $e$-th edge, $e \in \mathbf{E}$. Similarly, the variable takes on a value of 1 if the $e$-th edge is used and 0 if it's unused.
- $\psi_i$ - Binary variable used to define the usage of the $i$-th splitter, $i \in \mathbf{D}$. If the splitter is used, the variable takes on a value of 1. If unused, the variable is 0.

### D. Parameters

- $c_{\mathrm{CO}}$ - The fixed OLT cost incurred for each PON at the CO.
- $c_{\mathrm{SP}}$ - The cost associated with deploying a single splitter.
- $c_{\mathrm{ONU}}$ - The cost to deploy a single ONU in the field.
- $c_{\mathrm{T}}$ - Average cost per meter of trenching.
- $c_{\mathrm{D}}$ - Average cost per meter of distribution fibre.
- $c_{\mathrm{F}}$ - Average cost per meter of feeder fibre.
- $\sigma_{js}$ - Takes on the value of 1 if ONU $j$ is contained in scenario $s \in \mathbf{S}$ and 0 otherwise, $j \in \mathbf{U}$.
- $\kappa$ - The maximum number of ONUs that can connect to a single splitter, i.e. splitter capacity or maximum split ratio.
- $\ell_e, \ell_p$ - The length in meter of the $e$-th edge and $p$-th path respectively, $e \in \mathbf{E}, p \in \mathbf{P}$.

### E. Model Formulation

*1) Deterministic model (DETRM):* Firstly, we present an ILP model for the PON planning problem without taking uncertain demand into account, i.e. a deterministic model (henceforth DETRM). This model is a standard formulation as seen in a number of works [14], [20], with the addition of fiber duct sharing as shown in [12].

$$\min \quad c_{\mathrm{CO}} + \sum_{i \in \mathbf{D}} \psi_i c_{\mathrm{SP}} + |\mathbf{U}| c_{\mathrm{ONU}} + \sum_{p \in \mathbf{P_D}} y_p \ell_p c_{\mathrm{D}} + \sum_{p \in \mathbf{P_F}} y_p \ell_p c_{\mathrm{F}} + \sum_{e \in \mathbf{E}} x_e \ell_e c_{\mathrm{T}} \tag{1}$$

$$\text{s.t.} \quad \sum_{k \in \mathbf{K_D}(j)} \sum_{p \in \mathbf{P_D}(k)} y_p = 1 \quad \forall j \in \mathbf{U} \tag{2}$$

$$\sum_{k \in \mathbf{K_F}(i)} \sum_{p \in \mathbf{P_F}(k)} y_p = \psi_i \quad \forall i \in \mathbf{D} \tag{3}$$

$$\sum_{k \in \mathbf{K}} \sum_{p \in \mathbf{P}(k,e)} y_p \leq |\mathbf{P}(e)| x_e \quad \forall e \in \mathbf{E} \tag{4}$$

$$\sum_{k \in \mathbf{K_D}(i)} \sum_{p \in \mathbf{P_D}(k)} y_p \leq \kappa \psi_i \quad \forall i \in \mathbf{D} \tag{5}$$

The model minimises total deployment cost in (1), consisting of the fixed OLT cost, splitter deployment cost, ONU cost, distribution and feeder fibre cost, and trenching cost. Constraints (2) and (5) specify the distribution side of the PON network, including the fibre demand for each ONU and the maximum number of ONUs that can connect to each splitter. Equation (5) also ensures that the splitter usage variable $\psi_i$ is set if one or more ONUs are connected to a splitter $i$. Constraint (3) ensures that a fibre is present between each used splitter and the CO, while the inequality in (4) ensures that a trench is placed if at least one fibre crosses the edge.

*2) 2-stage stochastic model (2STOCH):* By including uncertain demand through the use of the 2-stage stochastic programming approach, we arrive at the following ILP model (henceforth 2STOCH).

$$\min \quad c_{\mathrm{CO}} + \sum_{i \in \mathbf{D}} \psi_i c_{\mathrm{SP}} + |\mathbf{U}| c_{\mathrm{ONU}} + \sum_{p \in \mathbf{P_D}} \hat{y}_p^{\mathrm{D}} \ell_p c_{\mathrm{D}} + \sum_{p \in \mathbf{P_F}} y_p^{\mathrm{F}} \ell_p c_{\mathrm{F}} + \sum_{e \in \mathbf{E}} x_e \ell_e c_{\mathrm{T}} \tag{6}$$

$$\text{s.t.} \quad \sum_{k \in \mathbf{K_D}(j)} \sum_{p \in \mathbf{P_D}(k)} y_{ps}^{\mathrm{D}} = \sigma_{js} \quad \begin{matrix} \forall j \in \mathbf{U}, \\ \forall s \in \mathbf{S} \end{matrix} \tag{7}$$

$$\sum_{k \in \mathbf{K_F}(i)} \sum_{p \in \mathbf{P_F}(k)} y_p^{\mathrm{F}} = \psi_i \quad \forall i \in \mathbf{D} \tag{8}$$

$$\sum_{k \in \mathbf{K_D}} \sum_{p \in \mathbf{P_D}(k,e)} \hat{y}_p^{\mathrm{D}} + \sum_{k \in \mathbf{K_F}} \sum_{p \in \mathbf{P_F}(k,e)} y_p^{\mathrm{F}} \leq |\mathbf{P}(e)| x_e \quad \forall e \in \mathbf{E} \tag{9}$$

$$\sum_{s \in \mathbf{S}} y_{ps}^{\mathrm{D}} \leq |\mathbf{S}| \hat{y}_p^{\mathrm{D}} \quad \begin{matrix} \forall p \in \mathbf{P_D}(k), \\ \forall k \in \mathbf{K_D} \end{matrix} \tag{10}$$

$$\sum_{k \in \mathbf{K_D}(i)} \sum_{p \in \mathbf{P_D}(k)} y_{ps}^{\mathrm{D}} \leq \kappa \psi_i \quad \begin{matrix} \forall i \in \mathbf{D}, \\ \forall s \in \mathbf{S} \end{matrix} \tag{11}$$

The objective function (6) of the model is the same as that of DETRM, since we also want to minimise the total deployment cost. Constraint (7) is adapted from (2) but now specifies the demand for ONU $j$ in scenario $s$. Equation (8) is identical to (3), while (9) is adapted to include a duct when a fibre from any scenario passes through an edge. The inequality (10) ensures that when a specific path is used in any scenario, only a single fibre is included in the total plan. This is valid since only a single scenario can realise eventually. Finally, (11) ensures a splitter is used when an ONU is allocated to it from within any scenario.

*3) Realisation model (REAL):* If we were to deploy the complete networks as given in the solutions to DETRM and 2STOCH, the network would likely be impractically oversized. This is due to excess capacity reserved for ONUs not contained in the realised scenario. Therefore, to test the practical cost of the plan, we need to be able to determine the minimum cost of the network, given a scenario $s$. This solution represents the second phase of deployment, using ducts that have been deployed according to the solutions of DETRM and 2STOCH in the first phase. At this point, both splitter and fibre placement need to be determined to complete the network. This model will henceforth be called the *realisation model* (REAL).

$\bar{\mathbf{X}}$ donates a subset of the set $\mathbf{X}$, e.g. $\bar{\mathbf{U}}_\mathbf{s}$ is the subset of $\mathbf{U}$ containing only the ONUs from scenario $s \in \mathbf{S}$ and $\bar{\mathbf{P}}_\mathbf{D}$ is the set of distribution paths between splitters and ONUs contained in scenario $s$. Similarly, the subset $\bar{\mathbf{K}}$ contains all commodities present in scenario $s$ and $\bar{\mathbf{E}}$ contains all edges used in paths between commodities in $\bar{\mathbf{K}}$.

$$\min \quad c_{\mathrm{CO}} + \sum_{i \in \mathbf{D}} \psi_i c_{\mathrm{SP}} + |\bar{\mathbf{U}}_\mathbf{s}| c_{\mathrm{ONU}} + \sum_{p \in \bar{\mathbf{P}}_\mathbf{D}} y_p \ell_p c_{\mathrm{D}} + \sum_{p \in \bar{\mathbf{P}}_\mathbf{F}} y_p \ell_p c_{\mathrm{F}} + \sum_{e \in \bar{\mathbf{E}}} x_e \ell_e c_{\mathrm{T}} \tag{12}$$

$$\text{s.t.} \quad \sum_{k \in \bar{\mathbf{K}}_\mathbf{D}(j)} \sum_{p \in \bar{\mathbf{P}}_\mathbf{D}(k)} y_p = 1 \quad \forall j \in \bar{\mathbf{U}}_\mathbf{s} \tag{13}$$

$$\sum_{k \in \bar{\mathbf{K}}_\mathbf{F}(i)} \sum_{p \in \bar{\mathbf{P}}_\mathbf{F}(k)} y_p = \psi_i \quad \forall i \in \mathbf{D} \tag{14}$$

$$\sum_{k \in \bar{\mathbf{K}}} \sum_{p \in \bar{\mathbf{P}}(k,e)} y_p \leq |\bar{\mathbf{P}}(e)| x_e \quad \forall e \in \bar{\mathbf{E}} \tag{15}$$

$$\sum_{k \in \bar{\mathbf{K}}_\mathbf{D}(i)} \sum_{p \in \bar{\mathbf{P}}_\mathbf{D}(k)} y_p \leq \kappa \psi_i \quad \forall i \in \mathbf{D} \tag{16}$$

REAL is identical in structure to DETRM, but uses subsets depending on the scenario that realises. This model therefore effectively solves a complete brownfield network.

## V. Results and Analysis

All models were implemented in C++ using the Concert extensions of IBM ILOG CPLEX and solved on an Intel Core i7 @ 2.67 GHz with 16 GiB memory running Windows.

Demand uncertainty is tested by solving both DETRM and 2STOCH to optimality. Then, REAL is solved for each scenario in $\mathbf{S}$, using the ducts from DETRM and 2STOCH.

To ensure the models can be solved in a feasible time, a very small GIS-mapped dataset known as *MicroNet* is used,

TABLE I: Total deployment cost results with $\kappa = 8$

| Model | Scenario | Objective (R) | Splitters used |
|---|---|---|---|
| **DETRM** | | **178,301.52** | **2** |
| DETRM + REAL | 1 | 100,456.66 | 1 |
| DETRM + REAL | 2 | 146,382.39 | 1 |
| DETRM + REAL | 3 | 72,152.84 | 1 |
| **2STOCH** | | **171,851.24** | **1** |
| 2STOCH + REAL | 1 | 100,456.66 | 1 |
| 2STOCH + REAL | 2 | 146,382.39 | 1 |
| 2STOCH + REAL | 3 | 72,152.84 | 1 |

TABLE II: Total deployment cost results with $\kappa = 4$

| Model | Scenario | Objective (R) | Splitters used |
|---|---|---|---|
| **DETRM** | | **202,565.98** | **3** |
| DETRM + REAL | 1 | 106,935.77 | 2 |
| DETRM + REAL | 2 | 159,882.78 | 2 |
| DETRM + REAL | 3 | 72,152.84 | 1 |
| **2STOCH** | | **179,101.68** | **2** |
| 2STOCH + REAL | 1 | 106,906.94 | 2 |
| 2STOCH + REAL | 2 | 154,433.00 | 2 |
| 2STOCH + REAL | 3 | 72,152.84 | 1 |

containing 10 ONUs, 3 possible splitter locations and 36 edges. Three scenarios were manually generated, containing between 3 and 6 ONUs each. To allow a fair comparison in deployment cost between DETRM and 2STOCH, it was ensured that every ONU is contained in at least one scenario.

Maximum split ratio was set to either 1:8 or 1:4 to test its influence on deployment cost. Even though such low maximum split ratios are rarely seen in practical network sizes, it is necessary to scale the constraint to the dataset size to illustrate the effect. In this case, if a split ratio of 1:16 were allowed, all ONUs could be connected to a single splitter, which will almost never be the case in practice.

The total deployment cost of each test run is given in tables I and II. Given a maximum split ratio of 1:8, DETRM gives a 3.8 % higher objective value than 2STOCH, even though all scenarios give the same deployment cost for both types of models. The ducts used in the solutions to both DETRM and 2STOCH are the same, even though fibres are connected differently, explaining the identical scenario results. DETRM also uses an extra splitter, indicating that extra capacity is reserved in the model, even though it is never used when a single scenario realises.

When the split ratio is more tightly constrained to a maximum of 1:4, the difference between models become more evident, with the total deployment cost of DETRM now 13.1 % higher than 2STOCH. Since the resulting network plan for DETRM and 2STOCH differs quite dramatically at this split ratio, the ducts available to REAL changes as well. This results in a slight increase in deployment cost of 0.02 % for scenario 1 and a substantial 3.5 % for scenario 2 when demand uncertainty is excluded, even though the same number of splitters are used.

Figures 3 and 4 show the resulting PON topology for DETRM and 2STOCH respectively, using a maximum split ratio of 1:4. In these figures, circles represent ONUs, triangles represent used splitters and the square represents the CO. Used

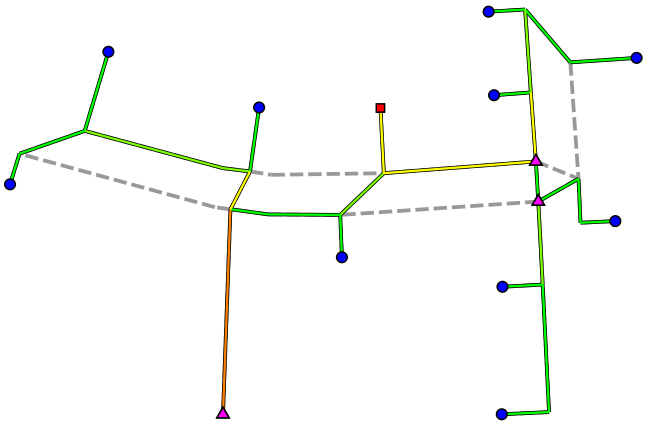


Fig. 3: Optimal solution for DETRM with $\kappa = 4$

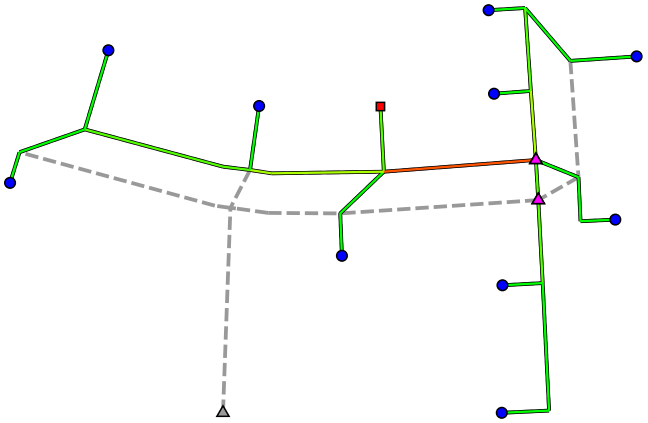


Fig. 4: Optimal solution for 2STOCH with $\kappa = 4$

trenches are coloured green through yellow to red, indicating, in ascending order, the number of fibres contained in each. Grey components indicate unused ducts or splitters. From the figures it is clear that DETRM tries to connect all ONUs without regarding the independent nature of scenarios.

## VI. Conclusion & Future Work

In this paper, demand uncertainty was incorporated in a model of the PON planning problem through the use of a discrete 2-stage stochastic programming approach. This ensures that a finite number of scenarios can be calculated to predict consumer demand, allowing for splitter over-utilisation and optimal duct planning. To the best of our knowledge, this type of formulation has not been previously published, likely due to the increased complexity that is inevitable when modelling stochastic parameters.

Using a small dataset, it was evident that the gains from this approach are very much dependent on the splitter capacity, with tighter restrictions allowing for more savings, up to 13 %. When re-optimising the second phase deployment, gains of up to 4 % were demonstrated. These figures are heavily dependent on the data, with larger datasets expected to show much larger savings due to the increased number of possible fibre paths.

The advantage of this approach stems from the fact that scenarios are independent, ensuring that excess capacity on

splitters are kept to a minimum and ducts are shared as much as possible across all scenarios.

Future work will include algorithmic and decomposition methods to improve computational performance of the modelling approach provided, allowing for much larger datasets to be solved. Furthermore, though scenario generation is discussed in detail in other works, it would be of interest to determine how the process can be tailored for the PON planning problem in particular. Handling of unequal scenario probabilities is also necessary to improve real-world applicability. Finally, a number of refinements can be made to the model to remove some of the more restrictive assumptions, including the use of multiple splitter types, multiple COs and network constraints such as optical power budget.

## ACKNOWLEDGEMENT

The authors thank the anonymous reviewers for their constructive feedback and acknowledge the financial support of the Telkom Centre of Excellence (CoE) at the North-West University, Potchefstroom Campus.

## REFERENCES

[1] TeleGeography. (2013, April) International bandwidth demand is decentralizing. [Online]. Available: http://www.telegeography.com/press/press-releases/2013/04/17/international-bandwidth-demand-is-decentralizing/index.html

[2] TeleGeography. (2013, October) Africa's international bandwidth demand to lead the world. [Online]. Available: http://www.telegeography.com/press/press-releases/2013/10/31/africas-international-bandwidth-demand-to-lead-the-world/index.html

[3] ITU-T, *Gigabit-capable passive optical networks (GPON): General characteristics*, Mar 2008, ITU-T Rec. G.984.1. [Online]. Available: http://www.itu.int/rec/T-REC-G.984.1/en

[4] IEEE, *Standard for Information Technology- Telecommunications and Information Exchange Between Systems- Local and Metropolitan Area Networks- Specific Requirements Part 3: Carrier Sense Multiple Access With Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications Amendment: Media Access Control Parameters, Physical Layers, and Management Parameters for Subscriber Access Networks*, IEEE Std. 802.3ah, 2004.

[5] ITU-T, *10-Gigabit-capable passive optical networks (XG-PON): General requirements*, Jan 2010, ITU-T Rec. G.987.1. [Online]. Available: http://www.itu.int/rec/T-REC-G.987.1/en

[6] IEEE, *Standard for Information technology - Telecommunications and information exchange between systems - Local and metropolitan area networks - Specific requirements Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications Amendment 1: Physical Layer Specifications and Management Parameters for 10 Gb/s Passive Optical Networks*, IEEE Std. 802.3av, 2009.

[7] M. Grötschel, C. Raack, and A. Werner, "Towards optimizing the deployment of optical access networks," ZIB, Takustr.7, 14195 Berlin, Tech. Rep. 13-11, 2013.

[8] M. Leitner, I. Ljubic, M. Sinnl, and A. Werner, "On the two-architecture connected facility location problem," ZIB, Takustr.7, 14195 Berlin, Tech. Rep. 13-29, 2013.

[9] I. Ljubić and S. Gollowitzer, "Layered graph approaches to the hop constrained connected facility location problem," *INFORMS J. on Computing*, vol. 25, no. 2, pp. 256–270, Apr. 2013.

[10] M. Chardy, M.-C. Costa, A. Faye, and M. Trampont, "Optimizing splitter and fiber location in a multilevel optical FTTH network," *European Journal of Operational Research*, vol. 222, no. 3, pp. 430 – 440, 2012.

[11] S. van Loggerenberg, M. Grobler, and S. Terblanche, "Optimization of pon planning for ftth deployment based on coverage," in *Southern African Telecommunications and Networks Access Conference (SATNAC), 2012 Proceedings of*, Sep. 2012.

[12] S. van Loggerenberg, M. Grobler, and S. Terblanche, "Optimization of pon planning for ftth deployment with fiber duct sharing," in *Southern African Telecommunications and Networks Access Conference (SATNAC), 2013 Proceedings of*, Sep. 2013.

[13] F. Effenberger, D. Clearly, O. Haran, G. Kramer, R. D. Li, M. Oron, and T. Pfeiffer, "An introduction to pon technologies [topics in optical communications]," *Communications Magazine, IEEE*, vol. 45, no. 3, pp. S17 –S25, March 2007.

[14] J. Li and G. Shen, "Cost minimization planning for greenfield passive optical networks," *Optical Communications and Networking, IEEE/OSA Journal of*, vol. 1, no. 1, pp. 17 –29, June 2009.

[15] S. van Loggerenberg, "Optimization of passive optical network planning for fiber-to-the-home applications," Master's thesis, North West University Potchefstroom Campus, April 2013.

[16] J. Y. Yen, "Finding the k shortest loopless paths in a network," *Management Science*, vol. 17, no. 11, pp. pp. 712–716, 1971.

[17] J. R. Birge and F. Louveaux, *Introduction to stochastic programming*. Springer, 2011.

[18] W. Römisch, "Scenario generation," *Wiley Encyclopedia of Operations Research and Management Science*, 2011.

[19] M. Kaut and S. W. Wallace, "Evaluation of scenario-generation methods for stochastic programming," *Stochastic Programming E-Print Series*, vol. 14, 2003. [Online]. Available: http://edoc.hu-berlin.de/browsing/speps/

[20] K. Poon, D. Mortimore, and J. Mellis, "Designing optimal ftth and pon networks using new automatic methods," in *Access Technologies, 2006. The 2nd Institution of Engineering and Technology International Conference on*, June 2006, pp. 49 – 52.

**Samuel van Loggerenberg** is a Telkom CoE student, currently pursuing his Ph.D in Computer and Electronic Engineering at the North-West University. He received his M.Eng (cum laude) in 2013, B.Eng in 2011 and a B.Sc in Business Mathematics and Informatics in 2008 from the same institution. His research interests include network optimisation, optical networks and peer to multi-peer video streaming networks.

# Performance Analysis of LDPC-Based IEEE 802.16e with Different Modulation Techniques

Abimbola I. Adedapo[1], Mzyece M. and Noel G.
F'SATI and Department of Electrical Engineering
Tshwane University of Technology, Private Bag X680, Pretoria 0001
Tel: +27 12 3825737, Fax: +27 12 3825688
email: idrisnau@gmail.com[1];{mzyecem, NoelG}@tut.ac.za

**Abstract-In order to achieve the ultimate capacity limit of a communication channel, Shannon demonstrated the use of infinitely long sequence of random codes. However, these codes are impractical as a result of their astronomical number of codewords. The performance of the different rates Low Density Parity Check (LDPC) codes together with the various modulation techniques defined for the IEEE 802.16e standard over the Binary Input Additive White Gaussian Noise (BI-AWGN) channel is investigated in this paper. The Bit error rate (BER) results obtained showed that reliable transmission of information can be achieved in proximity to the capacity of the BI-AWGN channel through the use LDPC codes.**

**Index Terms—LDPC codes, Belief propagation, Mobile WiMAX**

## I. INTRODUCTION

IEEE 802.16 or WiMAX (Worldwide Interoperability for Microwave Access) is a suite of communications standards that provides for wireless transmission of digital information in several ways, ranging from a point-to-point (P2P), to full mobile cellular-kind access. It offers the delivery of last-mile wireless broadband access (WBA) as an alternative to wireline broadband technologies such as Digital Subscriber Line (DSL) and traditional cable. The first version of IEEE 802.16 standard addressed primarily the line-of-sight (LOS) environments, and is designated to operate on 10 – 66 GHz frequency range [1, 2]. This is also referred to as the fixed WiMAX or IEEE 802.16-2004. The enhancement of fixed WiMAX with mobility support evolved into mobile WiMAX or IEEE 802.16e-2005. The mobile WiMAX is based on Scalable Orthogonal Frequency Multiple Access (SOFDMA) for optimum allocation of communication resources both in time and frequency domains. The standard is designed to operate in the 2 – 11 GHz frequency range.

As a result of an increasing demand for an efficient and reliable dissemination of information transmission for high-speed data network, such as the mobile WiMAX, the physical layer of the standard is equipped with several Forward Error Correction (FEC) coding and modulation techniques. The coding schemes include Convolutional Code (CC), Reed-Solomon (RS) codes, Convolutional Turbo Code (CTC), Block Turbo Codes (BTC), and the optional Low-Density Parity-Check (LDPC) codes. In recent years, there has been an accelerating dominance of LDPC codes in applications demanding error control coding. The main motivation for this interest is as a result of the engrossing performances accomplished with this coding technique. These include the near-Shannon capacity approaching capability, fast and practically implementable decoding algorithm. Due to these features, LDPC codes have been incorporated into many emerging wireless communication technologies, such as the IEEE 802.16e-2005, 802.11n [3], and Digital Video Broadcasting by Satellite (DVB-S2) [4].

LDPC codes are linear block codes first instantiated in 1962 by Gallager [5], and subsequently rediscovered by several authors in the late 1990s [6]. Since their rediscovery, a tremendous number of research efforts has been expended on the design, efficient encoding and decoding, construction, structural analysis, and performance evaluations of the codes [7-9]. Generally, a linear block code is uniquely defined either by a generator matrix or a parity-check matrix H. If the code is defined by a parity-check matrix H, such as in LDPC codes, the code is a null space of the parity-check matrix. Assuming a Galois Field GF (q) with q-elements, and q is a power of a prime. A regular q-ary LDPC code is defined by a sparse parity-check matrix over GF (q) in which the column weight and row weight are constant. If there are varying numbers of column and row weights in the parity-check matrix, the q-ary LDPC code is said to be Irregular. It is q-ary quasi-cyclic (QC) LDPC code if the parity-check matrix is an array of sparse circulants over GF (q) of the same size. An LDPC code is defined as a cyclic LDPC if the parity-check matrix is made up of a single sparse circulant or a column of sparse circulants over GF (q) of the same size. For a binary LDPC codes q equals 2 [10].

The design of any of the aforementioned LDPC codes is governed by the following properties: no two rows (or columns) in the H-matrix should have any nonzero element in common [11]. This is to ensure that Tanner graph [12] of the resulting LDPC codes is free of short cycles particularly, cycles of length four, for ease of convergence of the decoding algorithm. Additionally, this is done to ensure that the code has a good minimum distance. A good minimum distance either removes the error floor or pushes it down to a lower error probability. The rest of the paper is organised as follows: some related works are examined in section II. Thereafter, the various LDPC code classes defined in the mobile WiMAX and their iterative decoding based on the belief propagation (BP) are introduced in section III. A discussion on the platform utilised for the implementation is given in the section IV. Section V presents the description of

the various simulation results obtained. Finally, the last section provides a conclusion on the significance of the study.

## II. Related work

Ever since the development of the IEEE 802.16e-2005 standard, a great deal of research effort has been deployed on the performance analysis of the physical layer features. For instance, a simulation of the physical layer was performed in [13] with a focus on the performance of the RS codes together with some modulation techniques. Similarly, the authors in [14] demonstrated the performance of the physical layer with both RS and Convolutional coding schemes. A simulation using OPNET modeller was considered in [15] with a focus on the four differentiated services offered by the standard. In [16], the authors examined the various LDPC codes specifications for the standard, and provided a system platform for their simulation. However, actual implementation was not performed in that paper.

Furthermore, the authors in [17] demonstrated some methods of extending the rates of the IEEE 802.16e LDPC codes to cater for flexible link adaptation. However, system simulation was only performed with Binary Phase Shift Keying (BPSK) modulation over the BI-AWGN channel. Thus, the main contribution of the current paper is to demonstrate the performance of the different rates LDPC code classes for IEEE 802.16e for the three modulation techniques specified in the physical layer profile of the standard. In subsequent discussion, these codes will be referred to as WiMAX LDPC codes.

## III. The WiMAX Low-Density Parity-Check Codes

As mentioned earlier, there are many ways of designing an LDPC in order to obtain coding schemes with flexible encoding and decoding complexities. This section presents the description of the QC WiMAX LDPC codes and their most widely used iterative decoder, the BP algorithm.

### *A. The Design of the LDPC Codes*

There are six different LDPC code classes defined for deployment in the physical layer architecture of the mobile WiMAX (IEEE 802.16e). These codes are characterised with four rates ranging from 1/2 rate to 5/6 rate. Two of these code rates are designated as A and B (that is, 2/3A, 2/3B, and 3/4A, 3/4B). The parity-check matrix H for these codes is generally quasi-cyclic in structure. It consists of 24 columns and $(1-\mathrm{R})\times 24$ rows, where R represents the code rate. The first $\mathrm{R}\times 24$ columns symbolise the systematic information while the remaining $(1-\mathrm{R})\times 24$ columns symbolise the redundancy or parity-check bits. Furthermore, the parity-check matrix can be decomposed into $z\times z$ sub-matrices in which each of the sub-matrixes is either a permuted identity matrix or a zero matrix. The sub-matrices have varying lengths, ranging from 24-by-24 to 96-by-96 with an accretion of four [3]. There are 19 codeword block sizes for each of the six code classes, and 114 in total. Thus, the maximum and minimum block sizes are 2304 and 576 respectively. A generic structure of the parity-check matrix for the half rate class of the WiMAX LDPC codes is illustrated in Figure 1.

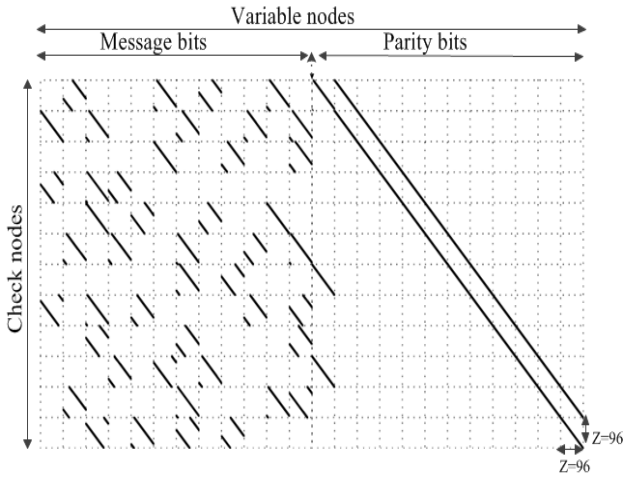



Fig. 1: The generic structure of half-rate WiMAX LDPC code

### *B. Iterative Decoding of LDPC Codes*

Tanner graphs are used to illustrate the decoding operation of an LDPC decoder. These are bipartite graphs with two set of nodes (variable and check nodes), in which undirected edges connect two edges residing in different set. The decoding algorithm for the LDPC codes is generally referred to as BP algorithm or Sum-Product Algorithm (SPA). This is an iterative decoding method wherein extrinsic information is exchanged back and forth among the two nodes that constituted the Tanner graph of the parity-check matrix of the code. Besides the BP algorithm, other decoding algorithms that can be utilised for decoding the WiMAX LDPC codes include the layered decoder and the min-sum algorithm. A comparative study is given in [18] on the various decoding algorithms for the WiMAX LDPC codes. For the purpose of the results obtained in this study, only BP decoding algorithm is considered.

As previously explained, BP decoding is an iterative technique, wherein each round of the algorithm proceeds in two stages. The first half consists of a variable node processing the received information from the channel, and passing the resulting output to the neighbouring check nodes. The other half consists of a check node processing the incoming information and passing the result back to the neighbouring variable nodes. By adopting the same notation as in [19], and assuming all the exchanged information to be Log-Likelihood Ratios (LLRs), the log-domain version of the BP algorithm can be summarised in five steps.

1. Initialisation: for $n=0,\ldots,N-1$, initialise LLR($q_{mn}$) according to the following equation

$$\mathrm{LLR}(q_{mn})=\mathrm{LLR}(c_n)=\frac{2y_n}{\sigma^2}$$

For all m, n for which $h_{mn}=1$, and where $\mathrm{LLR}\left(q_{mn}\right)$ denotes the LLR information sent from the variable node to the check nodes. Also LLR ($c_n$)

denoted the channel LLR information, that is,

$$\mathrm{LLR}(c_n) = \log\left(\frac{\Pr(c_n = 0 \mid y_n)}{\Pr(c_n = 1 \mid y_n)}\right)$$

2. Updating the check node: for $m = 0, \ldots, M-1$ computes the check node LLR ($r_{mn}$) according to the following equation.

$$\mathrm{LLR}(r_{mn}) = \left(\prod_{n' \in N(m) \mid n} \alpha_{m'n}\right) \phi\left(\sum_{n' \in N(m) \mid n} \phi(\beta_{m'n})\right)$$

Where $\alpha_{m'n} = sign(\mathrm{LLR}(q_{m'n}))$, $\beta_{m'n} = \left|\mathrm{LLR}(q_{m'n})\right|$

$$\phi(x) = \log \frac{e^x + 1}{e^x - 1}$$

3. Updating the variable node $LLR(q_n)$ using

$$\mathrm{LLR}(q_{mn}) = \mathrm{LLR}(c_n) + \sum_{m' \in M(n) \mid m} \mathrm{LLR}(r'_{mn})$$

4. Updating the $\mathrm{LLR}(q_n)$ using

$$\mathrm{LLR}(q_n) = \mathrm{LLR}(c_n) + \sum_{m \in M(n)} \mathrm{LLR}(r_{mn})$$

5. Performs hard decision to decode output data by computing the following, for $n = 0, \ldots, N-1$

$$\hat{c}_n = \begin{cases} 1 \rightarrow \text{if } \mathrm{LLR}(q_n) \le 0 \\ 0 \rightarrow \text{otherwise} \end{cases}$$

Finally, the algorithm terminates by computing the decoding syndrome $\hat{c}H^T = 0$ or the number of predefined iteration is reached.

## IV. Simulation Platform

The structure of the mobile WiMAX physical layer employed in the present study is depicted in Figure 2. It is segmented into two major halves, namely, the transmitter and the receiver. At the transmitter end of the model, the source is first randomised through a Pseudo-Random Binary Sequence Generator (PRBSG). It is made up of shift registers and exclusive-or gates. The randomization is performed in order to eradicate long sequence of consecutive ones and zeroes. Subsequently, the randomised output is encoded by each of the LDPC codes starting with the 1/2 rate. An interleaver is usually considered after the encoding operation; however, interleaver is omitted in this implementation as a result of encoding with the LDPC codes. The implementation of an interleaver in this scenario will amount to permuting the columns of the structured parity-check matrix of the code [19]. Similarly, de-interleaver is omitted at the receiving side of the simulation model.

Thereafter, the resulting output of the encoding process is mapped into each of the modulation constellations specified in the physical layer of the IEEE 802.16e standard. These include the Quadrature Phase Shift Keying (QPSK), 16-states, and 64-states Quadratic Amplitude Modulations (16-QAM and 64-QAM). Moreover, the modulated output of the symbol mapper is passed to the Orthogonal Frequency Division Multiplex (OFDM) modulator, which converts the symbols from frequency domain to time domain for ease of transmission across the channel. This conversion is achieved through the application of the Inverse Fast Fourier Transform (IFFT).

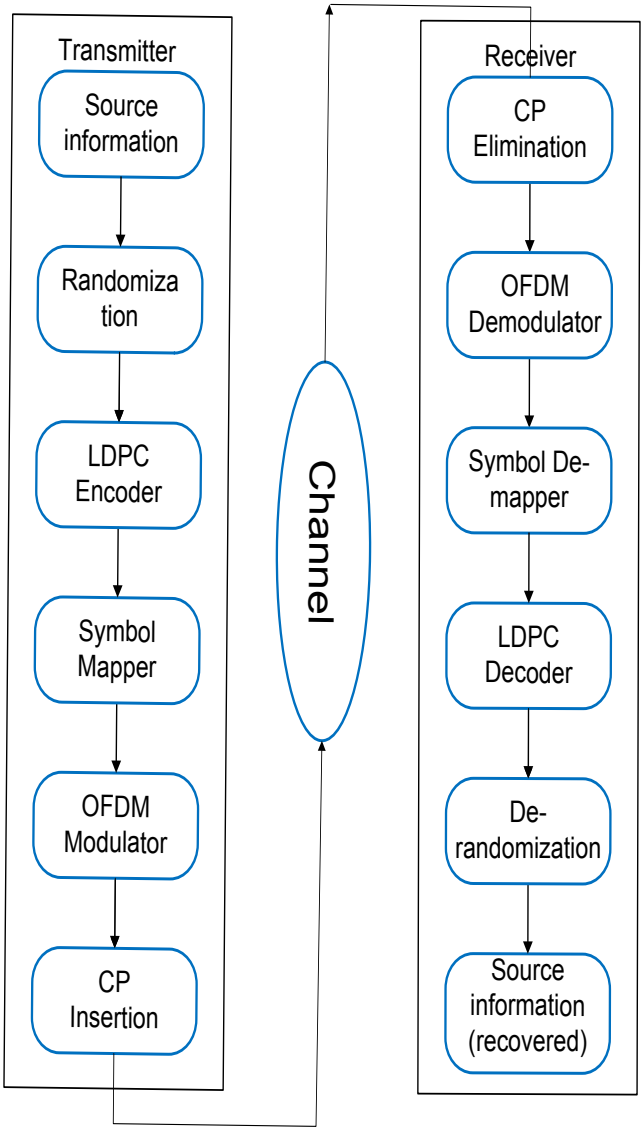



Fig. 2: The Simulation platform for IEEE 802.16e

At the initial stage of the OFDM modulator, the resulting symbol sequence from the mapper is first transformed from serial to parallel prior to the application of an N-point IFFT. Then, a Cyclic Prefix (CP) is appended to the signal stream from the OFDM modulator before being transmitted over the channel. The main goal at receiver side is to be able to recover the source information with a minimum error probability due to the channel imperfection. To achieve this objective, first the channel-corrupted received signal is converted from serial to parallel, followed by the elimination of the CP. Afterwards, an N-point Fast Fourier Transform (FFT) is applied to convert the signal back to the frequency domain by the OFDM demodulator. Subsequently, the signal is demodulated by the symbol de-mapper, and the BP algorithm is applied to the resulting log-likelihood ratios

from the symbol de-mapper. The LDPC decoder employs the LLR information to produce an estimate of the source information. Finally, the performance of the transmission technique is investigated using the BER performance metric.

## V. Discussion and Results

In this section, simulation results for the WiMAX LDPC codes are given with respect to each of the three modulation techniques. The simulation is aimed at analysing the performance of the coding classes for each of the modulation techniques. The LDPC codes considered are characterised with rates 1/2, 2/3A, 3/4A, and 5/6. The other rates (2/3B and 3/4B) were not considered because they yielded similar results compared to the 2/3A and 3/4A. In each cases, each of these code classes were simulated for block-length N = 2304. The maximum block-length specified for an IEEE 802.16e physical layer deployment. Also for each simulation, the maximum numbers of iteration have been chosen as the minimum number of iterations for which near-convergence of the BP algorithm is achieved. This was precisely achieved in 10 iterations for each of the scenarios considered. Table 1 shows the code rates as well as the OFDM parameters utilised for the implementation.

Table 1: System Parameters

| WiMAX LDPC Codes | Modulation Techniques | OFDM Parameters |
|---|---|---|
| R =1/2, LDPC (2304,1152) | QPSK | NFFT = 2048 |
| R =2/3, LDPC (2304,1536) | 16- QAM | N-used = 1536 |
| R =3/4, LDPC (2304,1728) | 64-QAM | CP = 1/4 |
| R =5/6, LDPC (2304,1920) | | N = 8/7 |

For each of the cases considered, a BER versus Energy per bit to noise spectral density ratio ($E_b/N_0$) curve was obtained using MATLAB. The error performances of the four WiMAX LDPC code classes with QPSK modulation are depicted in Figure 3.

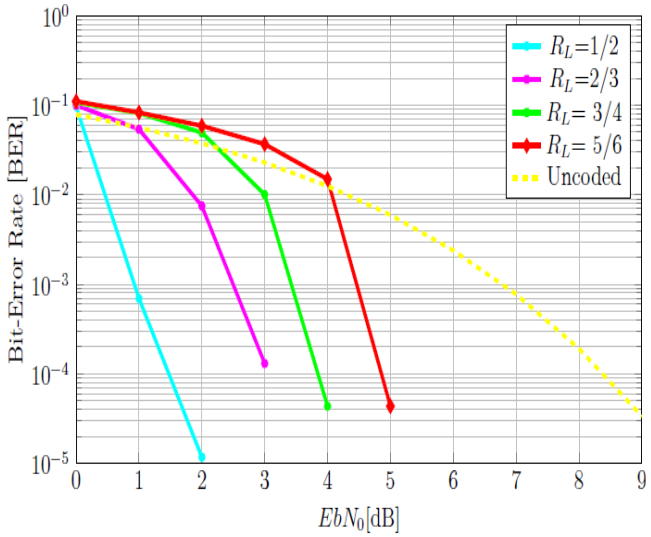



Fig. 3: Error performance of the WiMAX LDPC codes with QPSK modulation.

Moreover, the results obtained while employing 16-QAM and 64-QAM as modulation techniques are depicted in Figures 4 and 5 respectively. Also shown in the graphs are the results for the uncoded IEEE 802.16e system for each of the modulation schemes.

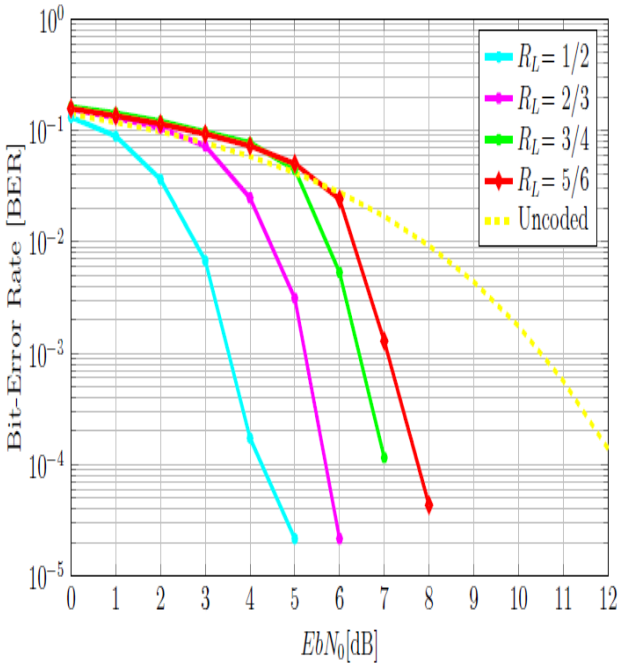



Fig. 4: Error performance of the WiMAX LDPC codes with 16-QAM modulation.

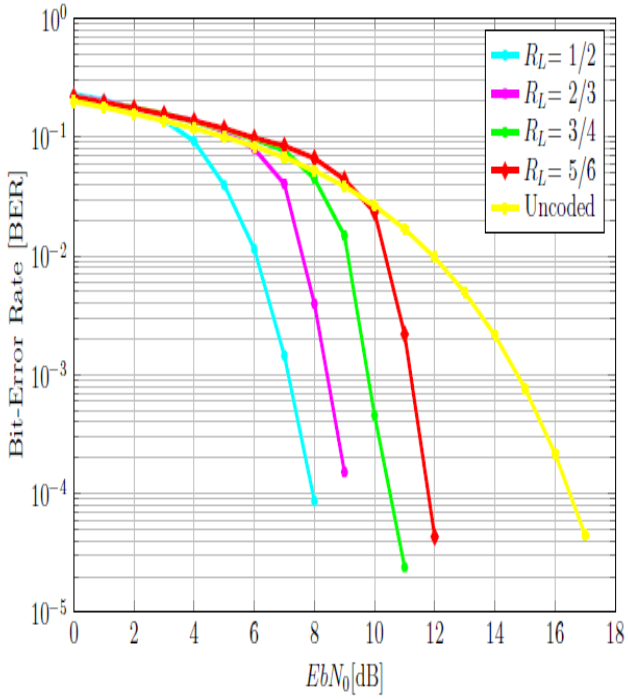



Fig. 5: Error performance of the WiMAX LDPC codes with 64-QAM modulation.

At a BER of $10^{-4}$, the approximate coding gains with respect to the uncoded system for each of the modulation techniques are given in Table 2.

Table 2: Coding gains for the modulation techniques

| Modulation Techniques | Coding gains | | | |
|---|---|---|---|---|
| | R =1/2 | R = 2/3 | R = 2/3 | R = 3/4 |
| QPSK | 6.8 dB | 5.4 dB | 4.5 dB | 3.5 dB |
| 16-QAM | 7.9 dB | 6.5 dB | 5.2 dB | 4.5 dB |
| 64-QAM | 8.5 dB | 7.5 dB | 6.0 dB | 4.6 dB |

It can be deduced from the Table and the graphs that the maximum coding gain can be achieved through the 1/2 rate WiMAX LDPC code category with any of the modulation techniques. This shows that the error performance of the coding scheme varies inversely with the code rate, and the number of bits per modulation constellation. The significance of this development is that the ultimate limit of the communication channel can only be approached at lower code rate and lower modulation. Alternatively, this implies that power efficiency can only be achieved at lower code rate and lower modulation.

## VI. CONCLUSION

Although, it is a challenging task to guarantee a reliable and an efficient transmission of information at high speed as a result of the physical characteristics of the wireless channel. This paper demonstrated that WiMAX LDPC codes can be deployed in this scenario to improve the overall system capacity as well as providing an efficient and a reliable transmission of information across the physical layer of IEEE802.16e. Future study will consider the deployment of these codes in a real mobile WiMAX test bed.

## VII. REFERENCES


[1] W. Fan, A. Ghosh, C. Sankaran, F. Hsieh, and S. Benes, "Mobile WiMAX systems: performance and evolution," *Communications Magazine, IEEE*, vol. 46, pp. 41-46, 2008.

[2] D. Pareit, B. Lannon, I. Moerman, and P. Demeester, "The History of WiMAX: A Complete Survery of the Evolution in Certification and Standardization for IEEE 802.16 and WiMAX," *Communications Surveys & Tutorials, IEEE*, vol. 14, pp. 1183-1211, 2012.

[3] T. Lestable and M. Ran. (2011). *Error control coding for B3G/4G wireless systems: paving the way to IMT-advanced standards*.

[4] A. Morello and V. Mignore, "DVB-S2: The Second Generation Standard for Satellite Broad-Band Service," *Proceedings of the IEEE*, vol. 94, pp. 210-227, 2006.

[5] R. G. Gallager, "Low-Density Parity-Check Codes," *IRE Transaction on Information Theory, IT-8*, pp. 21-28, 1962.

[6] D. J. C. Mackay and R. M. Neal, "Near Shannon limit performance of low density parity check codes," *Electronics Letters*, vol. 32, p. 1645, 1996.

[7] T. J. Richardson and R. L. Urbanke, "Efficient encoding of low-density parity-check codes," *Information Theory, IEEE Transactions on*, vol. 47, pp. 638-656, 2001.

[8] K. Yu, L. Shu, and M. P. C Fosserier, "Low density parity check codes: construction based on finite geometries," in *Global Telecommunucations Conference, 2000. GLOBECOM'00. IEEE*, 2000, pp. 825-829 vol.2.

[9] S.-l. Hwang, L. Hanho, and S. -11. Lim, "A novel method of constructing Quasi-Cyclic RS-LDPC codes for 10GBASE-T Ethernet," in *Circuits and Systems (ISCAS), 2012 IEEE International Symposium on*, 2012, pp. 1771-1774.

[10] H. Qin, D. Qiuju, L. Shu, K. Abdel-Ghaffar "Cyclic and Quasi-Cyclic LDPC Codes on Constrained Parity-Check Matrices and their Trapping sets," *Information Theory, IEEE Transactions on*, vol. 58, pp. 2648-2671, 2012.

[11] L. Shu and D. J. Costello. (2004). *Error Control Coding*.

[12] R. M. Tanner, "A recursive approach to low complexity codes," *Information Theory, IEEE Transactions on*, vol. 27, pp. 533-547, 1981.

[13] M. Patidar, *et al.*, "Performance analysis of WiMAX 802.16e physical layer model," in *Wireless and Optical Communications Networks (WOCN), 2012 Ninth Internation Conference on*, 2012, pp. 1-4.

[14] M. A. Mohamed, F. W. Zaki, and R.H. Mosbeh, "Simulation of WiMAX Physical Layer: IEEE 802.16e," *International Journal of Computer Science and Network Security*, vol. 10, pp. 49-55, 2010.

[15] B. Narendra, G. Anita, and S. Anurag, "Simulation of Physical layer of WiMAX Network using OPNET Modeller," *International Journal of P2P Network Trends and Technology* vol. 3, pp. 258-261, 2013.

[16] S. H. Gupta and B. Virmani, "LDPC for Wi-Fi and WiMAX technologies," in *Emerging Trends in Electronic and Photonic Devices & Systems, 2009. ELECTRO'09. International Conference on*, 2009, pp. 262-265.

[17] Z. Gao, Z. Fei, J. Kuang, and L. Wan, "Rate-Compatible Schemes for Link Adapted LDPC Codes in IEEE 802.16e Standard," in *Vehicular Technology Conference, 2007. VTC2007-Spring. IEEE 65th*, 2007, pp. 3155-3159.

[18] M. Daud, A. B.Suksmono, and S. Hendrawan, "Comparison of decoding algorithms for LDPC codes of IEEE 802.16e standard," in *Telecommunication Systems, Services, and Applications (TSSA), 2011 6th International Conference on*, 2011, pp. 280-283.

[19] W. E. Ryan and S. Lin. (2009). *Channel Codes*.



**Abimbola Idris** received his undergraduate degree in 2005 from the Ladoke Akintola University of Technology and is presently studying towards his Master of Technology degree at the Tshwane University of Technology. His research interests include error control coding, image analysis, and power control.

# Iterative zero-forcing MIMO decoder with symbol sorting

Philip R. Botha and B.T. Maharaj
Department of Electrical, Electronic and Computer Engineering
University of Pretoria
Pretoria, South-Africa
Email: prbotha@ieee.org, sunil.maharaj@up.ac.za

***Abstract*—In this paper the effect of sorting the order in which the Zero Forcing(ZF) decoder performs the decoding of a MIMO system in an iterative Turbo decoder structure is investigated. Two new metrics for sorting based on apriori information is presented and their performance evaluated. The one method sorts according to the absolute value of the Log-Likelihood Ratios (LLR), the second incorporates the norms of the columns of the channel matrix with the LLRs for the symbols. The apriori based metrics are shown to significantly outperform the other metrics. The other methods are the Sorted QR Decomposition (SQRD) and the suboptimal sorting according to the norms of the channel matrix without any apriori information. The decoders were tested on a MIMO threaded algebraic space-time code with quadrature phase shift keying and 16 level quadrature amplitude modulation. The LLR sorted ZF-DF and MMSE-DF decoders are shown to obtain up to an 11dB improvement versus the respective unsorted decoders in the test system. The sorting is shown to have very little effect on non-iterative decoding.**

***Index Terms*—MIMO, Zero forcing, Minimum Mean Squared Error, Decision Feedback, Iterative Decoding, STFC, TAST.**

## I. Introduction

The ever increasing demand for high performance wireless communications has spurred the development of various technologies such as Multiple Input Multiple Output (MIMO) systems. The reliability and capacity of a wireless communication can be significantly improved by the application of MIMO systems [1], [2]. To fully exploit the MIMO channel, various techniques and algorithms have been developed [3], [4]. Linear Pre-coding (LP) is one such technique. LP codes that manage to exploit fully the diversity and capacity available in a MIMO channel have been developed [5]. The mentioned LP codes normally require *joint decoding* which can be accomplished using techniques such as Zero Forcing (ZF), Minimum Mean Squared Error (MMSE) and Sphere Decoder (SD) [6], [7]. Should soft input and output MIMO and Forward Error Correcting (FEC) decoders be used, an iterative turbo structure decoder can be employed to further increase the performance as shown by the authors in [8]. In the iterative turbo decoder, the information obtained by the FEC is passed back to the MIMO decoder and used to improve the result of the MIMO decoder and consequently the information passed on subsequent iterations to the FEC decoder. In [8] the authors proposed a method for performing the ZF decoding with apriori information and compared it to an existing technique proposed in [9]. In this paper, the effect of changing the order in which the symbols are decoded by the ZF-DF and MMSE-DF decoder is investigated. It is known that the order in which the symbols are decoded can have a significant effect on the performance of the ZF decoder [10]. The focus of the paper is to determine what advantages including a priori information in determining the symbol order has on the resultant performance of the iterative decoders. In this paper, the symbols are sorted according to various metrics, such as, absolute symbol LLR, column norms of the channel and symbol LLR, column norms of the channel and by using SQRD.

The results, see section VII, show that including a priori information in the algorithm determining sorting order shows a marked improvement in performance. A difference in the region of 10dB can be seen between the a priori and standard metrics after ten iterations.

*Notation*: In this paper we use the following notation. Vectors are denoted by boldface lowercase letters. Matrices are denoted by boldface uppercase letters. Superscripts $\mathcal{T}$ and $\mathcal{H}$ denote the transpose and Hermitian transpose operations, respectively; $diag(d_1 \dots d_P)$ denotes a $P \times P$ diagonal matrix with diagonal entries $d_1 \dots d_P$. $\mathbf{F_P}$ is the $P \times P$ discrete Fourier transform (DFT) matrix and $(\cdot)$ denotes the dot product of two vectors.

## II. MAP MIMO decoding

The MIMO received signal can generally be described by the following set of linear equations:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}, \tag{1}$$

where $\mathbf{H}$ is the equivalent channel matrix, $\mathbf{x}$ is the transmitted data vector and $\mathbf{n}$ is the noise vector. In general $\mathbf{H}$, $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{n}$ are complex valued. The optimal decoding of the receive signal involves the Maximum Apostiori Probability (MAP) calculation of the value of a bit. This probability expressed in terms of log likelihood ratios (LLR) defined as:

$$\lambda = \ln\left(\frac{P(b=+1)}{P(b=-1)}\right), \tag{2}$$

where $\lambda$ is the LLR and $P(b = x)$ is the probability that $b = x$. BPSK signalling has been assumed for the individual bit values instead of 0 and 1.

Using eq. 2 and Bayes' rule, the MAP solution can be given as:

$$\mathbf{x}_{MAP} = \arg\min_{\mathbf{x}} \{-\ln P(\mathbf{y}|\mathbf{x},\mathbf{H})P(\mathbf{x})\}, \quad (3)$$

where the condition probability $P(\mathbf{y}|\mathbf{x},\mathbf{H})$ is given by:

$$P(\mathbf{y}|\mathbf{x},\mathbf{H}) = \frac{1}{\pi^{N_R}\det(\mathbf{\Sigma})}\exp\left[-(\mathbf{y}-\mathbf{H}\mathbf{x})^{\mathcal{H}}\mathbf{\Sigma}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x})\right], \quad (4)$$

where $N_R$ is the number of receive antennas and $\Sigma$ denotes the covariance matrix of $(\mathbf{y}-\mathbf{H}\mathbf{x})$ given by:

$$\Sigma = E\left[(\mathbf{y}-\mathbf{H}\mathbf{x})(\mathbf{y}-\mathbf{H}\mathbf{x})^{\mathcal{H}}\right]. \quad (5)$$

Since the noise $\mathbf{n}$ is assumed to be AWGN the following simplification may be made:

$$\mathbf{\Sigma} = \sigma_n^2\mathbf{I}_{N_R}, \quad (6)$$
$$\det(\mathbf{\Sigma}) = \sigma_n^{2N_R}, \quad (7)$$
$$\mathbf{\Sigma}^{-1} = \frac{1}{\sigma_n^2}\mathbf{I}_{N_R}, \quad (8)$$

where $\sigma_n^2$ is the noise power and $I_k$ is the $k \times k$ identity matrix. Assuming that the apriori probabilities for each symbol is independent of each other and that the probability of each symbol is the product of the apriori probabilities of each bit in the symbol:

$$P(\mathbf{x}) = \prod_{i=0}^{N_b-1} P(b_i), \quad (9)$$

where $N_b$ is the total number of bits represented by $\mathbf{x}$. The LLR of the apriori information can then be represented as:

$$\lambda_{\mathbf{x}}^a = \sum_{i=0}^{N_b-1} b_i\lambda_i^a \quad (10)$$
$$= \mathbf{b}\cdot\boldsymbol{\lambda}^a \quad (11)$$

Using equations 2 to 11 the MAP solution can be expressed as:

$$\lambda_i^p \approx \arg\min_{\mathbf{x}\in\mathbb{X}_i^{b=-1}}\left\{\frac{||\mathbf{y}-\mathbf{H}\mathbf{x}||^2}{\sigma_n^2} - \frac{1}{2}\mathbf{b}\cdot\boldsymbol{\lambda}^a\right\} - \arg\min_{\mathbf{x}\in\mathbb{X}_i^{b=+1}}\left\{\frac{||\mathbf{y}-\mathbf{H}\mathbf{x}||^2}{\sigma_n^2} - \frac{1}{2}\mathbf{b}\cdot\boldsymbol{\lambda}^a\right\}, \quad (12)$$

where $\lambda_i^p$ is the LLR of the MAP of the $i^{th}$ bit, $b_i$, with $\mathbb{X}_i^{b=c}$ denoting the set of all transmit vectors having $b_i = c$. $\mathbf{b}$ denotes the vector of bits associated with $\mathbf{x}$. BPSK signalling is used for $\mathbf{b}$, and $\boldsymbol{\lambda}^a$ is the vector of apriori LLRs for each corresponding bit in $\mathbf{b}$. Additional use of the max-log approximation:

$$\ln\left(\sum_i e^{x_i}\right) \approx \max_i x_i, \quad (13)$$

was made in the derivation of eq. 12. If the max-log approximation is not used, the decoding problem would be equivalent to an exhaustive search.

## III. ZF MIMO Decoding

Multiplying equation 1 by the inverse of the channel matrix, $\mathbf{H}^{-1}$ yields the ZF solution:

$$\mathbf{H}^{-1}\mathbf{y} = \mathbf{H}^{-1}\mathbf{H}\mathbf{x} + \mathbf{H}^{-1}\mathbf{n} \quad (14)$$
$$= \mathbf{x} + \mathbf{H}^{-1}\mathbf{n}. \quad (15)$$

In the event that $\mathbf{H}$ is non invertible the Moore-Penrose pseudo inverse can also be used. The Moore-Penrose pseudo inverse, denoted by $()^+$, being defined as [11]:

$$\mathbf{H}^+ = (\mathbf{H}^{\mathcal{H}}\mathbf{H})^{-1}\mathbf{H}^{\mathcal{H}}. \quad (16)$$

Alternatively, the QR decomposition of $\mathbf{H}$ may be used to obtain the ZF solution [11]. First the QR decomposition is used to transform eq. 1 into an upper triangular system:

$$\mathbf{Q}\mathbf{R} = \mathbf{H}, \quad (17)$$
$$\therefore \mathbf{Q}^{\mathcal{H}}\mathbf{y} = \mathbf{R}\mathbf{x} + \mathbf{Q}^{\mathcal{H}}\mathbf{n}, \quad (18)$$

where use of the fact that $\mathbf{Q}^{-1} = \mathbf{Q}^{\mathcal{H}}$ has been made. Furthermore, since $\mathbf{Q}$ has orthonormal columns, it is assumed that $\mathbf{Q}^{\mathcal{H}}\mathbf{n}$ is still AWGN.

Since equation 18 has been reduced to an upper triangular system, which is very easy to solve with back substitution, $\mathbf{x}$ can now be decoded on a symbol-by-symbol basis starting with the last symbol.

### A. Classical ZF

Thus, what will be termed the *classical ZF* estimate is given by:

$$\hat{x}_n = \frac{y_n' - \sum_{i=0}^{n-1}\hat{x}_{n-i}R_{n,n-i}}{R_{n,n}}, \quad (19)$$

where $y_n'$ is the $n^{th}$ element in the column vector given by $\mathbf{Q}^{\mathcal{H}}\mathbf{y}$. To obtain the final ZF output, a search for the closest matching transmit symbol is performed:

$$\hat{s}_n = \arg\min_{x_n\in\mathbb{X}} ||\hat{x}_n - x_n||^2, \quad (20)$$

with $\hat{s}_n$ the symbol number corresponding to channel symbol $x_n$ and $\mathbb{X}$ representing the set of all channel symbols, i.e. the signal constellation.

Alternatively, by combining equations 19 and 20 the following may be obtained:

$$\hat{s}_n = \arg\min_{x_n\in\mathbf{X}}\left\{\left|\left|y_n' - \sum_{i=0}^{n-1}\hat{x}_{n-i}R_{n,n-i} - R_{n,n}x_n\right|\right|^2\right\}. \quad (21)$$

### B. ZF with apriori information

Using equation 21 apriori information can be introduced:

$$\hat{s}_n = \arg\min_{x_n\in\mathbf{X}}\left\{\frac{||\hat{y}_n - R_{n,n}x_n||^2}{\sigma_n^2} - \frac{1}{2}\mathbf{b}_i\cdot\boldsymbol{\lambda}_i^a\right\}, \quad (22)$$

where $\hat{y}_n$ is given by:

$$\hat{y}_n = y_n' - \sum_{i=0}^{n-1}\hat{x}_{n-i}R_{n-i,n}, \quad (23)$$

$\mathbf{b}_i$ is the vector of bits corresponding to channel symbol $x_n$ and $\boldsymbol{\lambda}_i^a$ is the vector of apriori LLRs associated with the corresponding bits in $\mathbf{b}_i$. Equation 22 is effectively the MAP solution applied on a symbol-by-symbol bases.

### C. ZF-DF decoder with apriori information

Decision feedback (DF) can be introduced into the ZF decoder by using the channel symbols, $x$, of the previously decoded symbols, $\hat{s}$, in the summations of equations 19, 21, and 23. Specifically equation 23 becomes:

$$\hat{y}_n = y'_n - \sum_{i=0}^{n-1} \Omega(\hat{s}_{n-i}) R_{n-i,n}, \tag{24}$$

where $\Omega(s_n)$ is the function mapping the symbol number $s_n$ to channel symbol $x_n$.

### D. Minimum Mean Squared Error Decoder

The the ZF and ZF-DF decoder can be converted to an MMSE and MSSE-DF decoder by using an augmented channel matrix in the QR decomposition [12]:

$$\mathbf{H}_{aug} = \left[\mathbf{H}^{\mathcal{H}}\mathbf{H} + \frac{\sigma_x^2}{\sigma_n^2}\mathbf{I}_{N_T}\right]^{-1}\mathbf{H}^{\mathcal{H}}, \tag{25}$$

where $\sigma_x^2$ is the received signal power,$\sigma_n^2$ is the noise power and $\mathbf{I}_{N_T}$ is an $N_T \times N_T$ identity matrix. The resultant QR decomposition is based on the *thin QR factorisation* [11]:

$$\mathbf{Q}_1\mathbf{y} = \mathbf{R}_1\mathbf{x} + \mathbf{Q}_1^{\mathcal{H}}\mathbf{n}, \tag{26}$$

where $\mathbf{Q}_1$ is the upper left $N_R \times N_T$ part of $\mathbf{Q}$ and $\mathbf{R}_1$ of $\mathbf{R}$.

## IV. Symbol Sorting

The order in the symbols are decoded by the ZF decoder can have a great impact on the performance of the ZF decoder [10]. Generally the symbol decoding order is in descending order of the signal strength of each symbol. These methods generally do not take the apriori information into consideration. In this section, a few metric for sorting the symbols are investigated.

### A. Sorting with LLR Information

Two methods for calculating a metric based on the apriori information are investigated. The first only takes into consideration the apriori information, the seconds scale the contribution of the LLR with the norm of the appropriate column of $\mathbf{H}$.

*1) LLR Sort:* The LLR Sort metric is based on the sum of the absolute values of the apriori information:

$$C_n^{\lambda} = \sum_{i=0}^{N_b-1} |\lambda_i|, \tag{27}$$

where $\lambda_i$ is the apriori information of bit $i$ in symbol $n$ in LLR form.

*2) LLR & $\mathbf{H}$ Sort:* The combined LLR and $\mathbf{H}$ based sort works on the following metric based on the MAP decoding:

$$C_n^{H\lambda} = \frac{|\mathbf{h}_n|^2}{\sigma_n^2}\frac{1}{2}\sum_{i=0}^{N_b-1} |\lambda_i|, \tag{28}$$

where $\mathbf{h}_n$ is the column of $\mathbf{H}$ corresponding to symbol $n$.

### B. $\mathbf{H}$ Sort

The following metric is solely based on the norms of the columns of $\mathbf{H}$:

$$C_n^{H} = |\mathbf{h}_n|^2. \tag{29}$$

### C. Sorted QRD

The sorted QR decompostion proposed in [10] attempts to order the diagonal elements of $\mathbf{R}$ in increasing orders of magnitude. The algorithm is:

$\mathbf{R} = \mathbf{0}$, $\mathbf{Q} = \mathbf{H}$ & $\mathbf{P} = \mathbf{I}_{N_T}$
**for** $n = 0$ to $N_T - 1$ **do**
  $k_n = \underset{j=n,\cdots,N_T}{\arg\min} |\mathbf{q}_j|^2$
  exchange columns $n$ and $k_n$ in $\mathbf{Q}$, $\mathbf{R}$ and $\mathbf{P}$
  $R_{n,n} = |\mathbf{q}_n|$
  $\mathbf{q}_n = \mathbf{q}_n / R_{n,n}$
  **for** $j = n+1$, to $N_T - 1$ **do**
    $R_{n,j} = \mathbf{q}_n^{\mathcal{H}}\mathbf{q}_j$
    $\mathbf{q}_j = \mathbf{q}_j - R_{n,j}\mathbf{q}_n$
  **end for**
**end for**

where $\mathbf{q}_n$ is the $n^{th}$ column of matrix $\mathbf{Q}$. $\mathbf{P}$ is the permutation matrix by which the columns of $\mathbf{H}$ and the rows of $\mathbf{x}$ have been permuted.

## V. Linear Pre-coding

Linear pre-coding of the symbol constellation encodes the data symbols in a manner that extracts the most benefit from the diversity available in the channel. This is done by mapping a set of data symbols to a new set of encoded symbols that are transmitted. This process can be expressed as a matrix multiplication. Let $\mathbf{x} = [\mathbf{x_1}, \ldots, \mathbf{x_M}]^{\mathcal{T}}$ be a data vector of length $M$ complex symbols from a modulation alphabet (eg. QPSK, M-QAM). Let $\boldsymbol{\Theta}$ be a unitary matrix of size $M \times M$. $\boldsymbol{\Theta}$ may be defined as [13]:

$$\boldsymbol{\Theta} \quad = \quad \mathbf{F_M^{\mathcal{H}}}\mathbf{diag}(\mathbf{1}, \varphi, \ldots, \varphi^{\mathbf{M-1}}), \tag{30}$$

where $\varphi = \exp(j2\pi/4M)$. The linear pre-coding may now be expressed as:

$$\mathbf{s} \quad = \quad \boldsymbol{\Theta}\mathbf{x} \tag{31}$$

where $\mathbf{s}$ is the resultant encoded vector of complex symbols. A linear pre-coded (LP) data vector ($\mathbf{s}$) permits the correct decoding of all of the data symbols ($\mathbf{x}$) encoded by the linear pre-coder upon the reception of a single encoded symbol ($s_i$), thus yielding diversity equal to the rank of the rotation matrix ($R_\Theta$) [13]. Inorder to achieve maximum capacity a number of LP codes can be *layering* two or more LP codes [14]. Using Diophantine numbers aides in the separation of the LP layers at the decoder [14]. The Diophantine number for each layer is given by: $\phi_n = \varphi^n$, $n = 0, \ldots, N_L - 1$.

## VI. System Description

For the purpose of this paper the MIMO channel is modelled as an $N_R \times N_T$ matrix, $\mathbf{H}$, whose elements are each i.i.d complex Gaussian with unit variance and zero mean. This corresponds to the ideal MIMO Rayleigh fading channel, where there is no correlation between the antennas. The receive signal, $\mathbf{y}$, is given by:

$$\mathbf{y} = \mathbf{H}'\mathbf{\Theta}'\mathbf{P}\mathbf{x}' + \mathbf{n}, \tag{32}$$

where $\mathbf{x}$ is the channel symbols from the signal constellation and $\mathbf{P}$ is an arbitrary permutation matric chosen to provide full diversity. All the channel symbols have the same probability of occurring $P(\mathbf{x}) = \frac{1}{M}$, where $M$ is the constellation size. With layering, the matrices $\mathbf{H}, \mathbf{\Theta}$ and $\mathbf{x}$ are given as:

$$\mathbf{H}' = \begin{bmatrix} \mathbf{H_0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{H}_{N_H-1} \end{bmatrix} \tag{33}$$

$$\mathbf{\Theta}' = \begin{bmatrix} \mathbf{\Theta}\varphi^0 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Theta}\varphi^1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{\Theta}\varphi^{N_L-1} \end{bmatrix} \tag{34}$$

$$\mathbf{x}' = \begin{bmatrix} \mathbf{x}_0^{\mathcal{T}} & \mathbf{x}_1^{\mathcal{T}} & \cdots & \mathbf{x}_{N_L-1}^{\mathcal{T}} \end{bmatrix}^{\mathcal{T}}, \tag{35}$$

where $N_L$ is the number of layers and $N_H$ is the total number of orthogonal MIMO transmission made in the time and/or frequency domain. $\mathbf{H}_i$ is the MIMO channel matrix for the particular time and frequency over which the MIMO transmission was made and is i.i.d complex Gaussian. In this paper $N_L = 2$, $N_H = 4$ and the $R_\Theta$ = 4. The received vector $\mathbf{y}$ is then passed to the MIMO decoder, together with $\mathbf{H}'\mathbf{\Theta}'\mathbf{P}$ as can be seen in figure 1. The decoder uses this data to soft decode the received vector into a bitwise vector $\boldsymbol{\lambda}^p_{MIMO}$ the values of which are hard limited to $\pm 5$.

The information vector $\boldsymbol{\lambda}^p_{MIMO}$ is then passed to a binary LDPC decoder with $n = 82$ and $k = 273$ for error correction to yield $\boldsymbol{\lambda}^p_{LDPC}$.

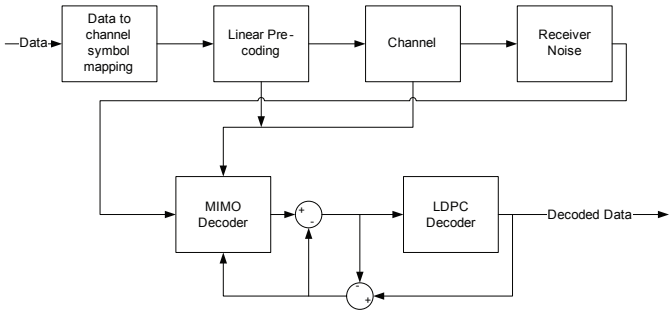



Fig. 1. Block diagram of the system.

The information vector $\boldsymbol{\lambda}^p_{MIMO}$ is then used as the apriori information in the MIMO decoder. This is shown as the feedback path between the LDPC decoder and the MIMO decoder in figure 1. The input to the MIMO decoder is given by:

$$\boldsymbol{\lambda}^a_{MIMO} = \boldsymbol{\lambda}^p_{LDPC} - \boldsymbol{\lambda}^a_{LDPC}. \tag{36}$$

The same is done at the LDPC decoder. The input to the LDPC decoder is given as:

$$\boldsymbol{\lambda}^a_{LDPC} = \boldsymbol{\lambda}^p_{MIMO} - \boldsymbol{\lambda}^a_{MIMO}. \tag{37}$$

## VII. Results

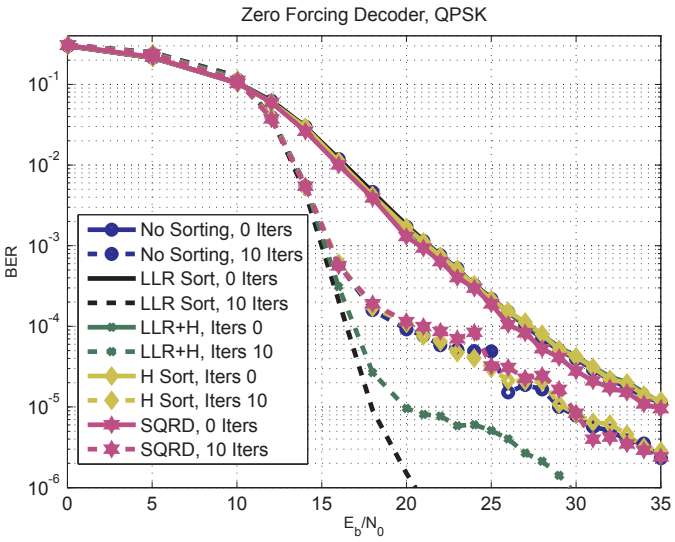



Fig. 2. Plot showing the relative performance of various sorting techniques in Iterative Turbo Zero Forcing Decision Feedback Decoding.

In figure 2 the BER performance of the ZF-DF decoder is shown. At $10^{-4}$ the iterative decoder provides between 7.5dB and 8dB gain. Figure 2 also shows that the gain reduces to approximately 5dB for the metrics that do not include apriori information. The metrics solely based on apriori information (LLR Sort) result in an additional gain of approximately 12.5dB at $10^{-5}$. It can be noted that the metric that incorporates the norms of the channel matrix (LLR+H Sort) does not provide better performance than the metric solely based on apriori information.

Figure 3 shows the performance of the various metric when decoding a 16-QAM constellation. It is evident that at $10^{-4}$ the iterative decoder provides a gain of approximately 5dB with the metrics that do not include apriori information. Including apriori information provides an additional gain of approximately 7.5dB. Once more the best performing metric is the metric exclusively based on the apriori information.

Figure 4 shows the performance of the various metric when decoding a QPSK constellation with the MMSE-DF decoder. It can be seen that LLR sorting provides approximately 12dB improvement over the other schemes, as measured at the $10^{-5}$ BER point. A small gain of approximately 2dB is observed for the un iterated SRQD decoder.

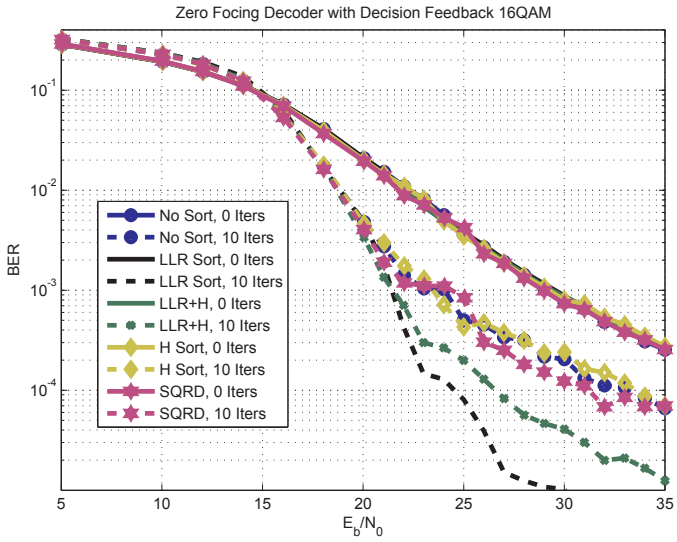



Fig. 3. Plot showing the relative performance of various sorting techniques in Iterative Turbo Zero Forcing Decision Feedback Decoding 16-QAM.

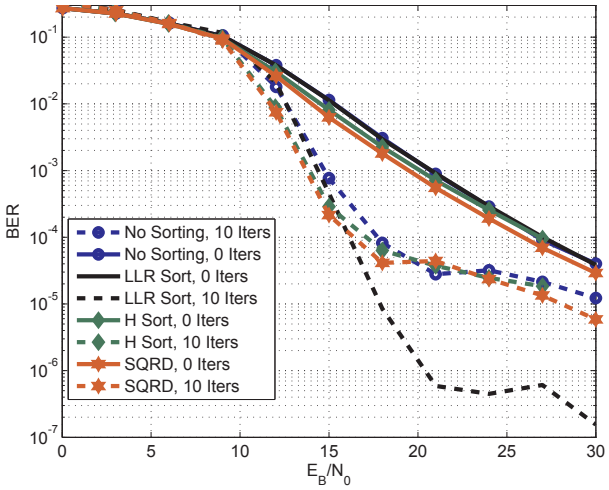



Fig. 4. Plot showing the relative performance of various sorting techniques in Iterative Turbo MMSE-DF Decoding of QPSK.

## VIII. Conclusion

In this paper the authors present various metrics by which to sort the order in which the symbols are decoded by the iterative ZF-DF and MMSE-DF decoders. The best performing metric is the metric based on the absolute LLR value of a symbol. It was shown that the ZF-DF and MMSE-DF decoder benefits significantly from sorting the symbol decoding order according to the apriori information. A gain of up to 10dB was observed for both QPSK and 16-QAM symbol constellations for the apriori information sorted decoder (LLR Sort) vs the unsorted decoder.

Though not shown here, the iterative ZF and MMSE decoders did not perform as well, gaining approximately 2.5dB over the uniterated decoders. The ZF and MMSE decoders also did not benefit significantly from sorting order.

## Acknowledgment

This research was supported by the Sentech Chair in Broadband Wireless Multimedia Communication at the University of Pretoria and the National Research Foundation of South Africa.

## References

[1] İ. E. Telatar, "Capacity of multi-antenna gaussian channels," *European Transactions on Telecommunications*, vol. 10, pp. 585–595, 1999. [Online]. Available: lthiwww.epfl.ch/~leveque/Projects/telatar.pdf

[2] T. Marzetta and B. Hochwald, "Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading," *Information Theory, IEEE Transactions on*, vol. 45, no. 1, pp. 139 –157, Jan 1999.

[3] H. Gamal and J. Harnmons, A.R., "On the design and performance of algebraic space-time codes for BPSK and QPSK modulation," *Communications, IEEE Transactions on*, vol. 50, no. 6, pp. 907 –913, Jun 2002.

[4] V. Tarokh, H. Jafarkhani, and A. Calderbank, "Space-time block codes from orthogonal designs," *Information Theory, IEEE Transactions on*, vol. 45, no. 5, pp. 1456 –1467, Jul 1999.

[5] W. Zhang, X. G. Xia, and P. C. Ching, "High-Rate Full-Diversity Space-Time-Frequency Codes for Broadband MIMO Block-Fading Channels," *IEEE Trans. Commun.*, vol. 55, no. 1, pp. 25–34, 2007.

[6] J. Yoo, J. Lee, and P. Sin-Chong, "Performance evaluation of various MIMO decoders for IEEE 802.11n WLAN system," in *Communication Technology, 2006. ICCT '06. International Conference on*, Nov. 2006, pp. 1 –3.

[7] Z. Safar, W. Su, and K. J. R. Liu, "A fast sphere decoding algorithm for space-frequency block codes," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 148–148, Jan 2006. [Online]. Available: http://dx.doi.org/10.1155/ASP/2006/97676

[8] P. Botha and B. Maharaj, "Turbo STFC decoding with the Zero Forcing decoder," in *AFRICON, 2011*, Sept. 2011, pp. 1 –5.

[9] R. Wang and G. Giannakis, "Approaching MIMO channel capacity with soft detection based on hard sphere decoding," *Communications, IEEE Transactions on*, vol. 54, no. 4, pp. 587 – 590, April 2006.

[10] D. Wubben, R. Bohnke, J. Rinas, V. Kuhn, and K. Kammeyer, "Efficient algorithm for decoding layered space-time codes," *Electronics Letters*, vol. 37, no. 22, pp. 1348 –1350, Oct 2001.

[11] G. H. Golub and C. F. van Loan, *Matrix Computations*, 3rd ed. John Hopkins University Press, 1996.

[12] D. Wubben, R. Bohnke, V. Ku?hn, and K. D. Kammeyer, "MMSE extension of V-BLAST based on sorted QR decomposition," in *Vehicular Technology Conference, 2003. VTC 2003-Fall. 2003 IEEE 58th*, vol. 1, 2003, pp. 508–512 Vol.1.

[13] Z. Liu, Y. Xin, and G. B. Giannakis, "Linear constellation precoding for OFDM with maximum multipath diversity and coding gains," *IEEE Trans. Commun.*, vol. 51, no. 3, pp. 416–427, 2003.

[14] H. El Gamal and M. Damen, "Universal space-time coding," *Information Theory, IEEE Transactions on*, vol. 49, no. 5, pp. 1097 – 1119, May 2003.

# Experimental Demonstration of Raman Amplification in Vertical Cavity Surface Emitting Lasers for Extended Reach Access Networks

E. K. Rotich Kipnoo, T. V. Chabata, R. R. G. Gamatham, A. W. R. Leitch and T. B. Gibbon
Physics Department, Nelson Mandela Metropolitan University, P. O. Box 77000, Port Elizabeth 6031, South Africa. Tel: +27 41-504-2141; Fax: +27 41-504-2573.
Email: Enoch.Rotich@live.nmmu.ac.za

**Abstract- we report on the use of distributed Raman amplifiers employing different pumping schemes for long access networks extending beyond 75 km fibre links using vertical cavity surface emitting lasers (VCSELs) as transmission sources. High on-off gains of 15 dB, 8 dB and 5 dB in the bi-directional, forward and backward pumping schemes respectively enabling an increase in the reach by up to 50 km, 27 km and 17 km respectively were obtained. The bi-directional scheme yielded error free transmission on a 4.25 Gbps data over 100 km G.655 fibre.**

**Keywords:** VCSEL, Stimulated Raman scattering, Amplification, Access network.

## I. INTRODUCTION

Vertical cavity surface emitting lasers (VCSELs) have dominated the access networks in the recent past due to their multiple merits over other sources. There has been an exponential growth of FTTX (X-home, building, node, premise, business, office, etc.) majorly in European, Asian and American countries [1] recently due to constant progress in the development of components and devices suitable for such application scenarios. The Fibre to the Hut concept for African perspective is currently taking shape and considers FTTX technologies suited for our continent [2]. Low power consumption is one outstanding feature of the VCSEL that make it suitable for relatively short distances preferably FTTX. Another important feature of VCSELs is the development of the device at 1550 nm window, for example, 40 Gbps 1550 nm VCSEL has been reported [3]. This has attracted attention not only due to low attenuation around 1550 nm region but also its potential in the wavelength division multiplexing (WDM) application. This has seen VCSELs thrive in WDM - passive optical networks (PONs) [4]. The need for upstream and downstream at tailored wavelengths in WDM-PONs qualify VCSELs for the application. This is due to relatively low cost hence affordability by many users thus the bi-directional transmission is fostered. Another advantage of the 1550 nm window is its ability to utilize the various amplification mechanisms; mainly Erbium Doped fibre amplification, semiconductor amplifiers (SOA) and Raman amplification [5, 6]. Distributed Raman amplifiers for WDM-PONs have been reported. For instance, demonstration of 2.5 Gbps linking remote areas over 60 km reach [7], 10 Gbps transmission was achieved on 80 km [8] and 120 km [9] using an externally modulated laser. The need to increase the reach demands for amplification so as to compensate the attenuation losses and cumulative dispersion penalties in long reach access networks.

In this study, the use of distributed Raman amplifiers in increasing the VCSEL transmission reach is experimentally demonstrated. We first present the theory of Raman amplification then research findings before conclusion.

## II. THEORY

Raman amplification uses stimulated Raman scattering (SRS) to achieve the power transfer from a pump to a signal. SRS is a nonlinear process [10] that occur into the fibre as shown by the quantum representation in Fig. 1.

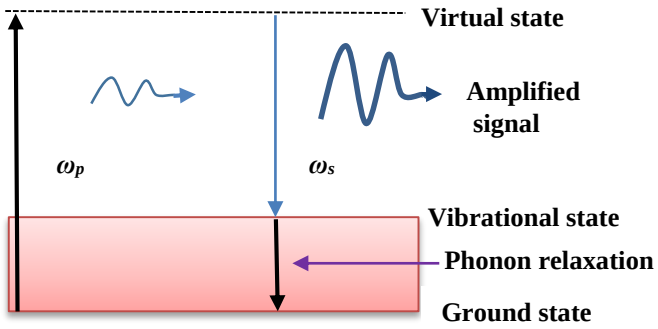



Figure 1: Illustration of stimulated Raman scattering

When a high power pump of wavelength, $\omega_p$ interacts with a signal of wavelength, $\omega_s$ in a fibre, a pump photon is converted to a signal photon replica of the first as the residual energy forms an optical phonon [10].

The pump power, Pp and signal power, Ps propagating in z direction of the fibre can be expressed mathematically [11] as

$$\frac{dP_s}{dz} = -\alpha_s P_s + \left(\frac{g_R}{A_{eff}}\right) P_p P_s \quad \text{..... (1)}$$

Stimulated Raman scattering is polarization dependent therefore polarization dependency can be expressed as,

$$\xi \frac{dP_p}{dz} = -\alpha_p P_p - \left(\frac{\omega_p}{\omega_s}\right)\left(\frac{g_R}{A_{eff}}\right) P_s P_p \quad \text{.... (2)}$$

Where $g_R$ is the gain, $P_p$ and $P_s$ are the pump and signal powers, $\omega_p$ and $\omega_s$ are pump and signal frequencies respectively. $\xi$ indicates the polarization dependency and

$A_{eff}$ is the effective area. These equations illustrate that the pump power provides the energy for amplification and the pump is depleted as the signal power increases. Attenuation in fibres reduce the power with increase in the fibre length. With amplification, the signal power transmitted for forward pumping can be expressed as

$$P_s(L) = P_s(0)\exp\left(\frac{g_R P_0 L_{eff}}{A_{eff}} - \alpha_s L\right) \quad .. (3)$$

Raman amplification has different pumping schemes [11]; co-pumping also referred to as forward pumping which the pump and the signal are coupled in the same direction in the fibre. The second is counter pumping which the pump is in opposite (backward) direction with the signal. The combined co-pumping and counter pumping leads to a bi-directional pumping.

## III. Experimental Research Design

The experimental set up is shown in Fig. 2.

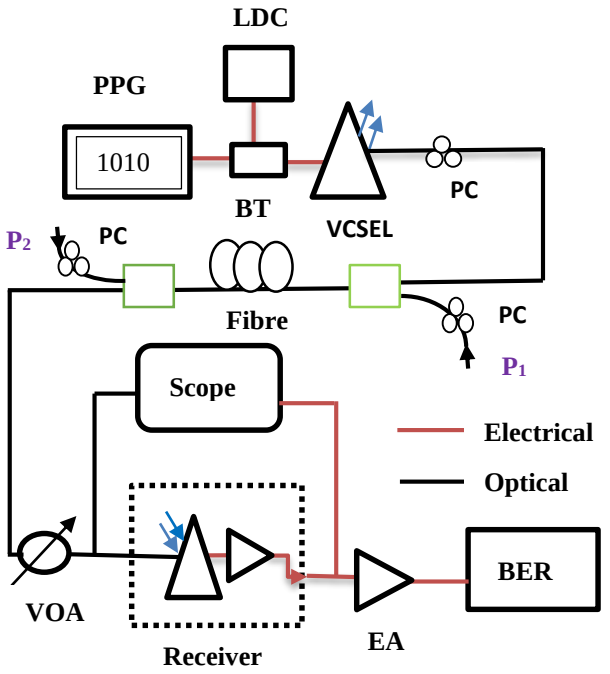



Figure 2: Experimental illustration of the set up. PPG is the programmable pattern generator, LDC is the laser diode controller, BT is the bias tee, PC is polarization controller, $P_1$ and $P_2$ are pump 1 and pump 2, VOA is the variable optical attenuator, EA is the electrical amplifier and BERT is the bit error rate tester. PPG generates the sequence, LDC tunes the bias current of the VCSEL and BT couples the DC and data onto the device.

A signal at wavelength, $\omega_s$ = 1548 nm from a VCSEL was coupled into a fibre. Raman pumps of wavelength, $\omega_p$ = 1448 nm at different pump powers were multiplexed onto the fibre and the signal output monitored. In the forward pumping scheme, only $P_1$ was used while in the backward scheme only $P_2$ was used. Polarization controllers were used to vary the orientation signal and pump states so as to ensure best coupling into the fibre. The on-off gain was established for various pump powers as well as the VCSEL wavelength tuneability range. For the bi-directional pumping both $P_1$ and $P_2$ were used. VCSEL was then modulated with a 4.25 Gbps $2^7$-1 pseudorandom bit sequence (PRBS) and the bit error rate (BER) measurements taken for different transmission distances. The VOA was used to vary the power received at the Avalanche photo diode (APD) receiver so as to measure the BER. EA amplified the received signal for BER measurements. On the BER measurement, different G.655, True wave non-zero dispersion shifted fibres (NZ-DSFs) of about 25 km each were joined together using optical connectors giving the various transmission lengths while on the gain characterization, only 25.3 km True wave Reach fibre [12] was utilized in this experiment.

## IV. Experimental Results and Discussion

The VCSEL is a low power device operating in the mA range as shown in Fig. 3. The lasing threshold is about 1.8 mA. The wavelength varies with change in the bias current. On the inset, the bias current was tuned from 4 mA to 9 mA giving a wavelength tuneability of about 3.5 nm (1547.5 to 1551 nm). The wavelength tuneability is an important feature for WDM-PON applications.

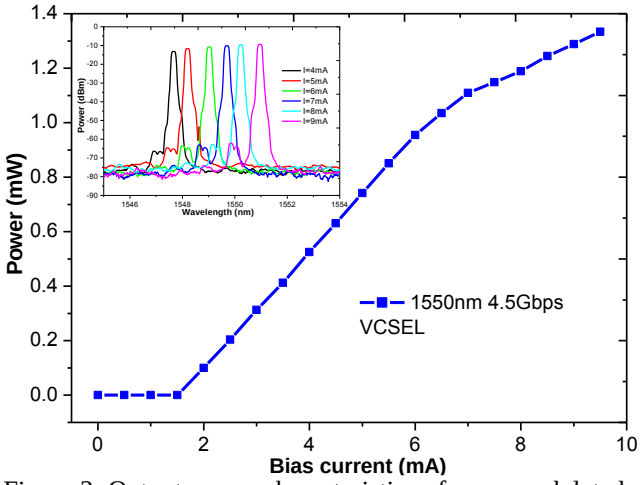



Figure 3: Output power characteristics of an unmodulated VCSEL. Inset: Wavelength tuneability with varying bias currents.

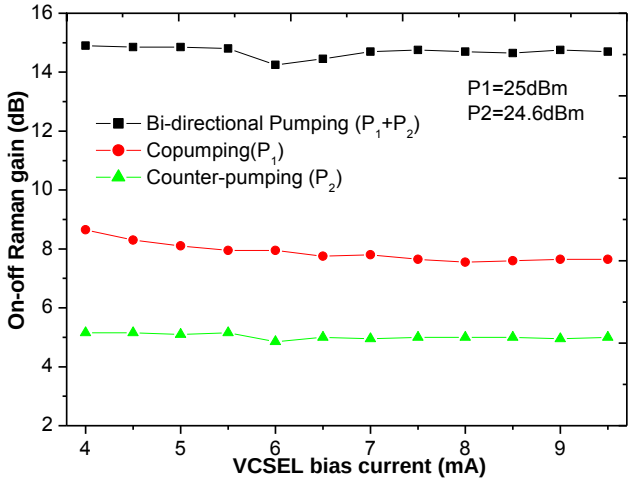



Figure 4: On-off gain for different bias currents for different pump configurations.

Fig. 4 shows on-off gain for counter, co- and bi-directional pumps. The input signal power was maintained at -11 dBm while the pump powers were

25 dBm and 24.6 dBm. The on-off gain refers to the difference between the signal power when the pump is on and when off. The condition for Raman amplification was achieved by setting the VCSEL bias current at 4.5 mA giving a frequency separation of 13.2 THz between the pump and the signal for optimum gain [11]. The gain spectrum appears flat because of the small wavelength range. In Fig. 4 and Fig. 5, a 25.3 km True wave-Reach fibre optimized for distributed Raman amplification [13] was used. The bi-directional scheme gave the highest on-off gain of about 15 dB while in the forward and the backward pumping schemes, 8 dB and 5 dB gains respectively were realized. In a transmission with a link loss of about 0.3 dB/km (fibre and connectors), the obtained gains can increase the reach by up to 50 km, 27 km and 17 km respectively.

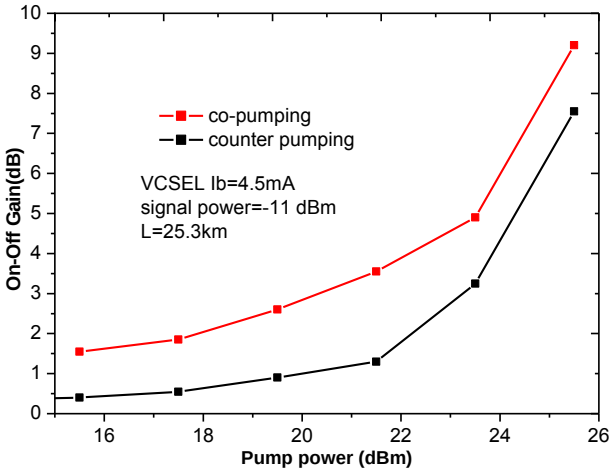



Figure 5: Relation between the gain and the pump power for different schemes.

In the bi-directional scheme, the two pumps provide more energy to be transferred to the signal thus reducing the depletion effect as compared to a one pump.

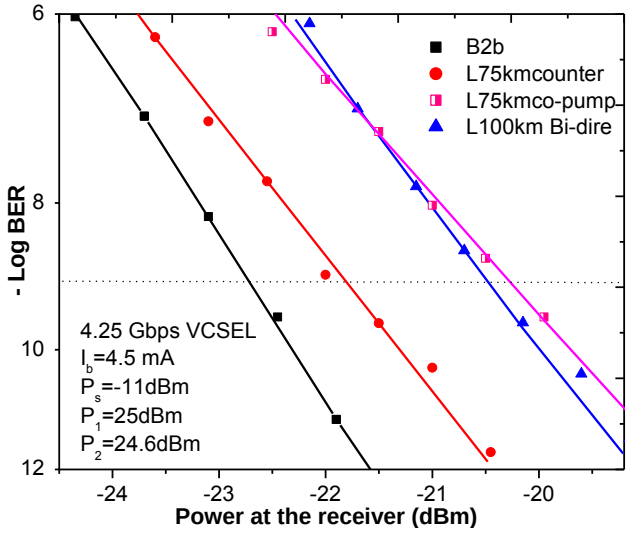



Figure 6: Experimental BER measurements of VCSEL transmission with Raman amplification.

The forward pumping has a higher on-off gain than backward pumping because of the enhanced signal - pump interaction. At the coupling end, the pump and signal are relatively strong and both gets weaker towards the end of the fibre. The energy transfer to a strong signal renders it the advantage of higher optical signal to noise ratio (OSNR) as compared to counter scheme [11]. In the backward configuration, a strong pump in the opposite direction interacts with a weaker signal thus the walk-off effect leads to a lower gain. This is because at the signal coupling end, the signal is strong when the pump is weaker thus reduced power transfer. However, for the counter scheme due to the pump direction, the pump to signal noise transfer are minimized. The bi-directional pumping balances the performance of forward and backward pumping. This is explained in detail with respect to amplified spontaneous emission noise, pump to signal noise transfer and multipath interference in the transmission medium [11, 14, 15]. The signal power increases with increase in pump power as shown in Fig. 5. The higher pump leads to more energy transfer to the signal thus more amplification.

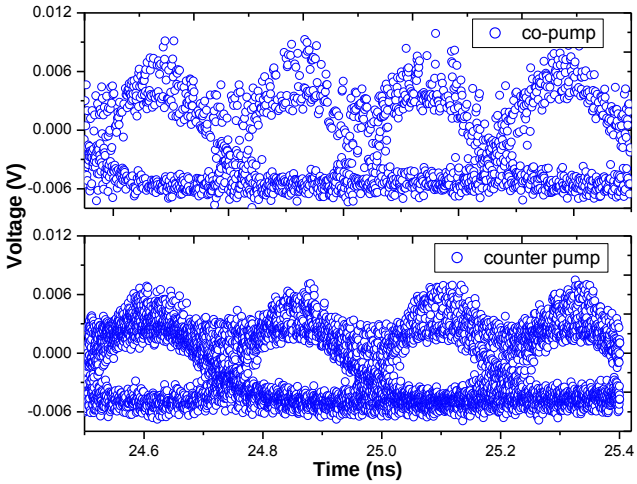



Figure 7: Eye diagrams of 75 km transmission with counter and co-pump schemes respectively.

Power penalties of 0.9 dB, 2.5 dB and 2.3 dB for 75 km counter pumped, 75 km co-pumped and 100 km bi-directionally pumped transmissions respectively were achieved. The penalty difference between the two pumping schemes at 75 km transmission is 1.6 dB. BER measurements show that the counter pump scheme has a better performance as compared to co-pumped at the 75 km fibre transmission as shown in Fig. 6. This may be attributed to walk-off effects that may reduce the pump to noise transfer to the signal giving a better signal [15].
The eye opening in Fig. 7 signify the signal clarity at 75 km transmission. The 100 km transmission shows a better performance as compared to 75 km co-pumped configuration. This signifies the advantage of bi-directional pumping in achieving a better signal amplification while balancing of co and counter scheme performance. It is expected that power penalties will increase at high transmission rates due to dispersive effects.

Generally, the combined transmission optimization reported in [16] and use of Raman amplification make the VCSEL performance suitable for long reach access networks. Raman pumps may however be costly as compared to employing distributed feedback lasers without amplification but the wavelength tuneability in

VCSELs makes it worth and suitable for access network application particularly WDM-PONs.

## V. CONCLUSION

Raman amplification has been tested successfully on VCSELs. High on-off gains of 15 dB, 8 dB and 5 dB in the bi-directional, forward and backward pumping schemes respectively enabling error free transmission were achieved. Transmission extending to about 80 km can be obtained using a single Raman pump in either scheme. Bi-directional scheme is an excellent configuration for longer reach extending up to 100 km. Therefore, Raman amplification proves vital and appropriate for VCSEL use in long access networks.

## ACKNOWLEDGEMENT

We are grateful for Research Funding and Support from: Telkom, Dartcom, Infinera, Ingoma Communication Services, NLC, NRF, THRIP and the ALC and Scholarship Funding from SKA/NRF SouthAfrica.

## REFERENCES

1. J. Prat, *Next-Generation FTTH Passive Optical Networks: Research towards unlimited bandwidth access*, Springer Science+Business Media B.V, Dordrecht, 2008. ISBN 978-1-4020-8469-0 e-ISBN 978-1-4020-8470-6.

2. R. Gamatham, E. Rotich, A. Leitch, and T. Gibbon, "Fibre-to-the-Hut: Research into tailored FTTH solutions for Africa," *AFRICON, 2013*, vol., no., pp.1-5, 9-12 Sept. 2013.

3. W. Hofmann, M. Müller, P. Wolf, A. Mutig, T. Gründl, G. Böhm, D. Bimberg and M.-C. Amann, "40 Gbit/s modulation of 1550 nm VCSEL," *Electr. Lett.* Vol 47, No. 4, pp. 270-271, (2011).

4. T. B. Gibbon, K. Prince, T. T. Pham, A. Tatarczak, C. Neumeyr, E. Rönneberg and M. Ortsiefer, "VCSEL Transmission at 10 Gb/s for 20 km Single Mode Fibre WDM-PON Without Dispersion Compensation or Injection Locking," *Elsevier. Opt. Fib. tech*. 17, pp. 41-45 (2011).

5. G. P. Agrawal, *Fibre-optic Communication Systems, 3rd Ed*. Chapter 6, pp. 232-260, John Wiley & Sons, Inc., New York, 2002.

6. E. Wong, "Next-Generation Broadband Access Networks and Technologies," *J. of Lightwave Technology*, Vol. 30, No.4, pp. 597-608, (2012).

7. K.L. Lee, J.L. Riding, A.V. Tran, R.S. Tucker, "Extended-Reach Gigabit Passive Optical Network for Rural Areas Using Distributed Raman Amplifiers," *Proc. of Opt. Fibre Comm. Conf. Nat. Fibre Optic Eng. Conf.* Paper NME3, 2009.

8. I. T. Monroy, R. Kjær, F. Öhman, K. Yvind, P. Jeppesen, "Distributed fibre Raman amplification in long reach PON bidirectional access links," *Opt. Fib. Techn.* 14, pp 41–44, (2008).

9. R. Kjaer, I. T. Monroy, L.K. Oxenloewe, P. Jeppesen and B. Palsdottir, "Bi-directional 120 km Long-reach PON Link Based on Distributed Raman Amplification," *Lasers and Electro-Optics Society, 2006. LEOS 2006. 19th Annual Meeting of the IEEE*, vol., no., pp.703-704, 2006.

10. G. P. Agrawal, *Nonlinear Fibre Optics*, 4th ed., chapter 8, pp. 298-319, Academic Press, San Diego, CA, 2007.

11. C. Headley and G. P Agrawal, *Raman amplification in fibre optical communication systems*, Elsevier Academic, Burlington, MA, 2005.

12. Truewave Reach fibre, low water peak, 2013. http://www.ofsoptics.com/resources/TrueWaveREACHFiber-124-web.pdf. Accessed 28th April 30, 2014.

13. T. Geisler and B. Pálsdóttir, "Selecting the Right NZDF Fibre For Distributed Raman Amplification," 2006. http://www.ofsoptics.com/resources/NZDF-for Distributed-Raman-Amplification.pdf. Accessed 28th April 30, 2014.

14. J. Bromage, "Raman Amplification for Fibre Communications Systems," *J. of Lightwave Technology*, Vol 22, No. 1, pp. 79-93, (2004).

15. Y. Emori, S. Kado and S. Namiki, "Independent Control of the Gain and Noise Figure Spectra of Raman Amplifiers using Bi-directional Pumping," *Furukawa Review*, No. 23, pp. 11-15, 2003.

16. E. K. Rotich Kipnoo, H. Kourouma, T. V. Chabata, R. R. G. Gamatham, A. W. R. Leitch and T. B. Gibbon, "Optimizing VCSEL Transmission for Longer Reach in Optical Access Networks," *Proc. 16th annual Southern Africa Telecommunication Networks and Applications Conference (SATNAC)*, Spier, Stellenbosch, Western Cape, South Africa, pp 27-30, 1-4th Sept 2013.

E. K. Rotich Kipnoo obtained a B.Sc. Hons (Physics/Maths) and an MSc Physics degree in 2008 and 2011 respectively at Moi University, Kenya. He is currently a PhD student at NMMU, working on SKA fibre Optical network. Research interests; Optical fibre communication optical networking, linear & nonlinear optical effects.

# Optimal Decoding of the Alamouti 4×2 Space-Time Block Coding

W.V. Kambale [1], K. Djouani [1, 2] and A.M. Kurien [1]

FSATI/Dept. of Electrical Engineering, Tshwane University of Technology [1],
P. O. Box X680, Pretoria 0001
Université Paris-Est [2], 12/LISSI Lab, Creteil, France

{kambalevw, djouanik, kurienam}@tut.ac.za

**Abstract-The extensive research that has been conducted on MIMO systems has demonstrated that Space-Time Block Coding is a pragmatic technique that successfully mitigates the effects of signal fading while efficiently increasing the channel capacity. While STBC possesses a relatively simple transmit encoding, the decoding process poses challenges with regards to complexity, particularly when more antennas are used on the receiver's side. This paper focuses on the optimisation of the decoding complexity of the Alamouti 4×2 STBC. It proposes the conditional optimisation of the Maximum Likelihood (ML) decoding of the Double Alamouti 4×2 STBC that leads to a reduced complexity from $O(N^4)$ to $O(N^2)$, where $N$ is the size of the underlying QAM signal constellation.**

**Index Terms— Conditional Optimisation, Double Alamouti coding, Maximum Likelihood (ML) detection, Space-Time Block Coding.**

## I. INTRODUCTION

Since the inception of wireless communications, there has been an endless strive to achieve high data rates without compromising the quality of reception. The birth of multi antenna systems, referred to as MIMO (Multiple Input Multiple Output), is a direct result of this quest. MIMO communications systems have therefore received extensive attention. By making use of antenna arrays at both the transmitter and the receiver, the aim has been to enhance the reliability of the wireless communication link and to increase the channel capacity without requiring greater power or extra bandwidth [1].

An effective and pragmatic way to benefit from the promises of MIMO systems is to employ Space-Time Coding (STC). Space-Time Coding is a MIMO technique that is based on introducing joint correlation in transmitted signals in both the space and time domains. By this approach, simultaneous diversity and coding gains can be obtained, as well as high spectral efficiency [2]. Numerous STC schemes have been proposed [3]. Space Time Block Coding (STBC) is one of these schemes that acts on a block of modulated symbols and then arranges them in a matrix format where the different symbols in the different columns are transmitted at different time slots while the symbols in different rows are transmitted over different antennas.

Of the many STBCs that have been designed, the Alamouti Code [4] is of particular interest. The Alamouti code is the first STBC scheme that provides full diversity for systems using two transmit antennas and one receive antenna.

Ensuring more reliable communication and enhanced communication performance implies increasing the number of antennas at the transmitter's and receiver's sides. However, the authors in [5] demonstrate that the Alamouti scheme, a full rate orthogonal design with complex symbols, is restricted to two transmit antennas. Besides, the increase in the number of antennas leads to decoding complexity at the receiver side.

This paper proposes an optimisation approach to the double Alamouti 4×2 STBC decoding complexity. The paper presents the details of the conditional optimisation of the Maximum Likelihood (ML) decoding of the Double Alamouti 4×2 STBC that results in a reduced complexity from $O(N^4)$ to $O(N^2)$.

The rest of the paper is organised as follows: Section II discusses related work. The MIMO System and channel model is presented in Section III. Section IV provides a discussion of the Alamouti 2×1 and the double Alamouti 4×2. In Section V, the fast optimal Maximum-Likelihood decoding of the double Alamouti 4×2 is discussed. Simulation results are discussed in Section VI and conclusions are given in Section VII.

## II. RELATED WORK

Numerous studies have been conducted on the decoding of 4x2 MIMO systems [6]-[9]. The authors in [6] provide a general design criteria for full-rate, fast-decodable STBCs and then introduce a family of 4x2 codes based on a combination of algebraic and quasi-orthogonal structures. However, in their design the full-diversity assumption is traded for simplified maximum-likelihood decoding of the code. The authors in [7] discuss a 4x2 MIMO scheme where a GLST-MIMO (Group Layered Space Time - MIMO) detector is adopted. The Transmit system of the scheme is combined with LST and STBC to claim a better trade-off between data rate and diversity while promising low-computational complexity. To achieve MLD (Maximum Likelihood Detection) reduced complexity, the authors in [8]

explore the use of antenna selection at the receiver side. The paper then proposes an alternative method that would yield processing complexity, the Constellation Partition (CP). The main idea of the method is the construction of the MLD search for the closest constellation point to the received signal in specific way that reduces the number of search operations. However, the particularity of the method provided in this work is the conditional optimization approach that leads to a fast optimal ML decoding.

## III. MIMO CHANNEL MODEL

This paper considers a MIMO communication system with $n_T$ transmit antennas and $n_R$ receive antennas as shown in Fig 1.

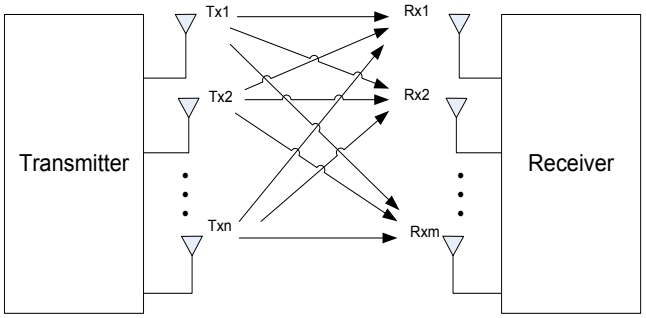



Figure 1: MIMO Block Diagram.

The MIMO channel model between the transmitting end and receiving end is considered as flat fading and can be represented by the following the $n_T \times n_R$ matrix:

$$H = \begin{bmatrix} h_{1,1} & \cdots & h_{1,nR} \\ \vdots & \cdots & \vdots \\ \vdots & \cdots & \vdots \\ \vdots & \cdots & \vdots \\ h_{nT,1} & \cdots & h_{nT,nR} \end{bmatrix}, \quad (1)$$

where $h_{i,j}$ denotes the channel response between the transmit antenna $i$ and the receive antenna $j$; with the received signal being expressed as:

$$y = Hs + n, \quad (2)$$

where $y$ represents the ($n_R \times 1$) matrix of the received symbols, $H$ the ($n_R \times n_T$) MIMO channel matrix, $s$ the ($n_T \times 1$) transmit symbol vector and $n$ the ($n_R \times 1$) additive noise vector. The channel $H$ is chosen randomly according to a Rayleigh distribution in a quasi-static channel since the channel is assumed to be constant over a block of $T$ symbol periods.

## IV. SPACE-TIME BLOCK CODING SCHEMES

Space Time Block Coding is a simple yet effective transmit diversity technique used in MIMO technology. A STBC scheme acts on a block of modulated symbols that it arranges in a matrix format in such a way that different symbols in the different columns are transmitted at different time slots and symbols in different rows are transmitted over different antennas. Two schemes are discussed below:

### A. Alamouti Code

Historically, the Alamouti Code [4] is known to be the first space-time block code to provide full transmit diversity for systems with two transmit antennas. The scheme exploits the spatial and time diversity to improve the quality of the received signal with a less complex processing at the transmitter and a linear decoding at the receiver.

Its encoding process takes a block of two *N-QAM* modulated symbols, $s_1$ and $s_2$, in each encoding operation and maps them to the transmit antennas according to a code matrix given by

$$S = \begin{bmatrix} s_1 & s_2 \\ -s_2^* & s_1^* \end{bmatrix}, \quad (3)$$

where * represents the complex conjugate. The encoder's outputs are transmitted in two consecutive transmission periods from two transmit antennas. From the code matrix, *S*, each row represents the first and second transmission periods respectively. The first and second columns correspond to the symbols transmitted from the first and second antenna, respectively.

At the receive antenna, the received signals over the two transmission periods, denoted by $r_1$ and $r_2$, can be expressed as

$$\begin{aligned} r_1 &= h_1 s_s + h_2 s_2 + n_1 \\ r_2 &= -h_1 s_2^* + h_2 s_1^* + n_2 \end{aligned} \quad (4)$$

where $h_i$ is channel response between the transmit antenna and the receive antenna. $n_1$ and $n_2$ represent the additive white Gaussian noise (AWGN) vectors.

At the decoder side, the transmitted symbols $s_1$ and $s_2$ are estimated by a maximum-likelihood detector from the combining of the received signals, from Eq.4, according to the decision statistics given by [4]

$$\begin{aligned} \tilde{s}_1 &= h_1^* r_1 + h_2 r_2^* \\ \tilde{s}_2 &= h_2^* r_1 - h_1 r_2^* \end{aligned} \quad (5)$$

### B. Double Alamouti 4×2 STBC

The double Alamouti 4×2 STBC is a variant of the Alamouti scheme. It consists in simultaneously transmitting two Alamouti codes on two blocks of two antennas as depicted in Fig. 2. Technically, it transmits four N-QAM symbols using four transmit antennas in two transmission periods making it a rate-2 code.

The double Alamouti code matrix is given by [10]

$$S = \begin{bmatrix} s_1 & -s_2^* \\ s_2 & s_1^* \\ s_3 & -s_4^* \\ s_4 & s_3^* \end{bmatrix} \quad (6)$$

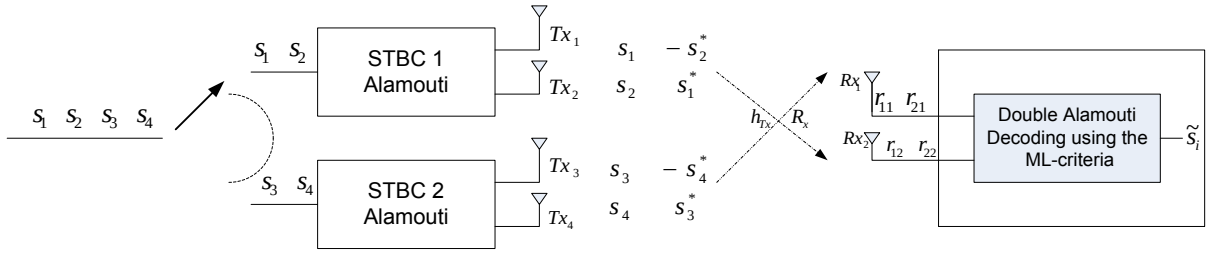



Figure 2: Double Alamouti 4×2 STBC coding scheme.

Since the channel model is considered quasi-static in this scheme, the channels are assumed to be constant across two consecutive symbol transmit periods.
The received signals obtained result from the matrix multiplication as shown in [10].

$$\begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix} = \begin{bmatrix} s_1 & s_2 & s_3 & s_4 \\ -s_2^* & s_1^* & -s_4^* & s_3^* \end{bmatrix} \bullet \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \\ h_{31} & h_{32} \\ h_{41} & h_{42} \end{bmatrix} + \begin{bmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{bmatrix} \quad (7)$$

For the purpose of signal decoding, (7) is rewritten as

$$\begin{bmatrix} r_{11} \\ r_{12} \\ r_{21} \\ r_{22} \end{bmatrix} = \begin{bmatrix} h_{11} & h_{21} & h_{31} & h_{41} \\ h_{12} & h_{22} & h_{32} & h_{42} \\ h_{21}^* & -h_{11}^* & h_{41}^* & -h_{31}^* \\ h_{22}^* & -h_{12}^* & h_{42}^* & -h_{32}^* \end{bmatrix} \bullet \begin{bmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \end{bmatrix} + \begin{bmatrix} n_{11} \\ n_{12} \\ n_{21}^* \\ n_{22}^* \end{bmatrix} \quad (8)$$

## V. Optimal Maximum Likelihood Of The Double Alamouti 4×2 STBC

The generic Maximum-Likelihood detection algorithm [11] is given by

$$\hat{s}_j = \arg\min_{s_j \in S} \|r - Hs\|^2 \quad (9)$$

where $\hat{s}_j$ is the estimated symbol, $S$ the set of N-QAM constellation, $r$ the received signal and $H$ the channel matrix. The detector calculates the Euclidian distance $\|r - Hs_j\|$ in which the signal with the minimal distance is retained and taken as the estimate of the transmitted signal.

However, this algorithm possesses $O(N^4)$ complexity since there are $N^4$ metric values from which the minimum metric needs to be chosen through ML decoding exhaustive search. To reduce the complexity, a fast optimal ML decoding approach is proposed. This is based on the conditional optimization [12] which is a decoding primitive that is well matched for high rates space-time codes obtained by multiplexing simple blocks. The decoding problem is defined as

$$\begin{aligned} r_1 &= cH_1 + sG_1 + n_1 \\ r_2 &= cH_2 + sG_2 + n_2 \end{aligned} \quad (10)$$

where $r_1$ and $r_2$ are the received signal vectors, $c = (x_1, x_2)$ and $s = (x_3, x_4)$ are pairs of transmitted symbols drawn from a QAM constellation $C$ of size $N$, the noise vectors $n_1$ and $n_2$ are complex Gaussian random with zero mean and variance $\sigma^2$ and $H_1$, $H_2$, $G_1$, $G_2$ represent the channels induced by the receiver referred to as "Alamouti blocks" expressed as

$$\begin{aligned} H_1 &= \begin{pmatrix} h_{11} & -h_{21}^* \\ h_{21} & h_{11}^* \end{pmatrix}, \quad G_1 = \begin{pmatrix} h_{31} & -h_{41}^* \\ h_{41} & h_{31}^* \end{pmatrix}, \\ H_2 &= \begin{pmatrix} h_{12} & -h_{22}^* \\ h_{22} & h_{12}^* \end{pmatrix}, \quad G_2 = \begin{pmatrix} h_{32} & -h_{42}^* \\ h_{42} & h_{32}^* \end{pmatrix} \end{aligned} \quad (11)$$

The fast optimal ML algorithm is implemented with complexity $O(N^2)$ as shown below. Primarily, (9) is condensed and rewritten as

$$r = cH + sG + n \ , \quad (12)$$

where $r = (r_1, r_2)$, $H = (H_1, H_2)$, $G = (G_1, G_2)$ and $n = (n_1, n_2)$. The likelihood function associated with (9) is expressed as:

$$p(r \mid s, c) \alpha \exp\left(-\frac{1}{2\sigma^2}\|r - cH - sG\|^2\right) \quad (13)$$

where $\|.\|$ is the Euclidian norm and the ML solution is given by

$$(\hat{c}, \hat{s}) = \arg\max_{c,s \in C^2} p(r \mid s, c) \quad (14).$$

The conditional optimisation decoding of the Double Alamouti is achieved by first choosing a symbol pair $s = [x_1, x_2]$ where $x_1$ and $x_2$ are from the constellation $C$. Then, given $s$, the symbol pair $\tilde{c}$ is calculated by

$$\tilde{c}(s) = \frac{rH^\dagger - sGH^\dagger}{\|H\|^2} \quad (15)$$

where $\|H\|^2 = \frac{1}{2}\|H\|_F^2$, and $\|.\|_F$ denotes the Frobenius norm. The next stage consists of finding the closest constellation pair given by

$$\hat{c}(s) = Q(\tilde{c}(s)) = (Q(\tilde{x}_3(s)), Q(\tilde{x}_4(s))) \ , \quad (16)$$

where $Q(.)$ is the quantizer for QAM constellation. Finally, the decision metric is computed by

$$M_s = \|r - sG - \hat{c}(s)H\|^2 \quad (17)$$

However, the algorithm needs to explore all the possible symbols pairs (combinations of $x_1$ and $x_2$) so that a decision can be made on the smallest decision metric. This is achieved by repetition of equations (14) and (16) and the smallest decision metric found by

$$\begin{cases} \hat{s} = \arg \min_{x_1, x_2 \in C} M_s \\ \hat{c} = Q(\tilde{c}(\hat{s})) \end{cases} \quad (18)$$

## VI. Discussion of Results

Conventional generic ML and conditionally optimised ML algorithms were applied to the double Alamouti 4x2 STBC. Performance results of both algorithms were implemented in MATLAB and are presented in this section. Perfect knowledge of the channel by the receiver was assumed and the channel was considered constant during two consecutive transmissions for both schemes.

In the conventional ML algorithm, an exhaustive search is performed through all possible transmitted vectors. That is, for the 4×2 STBC considering N-QAM, the complexity is of order $O(N^4)$, $N$ being the size of the QAM signal constellation. However, the conditionally optimised ML algorithm reduces the constellation alphabet by optimising over a subset of some symbols conditioned on other remaining symbols [12]. The algorithm, hence, takes 2 symbols from the constellation alphabet, finds the estimation of the other 2 symbols and then calculates a decision metric based on these 4 symbols. This process is repeated until all possible combinations ($N^2$) of symbols from the constellation map are scanned through. Finally, the minimum value of the calculated decision metrics is singled out and the actual transmitted symbols recovered. This way, the complexity is reduced to $O(N^2)$.

Fig. 3 illustrates the performance results of the conventional ML and optimal conditionally optimised ML decoding schemes in terms of Symbol Error Rate (SER) versus Signal to Noise ratio (SNR) for a 4-QAM signal constellation.

A small degradation in performance is observed in the conditionally optimised ML implementated scheme. However the trade-off is that the computational complexity has been reduced from $O(N^4)$ to $O(N^2)$ and especially many wireless communications (WiMAX, 3GPP LTE) standards may welcome this optimal conditionally optimised ML of the double Alamouti 4x2 STBC for implementation.

## VII. Conclusion

This paper presented the conditional optimization of the Maximum Likelihood (ML) decoding of the Double Alamouti 4×2 STBC. SER curves of the conventional ML and the conditionally optimised ML were presented. The proposed algorithm presents a decoding algorithm that demonstrates reduced complexity from $O(N^4)$ to $O(N^2)$, and hence termed fast optimal.

Both the conventional and the conditionally optimised ML algorithms present nearly similar curves but with small degradation in the proposed scheme. However, the proposed scheme presented an algorithm with reduced complexity.

Future work of this research will consider applying conditional optimisation to higher-order Alamouti-based MIMO techniques [13].

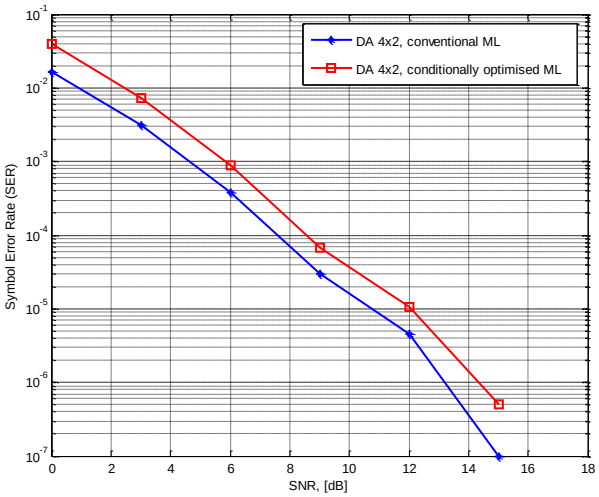



Figure 3: Performance results of the conventional ML and the conditionally optimised ML of the Double Alamouti 4×2 STBC scheme, 4-QAM.

## References


[1] D. Gesbert, M. Shafi, D.S. Shiu, P. J. Smith and A. Naguib, "From theory to practice: an overview of MIMO space-time coded wireless systems". *Selected Areas in Communications, IEEE Journal on*, 21(3), 281-302, 2003

[2] B. Vucetic and J. Yuan, Space-Time Coding, 2003: Wiley

[3] M. Jankiraman, Space Time Codes and MIMO Systems, Artech House, 2004

[4] S. M. Alamouti, "A Simple Transmit Diversity Technique for Wireless Communications," *IEEE Journal on Selected Areas in Communication*, Vols. 16, no. 8, pp. 1451-1458, October 1998.

[5] V. Tarokh, H. Jafarkhani and A. R. Calderbank, "Space-Time Block Codes from Orthogonal Designs," *IEEE Transactions on Information Theory*, vol. 45, no. 5, pp. 1456-1467, July 1999.

[6] E. Biglieri, Y. Hong, and E. Viterbo, "On fast-decodable space-time block codes," *IEEE Trans. Inform. Theory*, vol. 55, no. 2, pp. 524–530, Feb. 2009.

[7] K. Higashi, Y. Nagao, M. Kurosaki, and H. Ochi, "FPGA Implementation of a GLST-MIMO System," *The 9th International Conference on Advance Communication Technology*, no. 02E-03, pp. 282-285, Feb. 2007

[8] Y.A. Zahrani, S. Marshed, A. Dhofyan, A.I. Sulyman, S. Dosari, M. Elnamaky and S. Shebeili, "Design and FPGA Implementation of Reduced Complexity MIMO-MLD Systems." *Signal Processing and Information Technology (ISSPIT), 2010 IEEE International Symposium on* , vol., no., pp.348,353, 15-18 Dec. 2010

[9] V. Kambale, K. Djouani and A. Kurien,"Toward an FPGA Hardware Implementation of the Alamouti 4x2 Space-time Block Coding." *Procedia Computer Science, The 4th International Conference on Ambient Systems, Networks and Technologies (ANT)*, Volume 19, Pages 602–608, 2013.

[10] J.B. Bastos, and A. Gammeiro, "Performance of Extended Space-Time Coding Techniques for MIMO MC-CDMA Systems," *Proc European Conf. on Antennas & Propagation - EUCAP*, Nice, France, Vol. 1, pp. 1 - 5, November 2006.

[11] A. Van Zelst, "Space Division Multiplexing Algorithms," *in Proc. 10th Mediterranean Electrotechnical Conference*, vol. 3, pp. 1218-1221, 29-31 May 2000.

[12] S. Sirianunpiboon, Y. Wu, A. R. Calderbank and S. D. Howard, "Fast Optimal Decoding of Multiplexed Orthogonal Designs by Conditional Optimization," *IEEE Transactions on Information Theory*, vol. 56, no. 3, pp. 1106-1113, March 2010.

[13] S. Morosi, F. Argenti, M. Biagini and E. Del Re, "Comparison of channel estimation algorithms for MIMO downlink LTE systems," *In Wireless Communications and Mobile Computing Conference (IWCMC), IEEE*, pp. 953-958, July 2013

**Witesyavwirwa Vianney Kambale** received his undergraduate degree in 2010 from Tshwane University of Technology and is completing his Master of Science degree (ESIEE-Paris) and MTech Electrical Engineering at the Tshwane University of Technology. He is currently a Junior Lecturer in the Faculty of ICT at the same institution. His research interests include Signal Processing, Space Time Coding and FPGA Implementation of MIMO Systems.

# A Cross-layer Based Subchannel Allocation Scheme in Satellite LTE Networks

Gbolahan Aiyetoro[1] and Fambirai Takawira[2]

Department of Electrical, Electronic and Computer Engineering
University of KwaZulu-Natal[1], Durban 4041
Tel: +27 312602736, Fax: +27 312602500

School of Electrical and Information Engineering,
University of the Witwatersrand[2], Johannesburg, South Africa
email: g.aiyetoro@ieee.org[1], fambirai.takawira@wits.ac.za[2]

**Abstract- This paper proposes a new subchannel allocation scheme for satellite LTE with the adoption of MIMO technology. The Satellite Long Term Evolution (LTE) air interface will provide global coverage and hence complement its terrestrial counterpart in the provision of LTE services to mobile users. The user scheduling scheme is a vital element that is needed in order to effectively utilize the resources of the satellite LTE network. However, a user scheduling scheme without a scheme that will ensure effective allocation of available subchannels to the selected users will undermine the utilization of the resources of the satellite LTE network. The aim of this paper is to propose a new subchannel allocation scheme and compare with existing scheme through simulations, using throughput, spectral efficiency, and fairness index as the performance metrics. A land mobile dual-polarized GEO satellite system has been considered for this work.**

**Index Terms— GEO satellite, LTE, Subchannel allocation scheme, Packet scheduling**

## I. INTRODUCTION

The rapid growth in mobile users and continuous increment in the demands for different types of telecommunication services, like video streaming, video conferencing, Voice over IP (VoIP), Web browsing, multimedia messaging, video gaming and FTP downloads have necessitated the need for new technologies able to provide high data rates and also meet up with the demands of their respective Quality of Service (QoS) requirements. It is also worth to note that the available bandwidth is limited and this has made high spectrum efficiency an important target that must be addressed by future technologies.

The need to address these important challenges in future mobile networks formed the basis for International Telecommunication Union Radiocommunication sector Working Party 8F (ITU-R WP 8F) to define the future Fourth Generation Mobile (4G). The set of transmission capacity and QoS requirements are specified which allow any technology that meets up with these requirements to be included in the IMT-Advanced family [1]. This has led to the emergence of LTE and WiMAX 802.16x. Though, these two technologies do not fulfill the requirements, they are first steps towards the 4G [2]. The LTE technology, which is of interest to this paper, is made up of the radio access and packet core networks. LTE uses Orthogonal Frequency Division Multiple Access (OFDMA) as its multiple access technology and it also employs Multiple-In Multiple-Out (MIMO) technology [3].

In order to provide seamless mobile services to users irrespective of their locations, the satellite component of 4G systems will play a vital role, since the terrestrial component will not be able to provide a global coverage due to economic and technical limitations [4]. Therefore, future satellite air interfaces need to have a high-level of commonality with the 4G terrestrial air interface. Hence, both 3GPP LTE and WiMAX air interfaces have been proposed for the satellite scenario. An S-band GEO satellite system has been recommended for this purpose [5].

The ambitious 4G targets in terms of QoS, data rates and fairness can only be achieved with an effective scheduling scheme that will provide an optimal balance of all these requirements. This effective scheduling scheme is expected to comprise of user scheduling and subchannel allocation or Physical Resource Block (PRB) mapping schemes. Once, the user(s) have been selected by the user scheduler, the need for a subchannel allocation or PRB mapping scheme that will allocate or map the available subchannels/PRBs in order to effectively utilize the network is very crucial. This scheme must not only utilize the network but also be sensitive to the QoS requirements of the mobile users.

Several schemes have been proposed for terrestrial LTE or OFDMA based or MIMO based networks in the literature. In [6], an optimal solution called Hungarian method is used as an assignment problem solution; however, the major issue is that this solution is limited to a scenario where the number of users is equal to the number of channels. An exact solution called the Kuhn-Munkres algorithm is proposed [7], and near-optimal solution like Max Loss Delete (MLD) and Max Difference Top Two (MDTT) are proposed in [8-9]. However, the major challenge for all these solutions is also that the solutions are limited to scenarios where the number of users is equal to the number of channels or antennas. While [10] proposed a heuristic solution based on the auction algorithm which is applicable to the network

considered, only data rate is used and other QoS factors are not considered.

It is worth noting that exact solutions are only applicable to scenarios where the number of users is equal to the number of subchannels or antennas. Hence, for the network considered, where the number of subchannels are often more than selected number of users, only heuristic solutions are applicable. It is also worth noting that these solutions have been proposed for terrestrial networks but to the best of our knowledge, no solutions have been proposed for satellite LTE networks. It is against this background that this paper proposes a new subchannel allocation scheme that is derived using heuristic solutions with the aim of maximizing the utility of the satellite LTE network.

The rest of this paper is organized as follows: System description is presented in Section 2. In Section 3, the considered user scheduler and the subchannel allocation schemes are presented. Section 4 and 5 presents the simulation model and results respectively. Finally, Section 6 concludes the paper.

## II. System Description

The satellite LTE radio access technology is envisaged to use OFDMA for downlink transmission just like its terrestrial counterpart. OFDMA can be adopted for satellite as stated in [5], due to the fact that it easily exploits frequency selectivity and allows flexible bandwidth operation with low complexity receivers. It supports both Frequency Division Duplexing (FDD) and Time Division Duplexing (TDD) and allows for a wide range of different bandwidths (1.5, 3, 5, 10, 15 and 20 MHz) [11]. It also supports downlink multi-antenna schemes including both transmit diversity, spatial multiplexing and beamforming [12]. The spatial multiplexing, which includes single user and multi-user MIMO is of interest to this work. For the downlink of 3GPP LTE, the 2 x 2 MIMO is assumed to be the baseline configuration and 4 x 4 MIMO is also envisaged [13].

The transmission mode 5, which is for MU-MIMO, has been considered for this work since the focus here is to evaluate the performance of scheduling algorithms of a satellite LTE in MU-MIMO transmission mode. The details of these modes are presented in [13]. For this work, the evolved Node B (eNodeB), which acts as the base station in satellite LTE scenario, is located on the earth station and it is considered to be equipped with two transmit antennas; the User Equipment (UE) has two antennas as well, according to the 2 x 2 MIMO configuration.

### A. Satellite Air Interface

A transparent GEO Satellite has been adopted for this work. Dual-polarized antennas consisting of Right Hand Circular Polarized (RHCP) and Left Hand Circular Polarized (LHCP) antennas have been considered for both the GEO satellite and UEs. As shown in Fig. 1, the satellite eNodeB uses two satellite dishes to transmit via the dual-polarized antennas of the GEO satellite to mobile users, as proposed in [14]. Hence, the downlink is formed between the eNodeB and the UE. This allows simultaneous transmissions from the two polarized antennas of the GEO satellite to different UEs.

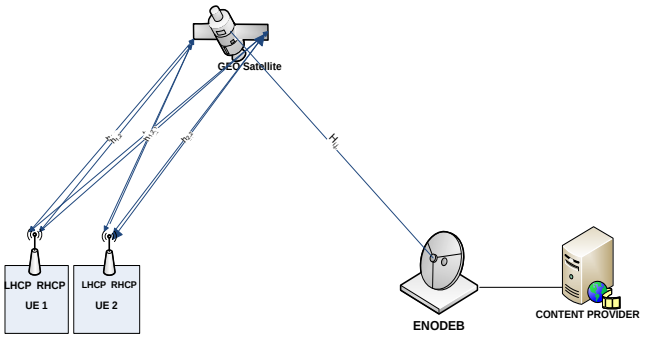


Figure 1 The system architecture of a satellite LTE network

This transmission mode is closed-loop, hence, there is a UE feedback for link adaptation purposes, which is very vital in determining the transmission rate. The UE will send the Channel Quality Indicator (CQI) message via the GEO satellite as recommended in [5] to the eNodeB on the earth station.. The reported CQI is used for transmission and scheduling purposes at the eNodeB.

At the Media Access Control (MAC) layer of the eNodeB, the packet scheduler works with the Link Adaptation (LA) module and Hybrid Automatic Repeat Request (HARQ) to schedule users on resources at every Transmission Time Interval (TTI) which is 1 ms, as specified in LTE. The basic time-frequency resource that is allocated is the Physical Resource Block (PRB). Each PRB consists of 12 consecutive subcarriers (180 kHz of the whole bandwidth) for duration of 0.5 ms for each slot [15]. It is worthy to note that the resource allocation is only finalized after every subframe of 1 ms. This means a pair of PRBs (i.e. scheduling block) is the resource allocation granularity and its on a TTI (= 1 ms) basis.

The users selected by the scheduler are mapped to the available pair of PRBs using the subchannel/PRB allocation scheme at every TTI (1 ms). The number of available PRBs in a scheduling interval depends on the size of bandwidth used and the number of antennas deployed (here 2 x 2 MIMO). The number of PRBs for a single antenna ranges from 6 to 100, depending on the bandwidth size, which ranges from 1.4 to 20 MHz [16].

### B. Channel Model

The channel model that is considered here is an empirical-stochastic model for LMS-MIMO [17]. This is based on the fact that the model is validated and compared to other existing models, it considers interdependence between small scale fading. The stochastic properties of this model are derived from an S-band tree-lined road measurement campaign (suburban area) using dual circular polarizations at low elevations [17]. The channel matrix, H, is made up of co-polar and cross-polar circularly-polarized channels and is represented as follows:

$$H = \begin{pmatrix} h_{RR} & h_{LR} \\ h_{RL} & h_{LL} \end{pmatrix} \quad (1)$$

The channel matrix, H, takes large scale fading (shadowing) and small scale fading (multipath) into account. A Markov chain is used to select between the possible regions of high and low shadowing values for both co-polar and cross-polar channels to model the mobile user movement across the buildings. There are four possible

Markov states as presented in Fig. 2. The four possible states are due to the high or low state of both the co-polar (CP) and cross-polar (XP) channels. State transitions in the chain in Fig. 2 occur on a TTI basis.

The 4 x 4 transition matrix, P, below is used to predict the next possible state. The columns of the matrix represent the probability of one state moving to another listed in the right hand column while the rows represent the probability of moving to the state on the right hand column from the previous state on the bottom row. State 1 is CP Low XP Low, State 2 CP Low XP High, State 3 is CP High XP Low and State 4 is CP High XP High. The probability matrix below is derived from the measurements obtained in [17]. The top right corner value of 0.1037 is the probability of "CP High, XP High" to "CP Low XP Low".

$$P_{ij} = \begin{bmatrix} 0.6822 & 0.1579 & 0.0561 & 0.1037 \\ 0.2887 & 0.2474 & 0.0447 & 0.4192 \\ 0.1682 & 0.0966 & 0.1745 & 0.5607 \\ 0.0098 & 0.0199 & 0.0150 & 0.9554 \end{bmatrix} \quad (2)$$

The large scale (shadowing) fading generation depends on the Markov chain. A high or low shadowing is generated on the basis of the state. The small scale fading is modelled using Ricean distribution. The Ricean fading for each of the MIMO branch is generated using Ricean factors. The details on how the large scale and small scale fading are obtained are shown in [17].

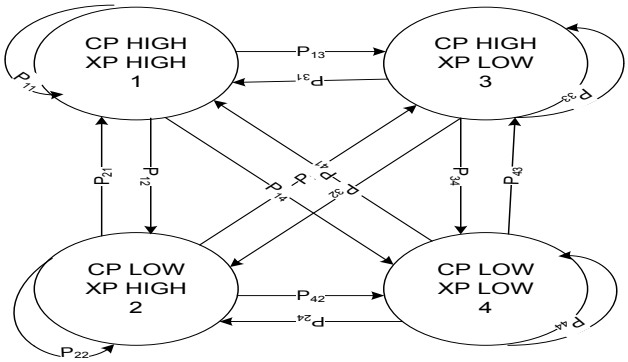



Figure 2 Four-state Markov model of an LMS-MIMO channel

Though the varying distance is of less significance to the total path loss, the path loss (in dB) at 2 GHz is computed as follows:

$$L_{Fs} = 190.35 + 20\log\left(\frac{38500+D}{35788}\right) \quad (3)$$

The large scale fading and small scale fading obtained in (1) above are considered together with the path loss ( $L_{FS}$ ) and polarization loss as part of the total loss experienced in the channel. The Signal-to-Noise-Interference Ratio (SNIR), which is obtained on a subchannel basis by dividing the received power by the noise power, can be expressed as follows:

$$SINR(dB) = EIRP + G_R - L_{Total} - (N + I)\ (dB) \quad (4)$$

The EIRP value of 63 dB, Polarization loss of 3.5 dB and a noise of 208.1 dBm for each subchannel is used to compute the SNIR. Also considered, it's the inter-spotbeam interference, *I*, as a result of power received from eNodeBs sharing the same frequency. The SNR-CQI mapping derived from [18] for a Block Error Rate (BLER) of $10^{-3}$ is used to determine the CQI from obtained SNR. This can be presented as follows;

$$\begin{gathered} if\ SNR < -3.8;\quad CQI = 1 \\ if\ -3.8 \le SNR \le 22.6; CQI = (0.55 * SNR) + 3.45 \\ if\ SNR > 22.6;\quad CQI = 15 \end{gathered} \quad (5)$$

The CQI distribution of a mobile user with speed of 30 km/h is presented in Fig. 3. Based on the reported CQI, an appropriate Modulation and Coding Scheme (MCS) is used to transmit the packets of the selected mobile users. A much lower BLER target of $10^{-3}$ has been considered as compared to the BLER target of $10^{-1}$ that is used for the terrestrial scenario, since if the first transmission is unsuccessful in the terrestrial scenario, retransmission can quickly be employed to recover the lost packets. However, this is not the case for satellite scenario due to the long RTPD experienced. This practically prevents the use of retransmissions to recover lost packets (real-time traffic).

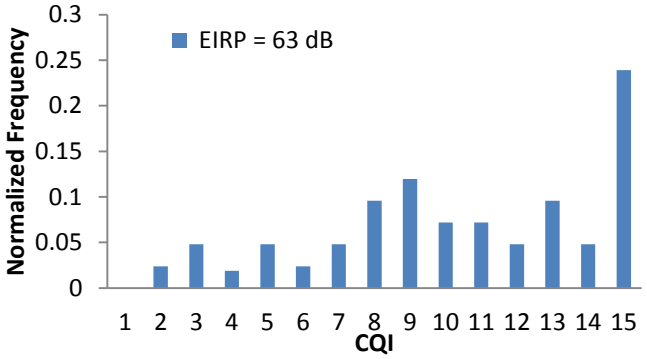



Figure 3. The CQI distribution of a UE at 30km/h

### C. *Traffic Model*

The video streaming and web traffic have been considered for this work. The video streaming traffic is modeled using realistic video trace files that are provided in [19]. The mean bit rate of the video traffic source is 440 kbps.

The Web traffic oscillates between an ON (packet call) state and OFF (reading time) state. During the packet call state the web source produces a number of messages that is geometrically distributed with the mean value of 300 and the inter-arrival time is also geometrically distributed but with the mean value of 0.5s. The OFF time is also geometrically distributed with the mean value of 2s during which no traffic is generated. The message length is $l_{w\_bytes}$ which denotes a floor function of a random variable *x* which have the following truncated Pareto pdf [20].

$$g(x) = \frac{\zeta . l_{w\_min}^{\zeta}}{x^{\zeta+1}} . \left[u(x - l_{w\_min}) - u(x - l_{w\_max})\right] + \left(\frac{l_{w\_min}}{l_{w\_max}}\right)^{\zeta} . \delta(x - l_{w\_max}) \quad (6)$$

Where ζ is 1.1, u(.) is the unitary step function, δ (.) is the Dirac Delta function, a $l_{w\_min}$ and $l_{w\_max}$ are the minimum (815 bytes) and maximum (664 Kbytes) message length respectively.

## III. SCHEDULING SCHEME

The packet scheduling scheme, which consist of the user scheduler and the subchannel allocation scheme, to be used for satellite LTE network, is meant to utilize the available resources, ensure fairness and satisfy the QoS requirements of varying users.

A single spotbeam is considered which consists of a base station (eNodeB) where the downlink bandwidth is divided into *J* subchannels. It is assumed that in this scenario the scheduler only schedules users, without assigning them to specific subchannels. A set of *N* users with the best priority index or utility function or metrics, selected by the scheduler from the total number of *K* users is considered for the purpose of subchannel allocation or mapping. The set of subchannels are denoted by $J = \{j | j = 1,2,3, \dots, J\}$ and the set of users are denoted by $N = \{n | n = 1,2,3, \dots, N\}$. $\lambda_{n,j}$ denotes whether user $n$ is mapped or assigned to subchannel $j$ or not; if assigned, it will have a value of 1 or else it will be 0. $\lambda_{n,j}$ represents the utility achieved to transmit the data of user $n$ over subchannel $j$ as shown in (10).

The assignment problem, which aims to maximize the total utility function of the network, can then be expressed as follows;

$$max \sum_{j=1}^{J} \sum_{n=1}^{N} U_{n,j} \lambda_{n,j} \qquad (7)$$

Subject to;

$$\sum_{j=1}^{J} \lambda_{n,j} \leq J \qquad \forall n \in (1, \dots, N) \qquad (8)$$

$$\sum_{n=1}^{N} \lambda_{n,j} \leq 1 \qquad \forall j \in (1, \dots, J) \qquad (9)$$

$$\lambda_{n,j} \in \{0,1\} \qquad (10)$$

Constraint (8) states that a user can be assigned to as many subchannels as possible depending on the number of subchannels available. It has also been shown that this constraint does not produce a significant reduction in the total utility of the network, and constraint (9) states that a maximum of one user can be assigned to a subchannel. Hence, a subchannel can only accommodate one user at a time (or each timeslot).

### A. User Scheduler

In order to achieve this objective, a user scheduler which use both channel conditions and QoS factors in making scheduling decisions must be considered. Hence, a common throughput-optimal scheduler, M-LWDF, has been considered in this paper. The M-LWDF scheduler provides good QoS performance with each user having its own probabilistic QoS requirement of the form:

$$\Pr(W_k > T_{k,deadline}) \leq \delta_k \qquad (11)$$

where $W_k$ is the waiting time of the Head Of Line (HOL) packet in user queue *k*, $T_{k,\ deadline}$ is the delay deadline for user $k$ and $\delta_k$ is the probability of exceeding the delay deadline. According to [21], the scheduler selects user k for subchannel $j$ based on:

$$U_{k,j} = max \left( \frac{R_{k,j}(n)(-\log \delta_k) w_k(n)}{T_k(n) \quad T_{k,deadline}} \right) \qquad (12)$$

where $W_k(n)$ is the waiting time of the HOL packet in user queue k at TTI n, $R_{k,j}(n)$ is the instantaneous transmission rate of user *k* for subchannel *j* at TTI *n*, $T_k(n)$ is the average transmission rate of user *k* over previous TTI before TTI *n* and $T_{k,deadline}$ is the delay deadline for the packet, this varies depending on the traffic type. $T_{k,deadline}$ for RT packet is assumed to be 160 ms. The $\delta_k$ varies based on the priority of the service being demanded by the user.

### B. Subchannel Allocation Schemes

The proposed subchannel allocation algorithm and the algorithm that we compared with are presented in this section.

*Cross-layer Based Subchannel Allocation Scheme*

The respective data rate $R_{n,j}$ and priority index $U_{n,j}$ for each flow or user *n* over subchannel *j* is known. The bidding function $\Delta_n(j)$ is computed for each user, which is the difference between the priority index or utility of the best subchannel and that of the second best subchannel of a particular user *n*. The user with the maximum or best bidding function is allocated the subchannel. The $\Delta_n(j)$ is assumed to be the maximum willingness to pay for subchannel *j* of user *i*. This process is repeated until all subchannels have been assigned. The algorithm is presented as follows;

Cross-layer Based Subchannel Allocation (CBSA) Algorithm

1. Initialize S = $\{1,2,3, \dots \dots \dots \dots \dots \dots \dots \dots N_s\}$ and

   $S_n = \phi$ for all selected user *n*
2. While $S \neq \emptyset$
3. For user *n* = 1 to N
4. $\Delta_n(j) = \max_j \{U_{n,j}\} - \max_{i \neq j} \{U_{n,i}\}$

   end
5. $j_n = arg \max_n \Delta_n(j)$
6. $R_{n*} = R_{n*} + R_{n,j}$
7. $S = S - j_{n*}$
8. $S_n = S_n + j_{n*}$

   end

Where $S$ is the set of all subchannels and $S_n$ is the set of subchannels allocated to user *n*. This algorithm will run for $S$ iterations. This means that the number of iterations depends on the number of available subchannels.

*Default Scheme*

The commonly used, default subchannel allocation scheme is actualized by computing the priority index of all users on a subchannel basis and the user with the best priority index

is allocated the subchannel [23]. This is repeated for each subchannel or PRB. Hence, the computation of the priority index or utility function is repeated for every subchannel.

Maximum Utility per Subchannel (MUS) Algorithm

1. For subchannel $j$ = 1 to J
2. $j_k = arg\max_k \{ U_{k,j} \}$
3. $R_{k*} = R_{k*} + R_{k,j}$

end

$J$ is the total number of subchannels and $K$ is the total number of users waiting to be served. It should be noted that $K$ is greater than $N$ that is used in the proposed subchannel allocation algorithm. $N$ is a group of users selected by the scheduler from the total number of users $K$.

## IV. Simulation Model

An event-driven-based open source simulator called LTE-Sim [21] is used for simulations in this paper. It is a standalone version of the LTE module in NS-3 and is written in C++. The simulator has been adapted for the satellite scenario by making necessary changes to its physical layer. A new channel model for satellite which includes shadowing, multipath fading and path loss was added and the propagation delay was modified. The details of the simulator parameters are provided in Table 1.

Table 1 Simulation parameters

| Parameters | Value |
|---|---|
| Simulation Time | 500 seconds |
| RTPD | 540 ms (GEO satellite) |
| Channel Model | 4 state Markov model |
| MIMO | 2 x 2 (2 antenna ports) |
| CQI Reporting Interval | 100 TTI |
| TTI | 1 ms |
| Frequency Re-use | 7 |
| Mobile user Speed | 30 km/h |
| RLC Mode | AM |
| Web Traffic Model | ON/OFF M/Pareto |
| Video Traffic Model | Trace-based @ 440 kbps |
| Scheduler | M-LWDF |
| Subchannel Allocation | CBSA & MUS |
| Bandwidth | 15 MHz |

A single spotbeam has been considered for this simulation in order to evaluate the performance of the proposed subchannel allocation scheme. The UEs are capable of rendering video streaming and web surfing uniformly distributed within the spotbeam footprint. The channel and traffic model presented in previous sections are adopted for the simulations. Each set of UEs is made up of 50% of web browsers and 50% of video streamers. Each UE is assumed to be reporting its channel condition (in terms of CQI) according to fixed intervals (100 TTI) to the eNodeB.

## V. Simulation Results

The simulation results obtained are presented below. The performance metrics considered are throughput, spectral efficiency and fairness.

As shown in Fig. 4, the total throughput for both video and web users for the proposed subchannel allocation scheme CBSA is better than that of MUS for all the number of users considered. This is due to the fact that the proposed algorithm is able to access all the subchannels at once before allocating to the user with more willingness to use the subchannel rather than allocating on a subchannel basis.

The users with best utility functions are first selected hence user with the highest priority, the user which difference between its utility function for a particular subchannel and that of another subchannel is maximum is first assigned subchannels before other users with less difference. This will allow subchannels to be rightly assigned to users with limited options of good subchannels allocation first before users with more options in order to utilize the available resources. It is on this basis that a better throughput is obtained.

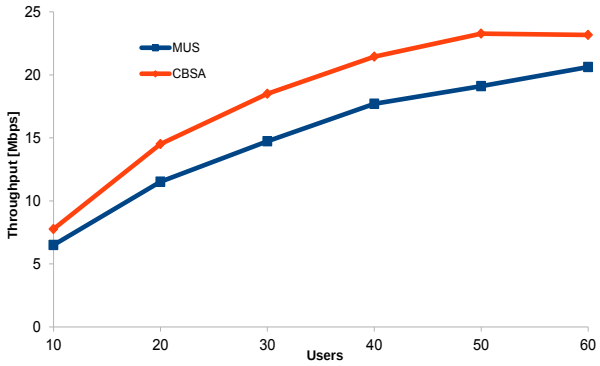



Figure 4 Total Throughput of Video and Web users

The spectral efficiency result shown in Fig. 5 follows the same trend with the throughput result. The proposed subchannel allocation algorithm called CBSA utilizes the spectrum better than the MUS algorithm for all number of users considered.

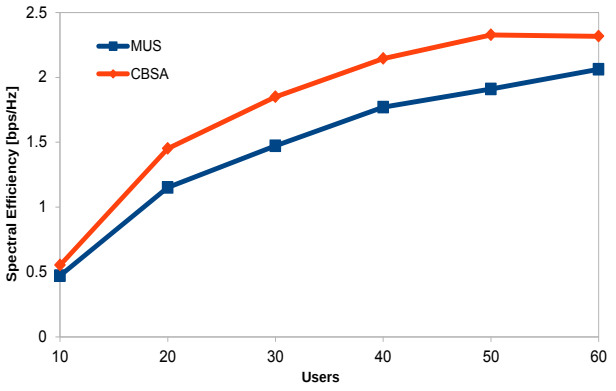



Figure 5 Spectral Efficiency for varying users

As depicted in Fig. 6, the Jain fairness index result shows that both CBSA and the MUS algorithms have a close fairness index performance at both 10 and 20 user, however, at 30 users and above, MUS has a little edge over the proposed CBSA algorithm.

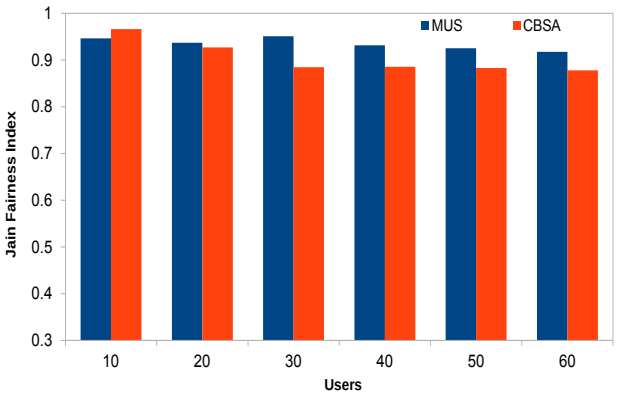



Figure 6 Fairness indexes of all users

## VI. CONCLUSION

This paper presents the proposed CBSA subchannel allocation algorithm in a satellite LTE network and the comparison of subchannel allocation algorithms using performance indices like throughput, spectral efficiency, and fairness index.

The results obtained shows that CBSA subchannel allocation algorithm provides a better throughput and spectral efficiency as compared to MUS algorithm without any serious compromise to the fairness index performance. This shows that the performance of the user scheduling scheme can be further improved by an effective subchannel allocation algorithm.

## REFERENCES


[1] D. Martín-Sacristán, J. F. Monserrat, J. Cabrejas-Peñuelas, D. Calabuig, S. Garrigas, and N. Cardona, "On the Way towards Fourth-Generation Mobile: 3GPP LTE and LTE-Advanced," *EURASIP Journal on Wireless Communications and Networking*, vol. 2009, Article ID 354089, 10 pages, 2009. doi:10.1155/2009/354089.

[2] J. Duplicy, B. Badic, R. Balraj, et al.,"MU-MIMO in LTE Systems," *EURASIP Journal on Wireless Communications and Networking*, vol. 2011, Article ID 496763, 13 pages, 2011. doi:10.1155/2011/496763.

[3] Q.H. Spencer, C.B. Peel, A.L. Swindlehurst, M. Haardt, "An introduction to the multi-user MIMO downlink," *IEEE Communications Magazine*, vol.42, no.10, pp. 60- 67, Oct. 2004 doi: 10.1109/MCOM.2004.1341262.

[4] F. Bastia, C. Bersani, E. A. Candreva, et al., "LTE Adaptation for Mobile Broadband Satellite Networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2009, Article ID 989062, 13 pages, 2009. doi:10.1155/2009/989062.

[5] ESA, "Study of Satellite Role in 4G Mobile Networks Final Report," May 2009.

[6] Choi, Young-June, Kim, Jongtack and Bahk, Saewoong. "Downlink scheduling with fairness and optimal antenna assignment for MIMO cellular systems." *IEEE Global Telecommunications Conference (GLOBECOM), 2004*, Vol. 5, pp. 3165 - 3169.

[7] Sun, Fanglei, et al. "*Multiobjective optimized subchannel allocation for wireless OFDM systems.*" IEEE 20th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), 2009, pp. 1863 - 1867.

[8] M. Torabzadeh, and Yusheng Ji**.** "A Near Optimal Antenna Assignment for MIMO Systems with Low Complexity." *15th IEEE International Conference on Networks (ICON)*, November 2007. pp. 336 - 341. doi: 10.1109/ICON.2007.4444109.

[9] M. Torabzadeh, and Yusheng Ji**.** "*Efficient Assignment of Transmit Antennas for Wireless Communications.*" 2nd IEEE/IFIP International Conference in Central Asia on Internet, Sepember 2006, pp. 1-5. doi: 10.1109/CANET.2006.279256.

[10] Sang-wook Han and Youngnam Han. "A Competitive Fair Subchannel Allocation for OFDMA System Using an Auction Algorithm." *IEEE 66th Vehicular Technology Conference, 2007 (VTC-2007 Fall )*, 2007, pp. 1787 - 1791 . doi: 10.1109/VETECF.2007.377.

[11] S. Parkvall, A. Furuskär, E. Dahlman, "Evolution of LTE toward IMT-advanced," IEEE *Communications Magazine*, vol.49, no.2, pp.84-91, February 2011 doi: 10.1109/MCOM.2011.5706315.

[12] Qinghua Li; Guangjie Li; Wookbong Lee; Moon-il Lee; Mazzarese, D.; Clerckx, B.; Zexian Li; , "MIMO techniques in WiMAX and LTE: a feature overview," *IEEE Communications Magazine*, , vol.48, no.5, pp.86-92, May 2010 doi: 10.1109/MCOM.2010.5458368.

[13] R. Ghaffar, R. Knopp, "Making multiuser MIMO work for LTE,", *2010 IEEE 21st International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC*, pp.625-628, 26-30 Sept. 2010 doi: 10.1109/PIMRC.2010.5671753.

[14] R.T. Schwarz, A. Knopp, B. Lankl, D. Ogermann, C.A. Hofmann, "Optimum-Capacity MIMO Satellite Broadcast System: Conceptual Design for LOS Channels," *Advanced Satellite Mobile Systems, 2008. ASMS 2008. 4th*, vol., no., pp.66-71, 26-28 Aug. 2008 doi: 10.1109/ASMS.2008.19.

[15] Tech. Specif. Group Radio Access Network; Physical Channel and Modulation (Release 8), 3GPP TS 36.211.

[16] Motorola, Inc (2007). Long Term Evolution (LTE): Overview of LTE Air-Interface Technical White Paper Retrieved from http://www.motorola.com.

[17] P.R. King, T.W.C Brown, A. Kyrgiazos, B.G. Evans, "Empirical-Stochastic LMS-MIMO Channel Model Implementation and Validation," *Antennas and Propagation, IEEE Transactions on* , vol.60, no.2, pp.606-614, Feb. 2012 doi: 10.1109/TAP.2011.2173448.

[18] C. Mehlf¨uhrer, M. Wrulich, J. C. Ikuno, D. Bosanska, and M. Rupp, "Simulating the long term evolution physical layer," in Proc. of the 17th European Signal Processing Conf., EUSIPCO, Glasgow, Scotland, 2009.

[19] Video trace library. [Online]. Available: http://trace.eas.asu.edu/.

[20] F. De Angelis, I. Habib, G. Giambene, S. Giannetti, "Scheduling for differentiated traffic types in HSDPA cellular systems," *Global Telecommunications Conference, 2005. GLOBECOM '05. IEEE* , vol.1, no., pp. 5, 28 Nov.- 2 Dec. 2005

[21] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R.Vijayakumar, "Providing Quality of Service over a Shared Wireless Link," IEEE Communications Magazine, vol. 39, pp. 150-154, Feb. 2001.

[22] G. Piro, L.A Grieco, G. Boggia, F. Capozzi, P. Camarda, "Simulating LTE Cellular Systems: An Open-Source Framework," *IEEE Transactions on Vehicular Technology*, vol.60, no.2, pp.498-513, Feb. 2011doi: 10.1109/TVT.2010.2091660.

[23] F. Capozzi, G. Piro, L. A. Grieco, G. Boggia, P. Camarda. "Downlink Packet Scheduling in LTE Cellular Networks: Key Design Issues and a Survey." *IEEE Communications Surveys & Tutorials*, Vol. 15, No. 2, Second Quarter 2013, pp. 678-700. doi: 10.1109/SURV.201.



**Gbolahan Aiyetoro** received the B.Sc. degree in Electronic and Computer Engineering from Lagos State University, Nigeria and MSc.Eng in Electronic Engineering from University of KwaZulu-Natal, Durban, South Africa. He is currently undergoing his PhD at the same University. His research interests is radio resource management in wireless communications.

**Fambirai Takawira** received the B.Sc. degree in Electrical Engineering (first-class honours) from Manchester University, Manchester, UK in 1981, and Ph.D degree from Cambridge University, Cambridge, UK, in 1984. He is currently a Professor of School of Electrical and Information Engineering at the University of Witwatersrand, South Africa.His research interests includes digital communications, data networks and radio resource management in wireless communications.

# CDMA-DCDM for Cognitive Radio Networks

Periola. A.A and Falowo.O.E, *Senior Member IEEE*
Department of Electrical Engineering
Communication Research Group, University of Cape Town, South Africa
email : {periola@crg.ee.uct.ac.za , bisi@ieee.org}

**Abstract - Secondary user cognitive radios are important entities in dynamic spectrum access networks. They are reconfigurable and capable of responding to environmental stimuli in meeting user objectives. This paper proposes the code division multiple access-duty cycle division multiplex (CDMA-DCDM) a hybrid multiplex scheme that leverages on the reconfigurabilty of cognitive radio. In CDMA-DCDM, the cognitive radio secondary user determines own signal duty cycle via interaction with the multiuser detector in the network domain. The functional design of CDMA-DCDM is presented. Additional results that are presented show the throughput comparison of CDMA-DCDM and CDMA for different secondary user transmit power and interference levels. The improvement in spectrum utilization is also investigated. Numerical solutions show that CDMA-DCDM has a improved throughput and spectrum utilization.**

**Key Words: Cognitive radio, Signal duty cycle, CDMA-DCDM, rate per code (RPC).**

## I. INTRODUCTION

The method of static spectrum allocation limits achievable spectrum utilization. Increased subscription to mobile communication services and higher demand for spectrum requires new allocation methods. Cognitive radio networks (incorporating dynamic spectrum access) have been proposed in this regard. Two types of users exist in the cognitive radio network, these are the primary and secondary users. These users are differentiated by their spectrum access priority. Primary users (incumbents) have a higher priority to access the spectrum while secondary users engage in opportunistic use and have a lower spectrum access priority. Secondary user cognitive radios (SUCRs) have a lower priority and are required to transmit without interfering with the primary users.

The cognitive radio network requires new techniques to maximize spectrum utilization and minimize interference to the primary user. Some of the solutions that have been proposed are cross layer optimization [1], sub carrier nulling [2] and deliberate nulling of antenna radiation in certain directions [3] for user satisfaction and interference prevention. Several multiple access techniques such as orthogonal frequency division multiplexing (OFDM) [4], non-contiguous OFDM (NC-OFDM) [5] and selective subcarrier multi-carrier code division multiple access (SS-MC-CDMA) [6] have also been recognized. The secondary user can adapt a limited number of parameters such as the transmit power, modulation scheme and subcarrier allocation to improve the spectrum utilization in these schemes. This paper proposes the code division multiple access- duty cycle division multiplex (CDMA-DCDM) to improve throughput and spectrum utilization

CDMA-DCDM uses the signal duty cycle of the secondary user as the spectrum access differentiator. CDMA-DCDM allows secondary users in a CDMA network to make use of the same code and different signal duty cycle. This paper makes the following contributions:

1. The proposal of user and network domain functions in CDMA-DCDM for connection oriented secondary user session.
2. Investigation of the performance benefit using the data rate (for given multiple access interference and transmit power levels) and spectrum utilization in CDMA-DCDM as metrics.

The paper is organized in the following manner. Section II discusses relevant literature. Section III describes the resource allocation strategy in CDMA-DCDM. Section IV presents user and network domain functions. Section V focuses on derivation of the performance benefits of the proposed scheme. Section VI describes the l simulation procedure and discusses the results. Section VII concludes the work.

## II. RELATED WORK

The research community has proposed different multiple access schemes for cognitive radio networks. These schemes permit the secondary user to vary some operating parameters. OFDM varies the subcarrier allocation, employed modulation scheme and transmit power. NC-OFDM [5] improves upon OFDM and makes use of discontiguous spectrum. SS-MC-CDMA [6] permits greater flexibility allowing two channel usage modes for data transmission. These schemes allow the modulation scheme, transmit power, subcarrier allocation and orthogonal code in [6] to be modified. In [7] Mackenzie *et al.* show that the cognitive radio is capable of finding suitable channel using the detected signal properties.

Hefnawi *et al.* [8] and Cardoso *et al.* [9] propose the space division multiple access (SDMA) and Vandermonde frequency division multiplexing (VFDM) multiplex schemes respectively for the cognitive radio network. SDMA leverages on the existence of a unique spatial signature to differentiate individual secondary users. VFDM uses the redundant cyclic prefix time to transmit the secondary user data. The secondary user transmits using a linear precoder designed using the Vandermonde matrix to transmit over the null space and prevent interference with

the primary user. VFDM is limited in the maximum data rate achievable because the cyclic prefix symbols are small in comparison to the OFDM data symbols.

The duty cycle is an important physical layer parameter that is useful in designing multiple access schemes. It is used in [10, 11, 12] for optical [10] and for wireless communications media [11, 12]. The work in [12] proposes the use of duty cycle division multiplex in wireless communications. Rele *et al.* in [13] use the signal duty cycle to classify transmitters in the WLAN environment to identify an interferer. However while their work shows that duty cycle detection is feasible in the cognitive radio network, it does not propose a multiplexing scheme. To the best of our knowledge this work is the first to suggest the usage of the signal duty cycle property for improving the spectrum utilization in cognitive radio networks.

In CDMA, the user code is the network resource and each secondary user is distinguished by the code. The introduction of duty cycle division multiplex in CDMA-DCDM permits a single code to be used by two users that are differentiated by using different signal duty cycles. This improves spectrum utilization. CDMA-DCDM also improves the throughput because the secondary user cognitive radio is able to transmit using different signal duty cycles on spectrum holes previously inaccessible in CDMA.

## III. CDMA-DCDM SYSTEM DESCRIPTION

This section describes CDMA-DCDM and explains how signal duty cycle awareness is incorporated in the cognitive radio network. The signal duty cycle to CDMA code allocation in CDMA-DCDM is shown in Figure 1.

In CDMA-DCDM the secondary user cognitive radio (SUCR) is aware of the neighboring user's signal duty cycle. This is used to determine the signal duty cycle. Two SUCRs share the same spread code and transmit using different signal duty cycles. The multiuser detector (MUD) is aware of the duty cycle and the spread codes. CDMA-DCDM improves spectrum utilization by permitting multiple users with different duty cycles to make use of the same spread code on the same spectrum hole.

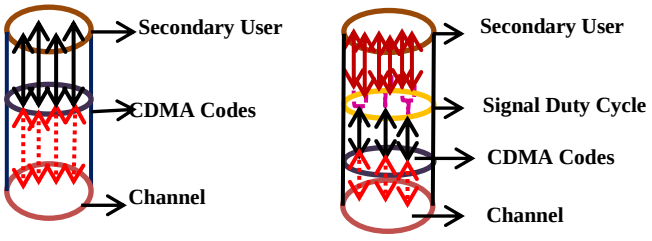



**Figure 1: Signal duty cycle to CDMA Code allocation in CDMA-DCDM.**

If the set of CDMA codes and two duty cycles is given as $[C_1, C_2, \dots, C_m]$ and $[d_1, d_2]$ respectively. This combination gives the set $U = [C_1 d_1 C_1 d_2, C_2 d_1 C_2 d_2, \dots, C_m d_1 C_m d_2]$.

CDMA-DCDM permits the same code to be shared by two users with different signal duty cycles. The proposed technique takes advantage of modulation symbols being sinusoidal pulses of duration $T_s$ defined at time instant *t*. The M-ary PSK $g_k^{psk}(t)$and QAM $g_k^{qam}(t)$ symbols for order *M* neglecting the angular offset is given as [14]:

$$g_k^{psk}(t) = rect\left(\frac{t-\frac{1}{2}T_s}{T_s}\right)\left(d\sqrt{\frac{2}{T_s}\cos\left(2\pi f_c + + \frac{2\pi k}{M}\right)}\right) \quad (1)$$

$$g_k^{qam}(t) = rect\left(\frac{t-\frac{1}{2}T_s}{T_s}\right)\left(D_k\sqrt{\frac{2}{T_s}\cos(2\pi f_c)}\right) \quad (2)$$

$T_s$ is varied by the duty cycle adaptation in CDMA-DCDM. The SUCR receives information on the active duty cycles for a sensed channel over the common control channel (CCCH) from the multi-user detector entity (MUD) on the secondary base station (SBS) in the network domain. A sensing SUCR ascertains the suitability of a spectrum hole using information accessed from the MUD. The spectrum usage profile (SUP) in CDMA-DCDM for three duty cycles $d_1$,$d_2$ and $d_3$ is

$$SUP = \begin{pmatrix} S_1 d_1 & S_1 d_2 & S_1 d_3 \\ S_2 d_1 & S_2 d_2 & S_2 d_3 \\ S_3 d_1 & S_3 d_2 & S_3 d_3 \end{pmatrix} \quad (3)$$

$S_1 d_1$ shows the use of duty cycle $d_1$ on first spectrum hole $S_1$. $S_3 d_2$ shows the use of duty cycle $d_2$ on third spectrum hole $S_3$. The other elements in (3) have similar interpretation.

The row elements in (3) show the usage of different duty cycles on a given spectrum hole (channel). The desirable spectrum hole $SH^{des}$ is:

$$SH^{des} = \max(C(d_n)) \quad (4)$$

Where C (.) is the cardinality and is the number of duty cycles that secondary users intending to transmit over the spectrum hole can use. A spectrum hole is suitable if C(.) > 0. A network scenario of CDMA-DCDM showing the interaction of two SUCRs, the MUD entity and the SBS is shown in Figure 2. The achieved rate and transmit power can be said to be attached to the code. Hence the user rate per code is taken as a metric in this work.

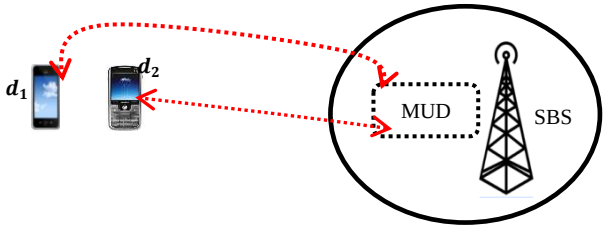



**Figure 2: Coordination of spectrum hole use involving users with different duty cycles.**

CDMA-DCDM achieves a higher rate per code than traditional CDMA assuming ideal multi user detection. The rate $R_{d_1}, R_{d_2}$ achieved by two users assigned duty cycles $d_1$ and $d_2$ for a given noise power $\sigma_n^2$, intra-cell interference $\sigma_{ICI}^2$ and multiple access interference $\sigma_{MAI}^2$ is obtained using the Shannon relation is:

$$R_{d_1} = \log_2\left(1 + \frac{P_{d_1}|h_d|^2}{\sigma_n^2 + \sigma_{ICI}^2 + \sigma_{MAI}^2}\right) \quad (5)$$

$$R_{d_2} = \log_2\left(1 + \frac{P_{d_2}|h_d|^2}{\sigma_n^2 + \sigma_{ICI}^2 + \sigma_{MAI}^2}\right) \quad (6)$$

$P_{d_1}$ is the secondary user transmitting power at duty cycle $d_1$. $P_{d_2}$ is the secondary user transmit power at duty cycle $d_2$. $|h_d|^2$ is the channel gain.

The rate per code (RPC) metric is used in this paper. CDMA differentiates users on the basis of used codes. Hence user throughput can be linked to the code used in transmission. The total SUCR rate per code in CDMA-DCDM is:

$$R_T^{d'} = R_{d_1} + R_{d_2} \quad (7)$$

CDMA rates for a single code given transmit power P and channel gain $|h|^2$ using Shannon relation is:

$$R = \log_2\left(1 + \frac{P|h|^2}{\sigma_n^2 + \sigma_{ICI}^2}\right) \quad (8)$$

Both CDMA and CDMA-DCDM suffer from intra-cellular interference. The multiuser detection functionality is introduced in CDMA-DCDM to obtain a minimal $\sigma_{MAI}^2$. The sampling rules provided in [12] and the use of multiuser detection schemes based on neural network algorithm such as those described in [15] enable the minimization of $\sigma_{MAI}^2$. Hence :

$$\sigma_{MAI}^2 \ll \sigma_n^2 + \sigma_{ICI}^2$$
$$\sigma_n^2 + \sigma_{ICI}^2 \approx \sigma_n^2 + \sigma_{ICI}^2 + \sigma_{MAI}^2 \quad (9)$$

Therefore $R_T^{d'} > R$ and it is concluded that the rate per code is increased in CDMA-DCDM.

## IV. CDMA-DCDM: SYSTEM DESIGN

This section discusses the user and network domain functions for the secondary user in CDMA-DCDM.

### A. *User Domain*

The user here refers to the transceiver entity. The transmitting secondary user cognitive radio (SUCR) performs spectrum sensing and secures the use of a spectrum hole. During link acquisition the SUCR interacts with the MUD entity on the SBS over the common control channel (CCCH). The MUD assigns a duty cycle to the SUCR for line encoding. This is followed by signal modulation and spreading for the actual transmission. The receiving entity also receives the duty cycle used in the line encoding stage from the SBS and is then able to demodulate and decode the signal.

### B. *Network Domain*

The multiuser detector entity (MUD) is located on the secondary base station (SBS) in the network domain. In the spectrum sensing phase prior to the commencement of data transmission in the network. The MUD allots signal duty cycle to incoming secondary users via (CCCH). It also holds the values of the duty cycles of present users. This takes place in both single hop and multihop networks.

Two duty cycles represented by binary values (1 and 0) are used. During data transmission in a single hop network the receiver obtains the duty cycle used via the MUD of the secondary base station.

The case of the multihop network is addressed using cross layer design technique shown in Figure 3. In establishing a TCP session for example, a source engages in a three way handshake with the intended destination before data transmission. The binary value describing the duty cycle is passed to the TCP layer via the physical layer of an incoming SUCR via interaction with the MUD on the CCCH (step 1). This is sent to the intended destination's TCP layer (step 2) and subsequently to the physical layer of the intended TCP destination for line decoding (step 3). This cross layer design is suitable for data transmission because TCP is widely used on the internet.

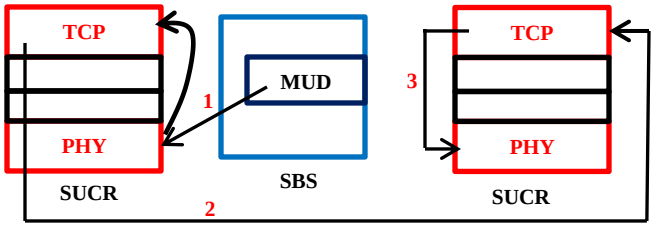



**Figure 3: Cross layer duty cycle exchange in CDMA-DCDM.**

## V. PERFORMANCE EVALUATION

In CDMA-DCDM, it is important that user information signals have distinct patterns during transmission. The reasoning here is that the presence of these unique patterns at different duty cycles facilitate signal reception. In this paper, the 16 QAM modulation scheme is considered to be employed by both users, the investigation is performed for different EbNo values of 2dB and 6dB. The RZ line encoding scheme is used. The duty cycle values are 50% and 75%. Figures 4 and 5 show the distinct pattern in the user signals at EbNo values of 2dB and 6dB respectively. A high EbNo value of 6dB has a small degree of overlap and a low EbNo value of 2dB has a higher degree of overlap as seen in Figures 4 and 5. This distinctiveness is important for the receiver design.

Figures 4 and 5 show a plot of 16 QAM modulated signals for a signal duty cycle values of 50% and 75% in a CDMA-DCDM system at an EbNo of 2dB and 6dB respectively. Signal points for the 50% duty cycle at the above specifications are shown in blue and 75% duty cycle signal points are in green. The plots in Figures 4 and 5 demonstrate that these signals are distinct from each other and can be distinguished at the receiver.

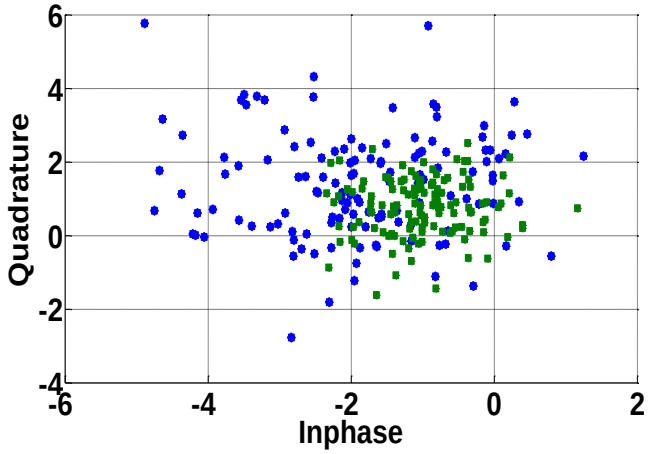



**Figure 4: Distinct User Patterns at EbNo of 2dB**

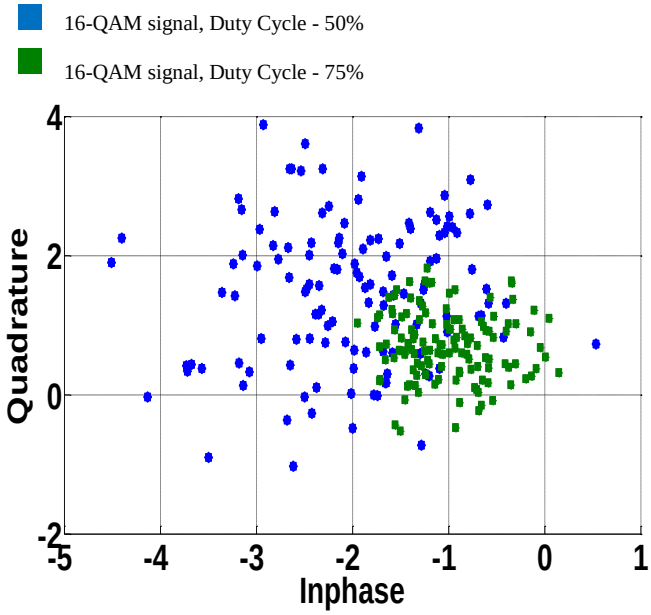



**Figure 5: Distinct User Patterns at EbNo of 6 dB**

The performance evaluation of the proposed scheme aims to investigate the user rate per code using (5) - (9). The investigation is done using different performance levels of the multi user detector in reducing the multiple access interference. The ratio of the multiple access interference $\sigma_{MAI}^2$ to the sum of other interference components $\sigma_n^2 + \sigma_{ICI}^2$ is used in the simulation. $P_T^d$ is the overall transmit power of two secondary users in a CDMA-DCDM network and P is the transmit power of a single user in the CDMA network. The evaluation is done using different values for the ratios $P_T^d/P$ and $\sigma_{MAI}^2/(\sigma_n^2 + \sigma_{ICI}^2)$.The improvement in spectrum utilization (mean channel occupancy to system bandwidth) is also investigated for a five channel scenario and is derived to be [16]:

$$\eta_1 = \pi_0 \frac{0.8\mu_s\lambda_s\lambda_p + 8\lambda_s^2\mu_p + 3.2\mu_s^2\mu_p + 3\mu_s\mu_p\lambda_p}{\mu_s\mu_p\lambda_p} \quad (10)$$

$$\eta_2 = \pi_0 \left(0.8 + 8\frac{\lambda_s}{\lambda_{d1}-\lambda_s}\frac{\mu_p-\lambda_{d2}}{\lambda_p} + 3.2\frac{(\lambda_s-\mu_s)(\mu_p-\lambda_{d2})}{\lambda_p\lambda_{d1}} + 3\frac{\mu_p-\lambda_{d2}}{\lambda_p}\right) \quad (11)$$

$\lambda_s, \mu_s$ are the secondary user arrival and departure rates. $\eta_1, \eta_2$ are the secondary user utilization in the case of CDMA and CDMA-DCDM networks respectively. $\lambda_{d1}$ and $\lambda_{d2}$ are the arrival rates of secondary user radios with duty cycles $d_1$ and $d_2$ respectively. $\pi_0$ is the initial state probability.

## VI. PERFORMANCE SIMULATION AND DISCUSSION

This section presents and discusses the performance results that aim to demonstrate the improvement in spectrum utilization and rate per code ratio when CDMA-DCDM is used. The rate per code metric is the achieved throughput when the user is assigned a code. In CDMA, the user is assigned a code and achieves a rate R bits per second. In CDMA-DCDM, two users can be assigned a code and transmit at two different duty cycles. In this case the spectrum utilization is improved and the total rate of the two users is $R_T^d$ bits per second.

The ratio $R_T^d/R$ in Figure 6 is the ratio of the throughput obtained in CDMA-DCDM to that obtained in CDMA. If this ratio exceeds unity then the secondary user throughput in CDMA-DCDM is improved over that of CDMA. In this case CDMA-DCDM has the advantage of improving both the spectrum utilization and the throughput. When $R_T^d/R$ less than unity, CDMA -DCDM is improves the spectrum utilization.

The multiple access interference $\sigma_{MAI}^2$ is related to the MUD capacity. The neural network algorithm can be trained to achieve good multiuser detection as seen in [15]. Hence $\sigma_{MAI}^2$ is sufficiently low and the projected performance in region A is therefore realistic.

The plot in Figure 6 is divided into four regions. The region A has the best performance as the multiple access interference is lowest. CDMA-DCDM achieves a higher throughput than traditional CDMA because of the higher rate per code (RPC) ratio.

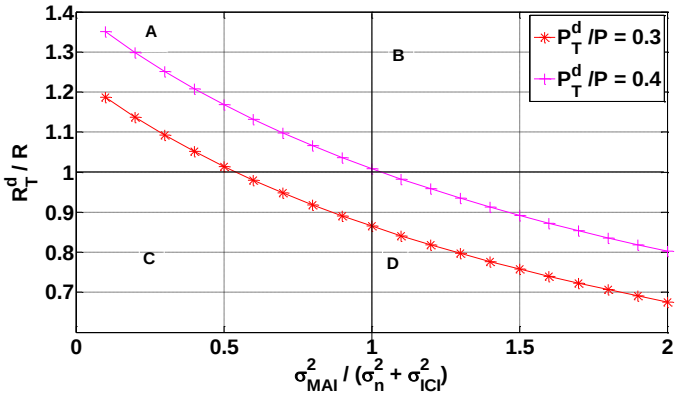



**Figure 6: Rate per code performance of CDMA-DCDM and CDMA.**

In region B, the multiple access interference is quite significant and a higher transmit power is required for any desirable improvement to be obtained in this region.

Region C shows the performance when the two variable signal duty cycle users in CDMA-DCDM are power

limited. $R_T^d/R$ falls short of unity and the performance can be improved by increasing transmit power. Region D shows the performance when the multiple access interference becomes significant. In region D, $R_T^d/R$ is considerably reduced because of higher $\sigma_{MAI}^2$.

The spectrum utilization is given as the ratio of the mean channel occupancy to the system bandwidth [16]. Figure 7 show that using CDMA –DCDM improves secondary user cognitive radios channel occupancy duration. This is because of the duty cycle flexibility introduced in CDMA-DCDM. This is also good for delay sensitive applications where the secondary user is required to respond to received messages in a given time period.

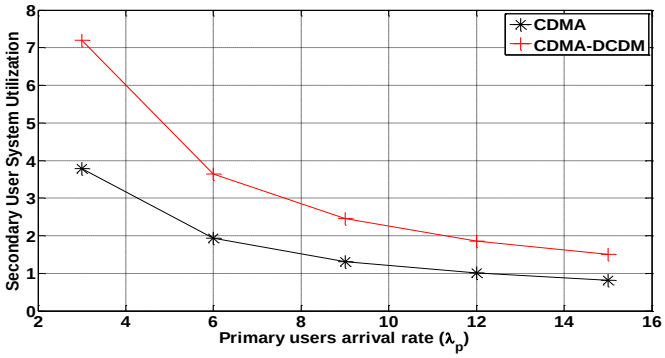



**Figure 7 : Improvement in secondary spectrum utilization when CDMA-DCDM is applied.**

The advantage of using CDMA-DCDM is that delay sensitive applications users that should respond within a given time have continuous spectrum access. Higher data rates are obtained by assigning multiple codes to a user [10] transmitting at a given signal duty cycle. These codes are shared by other users (desiring high data rates) who transmit at different signal duty cycles.

## VI. CONCLUSION

This work introduces the code division multiple access duty cycle division multiplex (CDMA-DCDM) for cognitive radio networks. CDMA-DCDM improves spectrum utilization by allowing two users in a CDMA network to share the same code by using different signal duty cycles. The secondary user cognitive radio acquires the signal duty cycle via interaction with a multiuser detector and communicates this information using cross layer design approach. Performance results show that CDMA-DCDM is suitable for delay sensitive applications. It also improves network capacity and spectrum utilization. The CDMA-DCDM receiver is under design and is the subject of future work.

## ACKNOWLEDGEMENTS

The indebtedness to the anonymous reviewers in the improvement of this paper is highly acknowledged and much appreciated. The financial support of the National Research Foundation (NRF) South Africa for this research is highly appreciated and acknowledged.

## REFERENCES

[1] M.V Nguyen, C.S.Hong and S.Lee '*Cross Layer Optimization for Congestion and Power Control in OFDM-Based Multihop Cognitive Radio Networks* ' IEEE Transactions on Communications , August 2012, Vol 60 No 8, pg 2101-2112.

[2] V.L.Nir and B.Scheers '*Cognitive Radio Systems: State of the Art*'[Online]ww.sic.ma.be/~vlenir/publications 2009

[3] J.Liu, W.Chen, Z.Cao and Y.Jun '*Cooperative Beanforming for Cognitive Radio Networks: A Cross Layer Design'* IEEE Transactions on Communications May 2012, Vol 60 No 5, pg 1420-1431.

[4] L.Haring *'Automatic Modulation Classification for Adaptive Wireless OFDM Systems'*(Ed) Ali Eksim. [Online] http://www.intechopen.com/books/wireless-communications-and-networks-recent-advances/automatic-modulation classification-for-adaptive-wireless-ofdm-systems,2012.

[5] R.Kapoor and P.Kumar *'Spectral Agility of NC-OFDM Transmission for Dynamic Spectrum Access'* International Conference on Computers and Devices for Communications, Kolkata, 17-19 Dec 2012, pp 1-4.

*[6]* D.Sarath, K.E Nolan, P.D.Sutton and Doyle.L.E *'Enabling Dynamic Spectrum Access using SS-MC-CDMA'* 1-3 August Orlando USA pp 193-198, 2007.

[7]A.B.MacKenzie,J.H.Reed,P.Athanas,C.W.Bostian,B.R,Michael,L.A.Da Silva,S.W.Ellingson,Y.T.Hou,M.Hsiao,J.Park,C.Patterson, Raman and C.DaSilva, *'Cognitive Radio and Networking Research at Virginia Tech'* IEEE Proceedings April 2009, Vol 97 No 4 pp 660-688.

[8] M.Hefnawi *'Space Division Multiplexing Access Aided Cognitive Radio Networks'* Biennial Symposium on Communications (QBSC), Kingston, ON 28-29 May 2012, pp 10-14.

[9] L.S.Cardoso, M.Kobayashi, F.Rodrigo,P.Cavalcanti and M.Debbah *'Vandermonde-subspace Frequency Division Multiplexing for Two Tiered Cognitive Radio Networks'* IEEE Transactions on Communications, June 2013, Vol 61 No 6, pp 2212-2220

[10] M.K.Abdullah, G.A Mahdiraji, A.M.Mohammadi, M. Mokhtar and A.F Abas *'Duty Cycle Division Multiplexing (DCDM): A New Electrical Multiplexing Technique for High Speed Optical Communication Systems'* National Conference on Telecommunication on Photonics and Telecommunication Technologies, Putrajaya, 26-28 Aug 2008, pp 71-74

[11] M.N Derahnam, K.Dimyati, A.M Mohammadi and M.K Abdullah *'Improvement of Decision Making Protocol for Duty Cycle Division Multiplexing Scheme (DCDM) system'* International Conference on Future Networks Sanya Hainan, 22-24 Jan 2010, pp 155-158.

[12] A. Malkmohammadi, M.K.Abdullah, A.F.Abas, G.A Mahdiraji and M.Mokhtar *'Absolute Polar Duty Cycle Division Multiplexing (APDCDM) Technique for Wireless Communications '* International Conference on Computer and Communication Engineering Kuala Lumpur ,13-15 May 2008, pp 617-620.

[13] K.Rele, D.Robertson, B.Zhang, L. L, Y.B.Yap, T.Taher, D.Ucci, and K.Zdunek *'A Two Tiered Cognitive Radio System for Interference Identification in 2.4GHz Band'* IEEE Consumer Communications and Networking Conference, 9-12 Jan 2010, Las Vegas, NV, pp 1-5.

[14] I.Otung, *'Short Course in Digital Communications'*, 2009, Pontypridd

[15] N.Taspinar and M.Cicek *'Neural Network Based Receiver for Multiuser Detection in MC-CDMA Systems'* Journal of Wireless Personal Communications Vol 68 No 2, pp 463-472, 2013.

[16] L.Li, S.Zhang, K.Wang and W.Zhou *'Queuing Method in Combined Channel Aggregation and Fragmentation Strategy for Dynamic Spectrum Access'* IEEE Personal Indoor and Mobile Radio Communications, Sydney NSW, 9-12 Sept 2012, pp 1214-1219.

## Biography

Periola Ayodele studied Electrical Engineering at the Bachelor (B.Eng.) and master degree (M.Sc.). He is currently a PhD candidate at the communications research group, University of Cape Town. His research interests are in multiplexing techniques for cognitive radios, using neural networks for prediction and decision making in cognitive radio networks.

# Design of a Cognitive Small Cell Backhaul System for Non-Line-of-Sight Deployment in Urban Canyons

Bessie Malila, Olabisi Falowo and Neco Ventura
Department of Electrical Engineering
University of Cape Town, Menzies Building Upper Campus, Rondebosh, Cape Town 7700
Tel: +27 21 650 2813, Fax: +27 21 6502465
Email: {bmalila,bisi,neco }@crg.ee.uct.ac.za

**Abstract-The unprecedented growth in network traffic accompanied by declining revenues has forced mobile network operators to look for cost effective ways to increase network capacity and coverage. In urban areas, small cell access networks deployed at street level are expected to address the capacity and coverage problem. This deployment strategy has created the need for new backhaul systems since the traditional systems based on fiber and copper are too costly to deploy, while broadband microwave solutions require clear line-of-sight (LoS), which is difficult to achieve at street level. In this paper we propose a new application of Cognitive Radio Technology that will enable non-line-of-sight (NLOS) wireless small cell backhaul communications in urban canyons. The solution uses millimeter wave technology to meet the capacity requirement and cognitive radio technology to address the coverage and hence LoS problem. In the solution, backhaul radio devices intelligently use diffracted signals to establish communication links under NLOS conditions.**

**Index Terms— Small Cells Backhaul, Cognitive Radio Technology, Millimeter Wave Technology**

## I. INTRODUCTION AND MOTIVATION

Network operators are currently faced by unprecedented growth in network traffic. This is attributed to the availability of low cost smart phones, tablet pc's and increase in the number of connected users [1]. The average traffic per user or device is also increasing exponentially with mobile video traffic having contributed up to 51% of network traffic by end of 2012 [1].The high penetration rate of wireless networks compared to wired networks has also driven the increase in the number connected devices and hence network traffic.

Figure 1 illustrates variation of traffic, costs and revenue with time as networks evolved from voice centric to data centric networks. The revenue gap is expected to widen with the evolution of mobile networks to 5G networks and the emergence of new technologies like machine type communications, cloud services and heterogeneous access networks. Communication networks have become critical in driving economies by providing connectivity to power, health and transport infrastructures. This will further put pressure on network operators to expand networks, leading to further increase in infrastructure and energy cost.

Network operators are therefore looking at ways in which they can increase network capacity and coverage in a cost effective way. Providing adequate network coverage also enhances user Quality of Experience (QoE) resulting in reduced customer defection and hence stable revenue streams. Some of the strategies to increase access network capacity have involved increasing the number of existing base stations, performance enhancements on the air interface and offloading traffic to WiFi networks. One solution that has however gained industry acceptance is the dense deployment of base stations with smaller coverage radius, called small cells. Unlike traditional macro-cell base stations installed at roof tops, small cell deployment strategy addresses the coverage problem by bringing the base station in closer proximity of the customer at street level. The resultant short transmission distance between the customer and the base station also translates to increased data transmission rate and improved user QoE. Dense deployment of small cell base stations in urban areas will however result in the need for cost effective, scalable and flexible small cell mobile backhaul systems. Any backhaul solution must also meet the capacity and coverage requirements of small cell base stations.

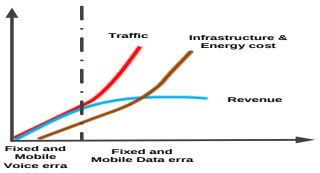


Figure 1 Costs and traffic increase vs declining revenues

Traditional mobile backhaul solutions include fiber, microwave, satellite and Digital Subscriber Line (DSL). Although fiber provides the high capacity and low latency expected of future mobile backhaul solutions costs associated with new fiber installation will be prohibitive since most small cells will be installed on new locations [4]. DSL has capacity limitations and also presents high cost of deployment where new installations are required. Microwave solutions provide cheaper backhaul access at roof top level due to the line-of-sight (LoS) requirement in the traditional 6-23GHz carrier frequency range; small cell deployment at street level makes this solution unsuitable. Furthermore, channel capacity of traditional microwave systems is limited to 2Mbps, which is far from the expected gigabit per second capacity of scalable SCB solutions.

Wireless Systems operating between 30 and 90GHz, commonly referred to as millimeter wave (mmW) technologies, have been identified as having the capability to meet the cost and capacity requirements for SCB systems.

However, street level deployment of systems operating at these frequencies also requires clear LoS for proper operation. A new innovative way to address the LoS problem is therefore required if street level deployment of mmW technology is to be fully realized. Cognitive Radio Technology (CRT) is viewed as an evolution of Software Defined Radio (SDR) towards a fully autonomic reconfigurable wireless transceiver which fully adapts its communication parameters to network conditions and user demands. Various applications of CRT have emerged [2], the most dominant being dynamic spectrum access and spectrum sharing. In these applications, CRT adds intelligence to radio devices, giving them the ability to autonomously interact with the wireless and network environment to achieve predefined network goals.

This paper investigates the possibility of using CRT to address the LoS problem in SCB deployment in urban environments. The technology can also be used to manage the complexity associated with densely deployed wireless networks by adding intelligence to the SCB radio devices. The system is expected to be deployed and operated in urban environments, specifically urban canyons, under non-line-of-sight (NLOS) conditions. The CSCB system will be based on technologies that are driving the evolution of 4G networks to 5G, i.e. mmW, CR and SDR [3]. In the solution we show that by harnessing these technologies, it is possible to develop a cost effective, flexible and scalable mobile backhaul solution that can be densely deployed in urban canyons to provide backhaul to small cell access networks. The rest of the paper is organized as follow: section II gives an overview of the small cells, key SCB backhaul requirements and technologies enabling development of the proposed solution. Section III reviews standardization work and related literature. Section IV discusses the proposed solution and section V concludes the paper and discusses future work to be done in order to implement the solution.

## II. Small Cells And Backhaul Networks

### A. Definition of Small Cells

According to the Small Cell Forum (SCF), small cells are defined as "high capacity, low-power and low cost access points that operate in the licensed spectrum and are managed by network operators [4]. They include various technologies known as Femto, Pico, Micro and Metro cells". Small cells are part of the 3GPP Evolved Packet System (EPS) heterogeneous radio access network technologies (HetNets) designed to improve capacity and coverage in homes, enterprises, public places, metropolitan and rural areas. Figure 2 illustrates the end-to-end architecture of a typical mobile system. Only small cells and macro cells are shown for clarity. The SCB can connect directly into the core network or via macro base station aggregation points at roof tops. The SCB must have the flexibility to provide backhaul to other access networks like Wi-Fi and WiMax.

Femto and Pico cells are designed to provide indoor coverage; micro and metro cells on the other hand are designed for outdoor coverage of a smaller area compared to a macro cell. This paper focuses on the outdoor backhaul for micro and metro cells. Outdoor small cells have a coverage

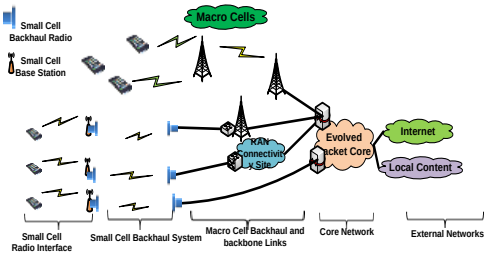


Figure 2: End-to-end System Architecture

radius of 50-500m and metro cells can provide coverage up to 2000m. In contrast, macro cells provide coverage for up to 30km and are mounted on masts or roof tops. Small cells will be deployed at street level approximately 3-6m above ground on street poles and building sides.

### B. Small Cells Backhaul Requirements

The major requirements of SCB have been specified by the SCF and the Next Generation Mobile Network (NGMN) alliance [4], [5] as coverage, capacity, availability and synchronization. Backhaul coverage refers to locations where backhaul connectivity of the required quality must be provided to small cell access points. The peak and average throughput of a SCB must be more than that of the small cell access point it is serving and defined as 150Mbps and 50Mbps respectively. For scalability, the SCB must provide for Gigabit per second (Gbps) capacity to cater for improved capabilities of new user devices and future small cell capacities. Network synchronization traffic, like real time video applications is sensitive to delay and delay variation (jitter). The SCB must therefore comply with the maximum specified delay for the IP backhaul segment of 60ms.

The SCF also specifies that the SCB must support QoS and traffic classification. The 3GPP specifies nine Quality Class Identifiers, the SCB is however expected to support four QoS classes i.e. Class 1: real time e.g. voice, real-time gaming and control traffic, Class 2: 2G data and real-time video, Class 3: buffered video and non-real-time GBR, Class 4: rest of traffic and best effort. Finally, a SCB must be self-organizing to reduce installation costs. The self-organizing requirement is in line with the concept of self-organizing capability expected of future mobile networks [6].

### C. SCB Enabling Technologies

This section gives an overview of mmW and CRT technologies and how they will enable the development of the CSCB system.

#### 1) Millimeter Wave Technology

The capacity of a wireless link varies directly with the system bandwidth. Wireless systems operating at mmW frequencies have therefore gained popularity due to the inherently wide bands which do not need any sophisticated techniques to achieve Gbps capacity. Applications of the technology have however been limited mainly due to costs previously associated with radio devices operating at these frequencies. The availability of low-cost semiconductor technologies like Gallium Arsenide (GaAs) and Silicon Germanium (SiGe) has resulted in reduced cost of the devices. High signal attenuation caused by atmospheric

absorption due to the small signal wavelength has also limited use in mobile networks [7]. Figure 3 illustrates signal attenuation of frequency bands up to 400_GHz. The discussion in this paper is limited to the 30-100_GHz spectrum range. The white circle represents sub-6_GHz spectrum currently used for access networks and point-to-multipoint backhaul systems. It also covers the traditional 6-23_GHz microwave backhaul spectrum. The blue circles represent spectrum suitable for short indoor links due to high attenuation. The 38_GHz, 42_GHz, 54-66_GHz and 70-90_GHz bands are earmarked for high capacity mobile access and backhaul systems.

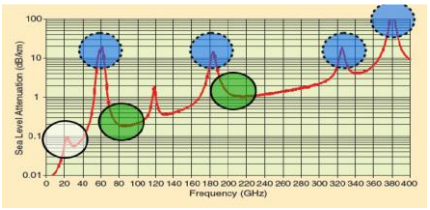



Figure 3 Atmospheric attenuation at mmW frequencies [7]

The high attenuation of mmW frequencies requires use of high-gain antennas. Such antennas have pencil-like beams which require careful antenna alignment for link setup. This adds a new dimension in the requirements of mmW-based backhaul solutions, i.e. automatic antenna steering, for reduced installation time and hence costs in densely deployed networks. The high attenuation also results in short transmission distances. This makes the technology suitable for small cell backhaul deployment which will be characterized by short hop links of 50-2000m. The short transmission distances will also result in increased spectral efficiency and reduced interference. The licensing regimes of the 60_GHz (unlicensed) and 80/90_GHz bands (light-licensed) have also added to the interest in the spectrum since this contributes to reduced operational costs.

*2) Cognitive Radio Technology*

The concept of Cognitive Radio (CR) was coined by Joseph Mitola in his ground breaking dissertation [8]. According to [9], a Cognitive Radio is "an intelligent wireless communication system that is aware of its surrounding environment (i.e., outside world), and uses the methodology of understanding-by-building to learn from the environment and adapt its internal states to statistical variations in the incoming RF stimuli by making corresponding changes in certain operating parameters (e.g. transmit-power, carrier-frequency, and modulation scheme) in real-time, with two primary objectives in mind: (i) highly reliable communications whenever and wherever needed; (ii) efficient utilization of the radio spectrum". The technology therefore adds intelligence to radio devices, giving them the capabilities to interact with the external environment through sensing, learning and making the required decisions, through the cognitive cycle.

Several measurement campaigns have proved the existence of usable high-capacity links based on mmW frequency bands under non-line-of-sight conditions in highly built-up urban environments. To fully operationalize such solutions, CRT can be used to give radio devices the intelligence required for device autonomy and reconfigurability.

Self-Organizing Network (SON) technology, with concepts closely related to Cognitive Wireless Networks, is being proposed for use in mobile access and core networks to manage network complexity. The SCB is no exception to this complexity. Furthermore, deployment of backhaul networks in urban canyons increases complexity due to the need for line-of-sight for systems based on mmW spectrum. CRT can therefore be used to introduce aspects of SON on the SCB segment, creating synergies with the access and core networks. The following section outlines the similarities and differences between Cognitive Networks (CN) and Cognitive Radio Systems (CRS), and the CSCB. The aim is to give a clear understanding of the CSCB system as it relates to existing networks based on CRT.

### *D. CSCB as a Cognitive Radio System*

CRT has been applied to CN and CRS. A CRS uses CRT to gather geographical and radio system parameters to achieve performance goals of a system of interconnected radio nodes. On the other hand, a CN uses CRT to achieve end-to-end network performance goals. Unlike the CRS, the network nodes of a CN include routers, switches servers etc participate in the cognitive process.

CRS and CN have end-to-end goals of improving network performance, while a CR acts autonomously to achieve its own goals, e.g., deciding output power, modulation techniques, operating frequency access network. CR, CN and CRS use cognitive technology to achieve their objectives. The cognitive cycle in both cases includes obtaining knowledge, making decisions and learning from previous experience. It can therefore be concluded that any system based on CRT, like the CSCB system, should be characterized by having some form of intelligence or cognisance and must also achieve its objectives by following the cognitive cycle. The following section outlines standardization work and literature related to SCB solutions.

## III. RELATED WORK

### *A. Standardization Work*

The 3GPP standardization for mobile networks focuses on the access and core networks but does not address the small cell backhaul segment. To date, the International Telecommunications Union (ITU), which is responsible for backhaul and backbone standardization has not yet developed standards for the SCB. The IEEE has however set up a working group to specify amendments to the IEEE802.16-2012 standard [20] to include specifications for SCB solutions. This standard is based on sub-6_GHz spectrum, which has capacity limitation and is also reserved for access networks. The SCF and NGMN alliance have published performance requirements and recommended deployment strategies for small SCB solutions, some of which have been outlined in section II above. The proposed specifications and recommendations highlight the need for a high-capacity, self-optimizing, self-configuring, scalable and flexible SCB. The recommendations do not however fully address NLOS deployment of the SCB.

### B. *Literature Review*

Since little is known about propagation models for mmW systems in urban environments, recent research work on SCB solutions has focused on channel characterization of 28-90_GHz spectrum in urban environments [7], [10-12], [16]. These measurement campaigns have shown that it is possible to create usable links under NLOS conditions at mmW frequencies based on diffracted signals. Use of automatic antenna steering technology for signal searching has been recommended. This information is the basis for the use of diffracted signals in the proposed solution. The work is however, limited to propagation studies; no mention is made of how the complexities associated with dense deployment of such systems can be managed, creating an opportunity for further research.

A number of researchers have proposed a number of SCB solutions. In [13], the researchers propose the use of existing fiber to augment mmW backhaul links to improve coverage in urban environments. This solution is characterized by use of short hop links to minimize co-channel interference, use of a Bandwidth Resource Manager (BRM) to reroute traffic between the fiber and wireless networks according to traffic volumes. However, the solution assumes line-of-sight operation and the issue of antenna alignment, which is critical for mmW radios, is not addressed. In [14], the authors propose a point-to-multipoint architecture for a SCB which is designed to address the cost, capacity and coverage SCB requirements. However the solution also assumes line-of-sight operation and device intelligence is not addressed. In [15], the NLOS problem is addressed creating communication paths based on a knife edge diffraction model in a CN. The CN provides connectivity to small cells not covered by the primary backhaul network. Limitations of the solution include use of the low-capacity 10-20_GHz spectrum, reliance on a single communication path and lack of clarification on autonomy and reconfigurability of the radio devices. Since small cell deployment is already underway, a number of vendors have noticed the SCB technology gap and are coming up with their own solutions [19]. These solutions are however, proprietary and do not fully address the requirements of the SCB outline by the SCF and the NGMN alliance.

## IV. IMPLEMENTATION

In the absence of standardization for SCB solutions, the CSCB solution is based on 5G mobile transport network requirements defined for HetNets by the 3GPP and recommendations for small SCB solutions outlined by the SCF and the NGMN alliance. Outputs of the ETSI Future Mobile summit [17] are taken into consideration. In order to synthesize the concepts and technologies discussed in the previous sections, this section investigates how the proposed CSCB system can be implemented. A framework for link establishment is constructed and critical features of the solution are identified.

Figure 4 illustrates the development of multiple signal paths in a typical urban environment. Point-to-point radio links are created by diffraction and reflection of radio signals at the edges of building sides and roof tops and metallic objects in the environment, e.g., steel and reinforced

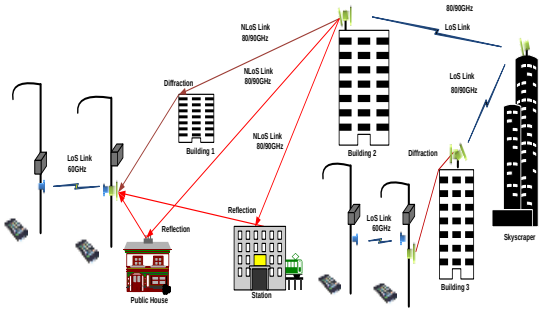



Figure 4: NLOS SCB deployment architecture for small cells.

concrete surfaces respectively. For diffraction, the single knife edge model for urban environments is used to estimate computation of the diffraction process [22]. Wherever possible, direct LoS links are used. As explained above, use of 60_GHz spectrum is limited to 500m links at street level due to the higher oxygen attenuation while the 80/90_GHz spectrum is used for longer links since they are less prone to attenuation. Connection to the core network is via fiber links terminating in the buildings.

High-gain antennas are used on the radio devices to compensate for the high atmospheric attenuation. The resultant pencil-like beam is immune to interference but requires careful antenna alignment. Inbuilt more automatic antenna steering is proposed to aid the antenna alignment process to reduce installation time, hence operational costs. The CSCB uses CRT, through the cognitive cycle, to obtain the required environmental information, makes decisions based on predefined performance requirements, and learns from processed information to minimize future decision-making processes. This paper focuses on links created due to diffraction only. It is assumed that a database already exists with initial information required to initiate link establishment. An architecture of the CSCB system is shown in figure 5. Radio 1 uses the $X_2$ [21] interface to connect and register itself on the core network. Radio 2 connects to the core network via an existing aggregation point at the roof top. Once registered on the network, the two radios iteratively point to the same diffraction point to establish a link.

### A. *Link Establishment*

Figure 6 summarizes the link establishment process. Once connected to the network, the radios download configuration information associated with current network, e.g., operating frequency, link performance requirements and applications QoS requirements. After self-configuration, each radio obtains geo-location information and that of its peer and possible diffraction points from an existing network database. Before data transfer between the small cell and the core network commences, the backhaul radio system

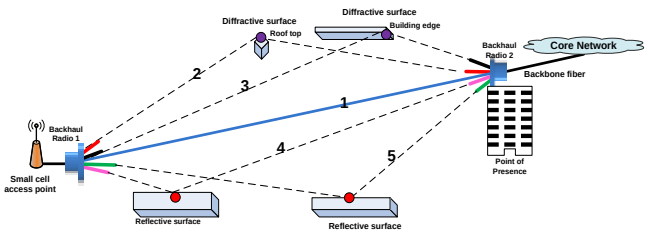



Figure 5: Link establishment in a multipath environment

performs the cognitive cycle to establish communication links.

### *B. System Cognitive Cycle*

The first stage of the cognitive cycle involves obtaining information from the network database and the environment. The decision-making stage involves using the obtained information to identify diffractive points based on predefined performance criteria. The learning process involves referring to previously made decisions to act on newly available information.

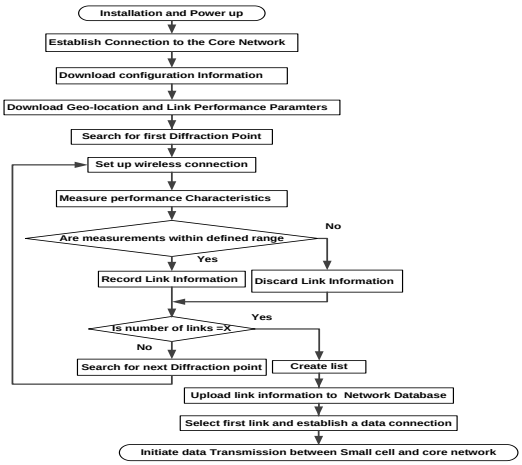


Figure 6: Link Establishment Process

#### *1) Obtaining Knowledge*

The CR uses the capabilities of the SDR such as automatic power control and adaptive modulation to adjust its internal state. Operational and protocol information is obtained from the network database. Operational information includes the wireless system operating frequency. Geographical information such as location of backhaul radios, the small cell access point, aggregation point and possible diffraction points is predefined and available for down load from the network database. Network and applications QoS specifications are also downloaded from the network database. Initial antenna orientation information must also be available for download to minimize the signal search process.

#### *2) Decision-making*

After information download, the radios point to the general link direction as defined in the initial antenna orientation information. The locations of the possible diffraction points must appear in the same order in both radios. After pointing to the first point of diffraction, the two radios must identify each other using location information, operating frequency, and MAC or IP addresses and establish their initial link. If the performance characteristics of the link are within predefined values, the link information is recorded in the local database of the radios. Such information includes link capacity, delay, jitter and throughput. The radios search for the next diffraction point and lock their beams to the point. Again, they establish connection, obtain performance information and compare this with available information. They record the links details if the parameters are within the defined range or discard the link details if the characteristics do not meet the link performance requirements. The process continues until the last diffraction point is identified and a decision is made whether to discard or keep the link details. This information is also uploaded to the network data base for future use.

The next stage involves minimization of the number of usable links to a predefined number, e.g., five. The minimization process is a multi-objective optimization process which involves checking link capacity, interference levels, link delay and jitter. After determination of the minimum number of possible links, a list of the links is created in the local database according to their performance characteristics. The link with the best characteristics is selected as the primary link to be used for communications. This information is also uploaded to the network database and can be used in the learning process. Environmental information such as weather conditions, time of day, properties of the diffractive surface is also recorded. Wireless environment information such as operating frequencies of neighboring devices and location of other SCB radios is also recorded for possible use in future decision-making process.

#### *3) The Learning Process*

The radio devices constantly monitor the performance of the primary link and other environmental information. Should the performance of the links on the current list fall below the initially measured values, the radios switch to the next best link on the list and update the list. The learning process improves system performance by first recalling the next best link from the existing database information, instead of performing a search again. If it is necessary to perform the search process, the radio nodes can recall the information about all the possible diffraction points in the system database and create a new list based on the latest recorded performance information During quiet times, the radios can rescan the diffraction points and record performance values. Environmental information is also recorded. The network database can be populated in the long term and with the performance of each link at particular times of the day, at different weather conditions and also based on the network traffic variations.

#### *4) System Reconfigurability and Autonomy*

The system self-configuration is demonstrated by the ability of the radio devices to download configuration information from an existing network database. Self-optimization is demonstrated by the ability of the radio nodes to measure the QoS parameters of a link and use the results to decide whether to keep or discard the link. The nodes further minimize the number of links to a list of five. Using the QoS values, a node selects the best link as a primary link for operation. During these processes, the radios adjust their internal parameters like operating frequency, modulation strategy, transmission power and external parameters like the diffraction point until the optimization goals are achieved.

### *C. Implementation of the Cognitive Process*

Execution of the cognitive processes requires use of existing technologies. The information used in the solution can be

obtained by downloading performance parameters, location information and weather information from a network database. This information is prerecorded and must be updated in real time on the network. The decision-making process can be implemented in algorithms. Genetic Algorithm (GA), Case-Based Reasoning (CBR) and Expert Systems (ES) are some of the technologies to implement algorithms for decision-making. GA has popularly been used in CR due to its ability to optimize more than one function objective in the optimization process, and will therefore be used in the CSCB system. GA can also be used to affect the learning process in cognitive radio due to its capability to change the link performance parameters to required values. The CSCB solution will therefore use GA for the learning process.

## V. EVALUATION AND PROOF OF CONCEPT

### A. *Experimentation and Test bed*

The evaluation tests will be carried out in three stages. The first stage will involve practical experiments using proprietary small cell backhaul radios operating in the license exempt 60_GHz frequency band. Since equipment in this band is still limited, issues of interference usually associated with license exempt equipment will be minimal. Outdoor experiments will be carried out in a typical urban canyon between UCT faculty of engineering building blocks to perform link quality test. The aim of the tests will be to evaluate the capability of the system to deliver a video application using diffracted signals. Performance measurements will include link capacity, throughput, delay and delay variation.

The second stage will involve evaluation of the developed decision-making and learning algorithms using the GA in Matlab. The system's capability to optimize the number of links and performance of each link will be tested. The final stage of the tests will involve emulation of the CR's in a Linux-based system and evaluation of the performance of the developed cognitive engine and learning and decision-making algorithms. The purpose of the tests will be to evaluate performance of the emulated system in a practical test bed environment consisting of an OpenEPC, real WiMax and WiFi networks and an emulation of a small cell access network. The machines will be interconnected to an existing real test environment called the OpenEPC. This is a prototype implementation of the 3GPP EPC. The implementation was developed by the Fokus Fraunhofer Institute [18]. A functional architecture of the EPC is shown in figure 7.

The test bed has three gateways that are used by various access networks to connect to the EPC. The S-Gw gateway allows access into the EPC by 3GPP defined access networks like the Small Cells. The ePDG allows access to IP-based networks like WiFi, while the AN-Gw allows access to non-3GPP access networks that are assigned an IP address by the EPC. The CSCB interconnects the various access networks to the EPC. Although designed primarily for small cell access points, the solution has the flexibility to be used for other access networks like WiFi and WiMax.

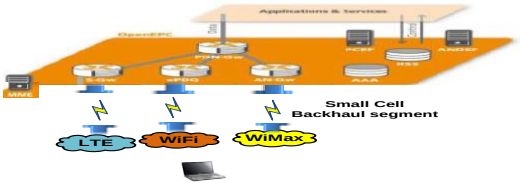


Figure 7: Test bed

## VI. CONCLUSIONS AND FUTURE WORK

In this paper we have proposed a new application of Cognitive Radio Technology (CRT) to solve the coverage problem for small cell backhaul deployment in urban canyons. We have proposed the development of a cognitive engine and supporting algorithms to enable backhaul radio devices based Software Defined Radio technology operating at millimeter wave frequencies to establish communication paths based on diffracted or reflected radio signals. The proposed solution is expected to address the coverage and capacity requirements of small cell backhaul networks in urban environments.

## ACKNOWLEDGEMENTS

The authors are thankful to Telkom-SA, Jasco/Telesciences, MimoTech, THRIP-NRF and DTI for their research support

## REFERENCES

[1] Cisco, "Cisco Visual Networking Index: Global Mobile Data," *Cisco Systems*, 2013.

[2] A. Madeisis et,al "Taxonomy of Cognitive Radio Applications," *IEEE ISDSAN*, 2012.

[3] J. Laskar, "CMOS, Cognitive and mmW: A Wireless Revolution," *Radio Wireless Symposium*, Santa Clara CA, 2012.

[4] "Backhaul Technologies for Small Cells Use Cases, Requirements and Solutions," Small Cell Forum, 2013.

[5] J. Robson, "Small Cell Backhaul Requirements," *NGMN Alliance*, Frankfurt, 2012.

[6] Shahryar Khan et al, "The Benefits of Self Organising BAckhaul Networks," *Ericsson Review*, online 2013.

[7] G. R. Maccartney.et. al, "Path Loss Models for 5G Millimeter Wave Propagation Channels in Urban Microcells," *IEEE GC*, 2013.

[8] J. Mitola, "The Software Defined Radio," *IEEE Communications Magazine*, pp. 26-38, 1995.

[9] S. Haykin, "Cognitive Radio: Brain Empowered Wireless Communications," *Selected Areas in Communications*, vol. 23, no. 2, 2005.

[10] T. S. Rappaport, et. al., "38 GHz and 60 GHz Angle dependent Propagation for Cellular and Peer to Peer Wireless Communications," *IEEE ICC*, 2012.

[11] Y. Azar et. al, "28GHz Propagation Measurements for Outdoor Cellular Communications Using Steerable Beam Antennas in New York City," *IEEE ICC*, 2013.

[12] M. Coldrey, "Non-Line-of-Sight Small Cell Backhauling Using Microwave Technology," *IEEE Communications Magazine*, Sept 2013.

[13] D. Bojic. e. al, "Advanced Wireless and Optical Technologies for Small Cell Mobile Backhaul with Dynamic Software Defined Management," *IEEE Communications Magazine* Sept 2013.

[14] R. Villar et. al, "Wireless Backhaul Architecture for Small Cells Deployment exploiting Q-band Frequencies," *Future Network & Mobile Summit*, Lisbon, 2013.

[15] P. Lakhani. et.al, " Optimization of Cognitive Radio in NLoS Backhauling of Small Cells," online 2013.

[16] J. Edmond et. al, "Millimeter Wave Propagation in an Urban Environment," *IEEE TGRS*, May 1988.

[17] J. Zander, "Spectrum for 5G - A Big Deal?," in *ETSI Future Mobile Summit*, Mandelieu, 2013.

[18] Fraunhofer Fokus, "Fraunhofer FOKUS OpenEPC toolk," http://www.openepc.net .

[19] M. Paolini. e. al, "Small-cell backhaul:Industry trends and market overview," *Senza Fili Consulting*, 2013.

[20] IEEE, "IEEE Project P802.16r: Ammendement for Small Cell Backhaul (SCB*)," IEEE Wirelessman 802.16*, https://mentor.ieee.org/802.16/dcn/16-14-0019, 2014.

[21] 3GPP, "X2 general aspects and principles", *3GPP TS 36.420 V9.0.0* (2009-12)

[22] ITU, "Propagation by Diffraction", *Rec. ITU-R P.526-5*, 1997

## BIBLIOGRAPHY

**Bessie Malila** Obtained her BSc degree from the University of Zimbabwe, and her MSc (Eng) from UCT in 2012 and is presently studying towards the PhD degree at the same. Her research interests include Cognitive Radio Technology, Millimeter Wave Communications and Mobile Backbone Networks.

# Optimisation of SlotTime for a single-radio Mid-Range Multi-hop Wireless Mesh Network

**Carlos Rey-Moreno**[1], William D. Tucker[1] and Javier Simó-Reigadas[2]
Computer Science,
University of the Western Cape[1]
Private Bag X17, Bellville 7535, South Africa
Tel: +27 21 959 3010, Fax: +27 21 959 1274
and Signal and Communications Theory
Rey Juan Carlos University, Spain[2]
email: {**crey-moreno**, btucker}@uwc.ac.za[1]; fjsimo@urjc.es[2]

**Abstract - This paper presents the business context and results of an optimisation exercise for a single-radio mid-range multi-hop wireless mesh network for the provision of VoIP services. Tweaking state of the art technology, this wireless mesh network (WMN) physically covers 30 square kilometres in rural South Africa with a dozen solar-powered nodes. Measurement of a range of values revealed a SlotTime setting that maximises aggregated throughput in this network by 115%. These results have allowed several simultaneous calls in the case study presented. Following the same optimization strategy similar improvements are expected in WMN sharing the same characteristics. We argue that this architecture is comparable yet cheaper and easier to install and maintain than multi-radio systems with directive antennas. We leverage this finding by proposing a win-win business case for a ground up community-based mesh network. Local residents benefit from free internal calls backed up by revenues from low cost voice breakout, while operators benefit from Internet provision, VoIP upstream provision and increased revenues from Mobile Termination Rates (MTR). Our novel approach offers an accessible and affordable alternative business model for residents in a rural area that have mobile connectivity yet cannot afford to use it.**

**Index Terms - Limited Range Communications, Ad-hoc, Wi-Fi, 802.11.**

## I. INTRODUCTION

Access to information and communication technologies (ICT) in Southern Africa is quite uneven. While in cities first class services can be experienced, the reality in rural areas is vastly different. Although the situation may appear to have changed dramatically in the last few years with the increasing availability of 3G and 4G services, both Internet and voice services remain unaffordable for most people.

The literature is abundant with regards to the application using wireless mesh networks (WMN) to tackle this situation [1-10]. Nowadays, WMN based on the 802.11 protocol family can provide a mature solution with several off-the-shelf products available in the market, e.g. Locus, Mesh Potato, Skylink and Tropos. However, it is difficult to compare the different real deployments reported in the literature given the various architectures available to create a mesh network.

One of the key factors to determine an architecture is the distance among the nodes forming it. There are two bodies of work; the first one concerns about long-range WMN [1-4]; where long-range refers to tens of kilometres (km). In most of those interpretations, long-range comes together with directional antennas due to the link budget and the power limitations of non-licensed bands. With directional antennas, static WMN are only feasible if each mesh node has installed several radios in order to connect to several neighbours. A similar architecture is also used when distances are smaller [5], where limitations on available non-overlapping channels, number of radios, and power consumption strongly limit possible topologies.

A second body of work considers that a mesh node is a low-cost low-power node with only one radio [7-8], so the whole WMN operates on a single channel. Hence, performance is sacrificed as each node acts as a source/sink/router and packet forwarding over a single wireless interface halves the performance. On the other hand, the gains include being cheaper, easier to install and maintain, and more flexible. From this second body of work, only [8] works on the range of few km (with a link up to 5km). However, in that work there is no reference to the effect that distance over the standard limits may have played upon the performance of the network under analysis.

In both bodies of work presented, there is no particular service that they aim to target, making it challenging to compare the business case they present.

In this paper we introduce an alternative to both bodies of work described above by proposing the use of single-radio multi-hop WMN composed entirely of up to 5km links, defined herein as mid-range. This paper expands the results presented on [9] by discussing the effect of multihop routes when optimizing the SlotTime value; and leverages the performance improvement to suggest a revised business model that commenced with [10]. The business model has evolved considerably after two years of engagement with rural community in which we came to better understanding their motivations and communications needs. In this community a mid-range architecture such as that described above is used to provide voice over Internet Protocol (VoIP) services. This paper presents this architecture in detail, together with a preliminary optimisation of its performance. We believe that the methodology to optimise the performance and the business case presented can increase the range of options to provide low-cost communications services to rural areas in Southern Africa, where such distances are common between homesteads.

## II. RELATED WORK

Using VoIP services over single-radio WMN to provide an alternative to cellular technologies has been explored by [6]. However, their work focused on urban scenarios where the density of users and the lack of line of sight (LoS).

requires a higher number of nodes, driving the cost of this architecture those of cellular coverage. The reduced population density and lack of buildings in rural areas could reverse these economics and worth exploring.

Eventually, the limitations of single-radio mesh networks also led some authors to discard this architecture as a feasible scalable solution to provide access in rural areas [4]. They acknowledge the potential benefits of combining both approaches described in the introduction, using the single-radio approach for the access tier of the architecture, and the multiple-radio to interconnect mesh cells. Although not explicitly stated, the mesh cells are considered to be made of links of tens of meters, like the ones in the Roofnet network provided as an example [7]. We fundamentally agree with the tiered architecture, however we propose that by widening the range of the links which make up the mesh cell it is possible to provide connectivity to rural villages with a dispersed population using a single radio WMN at a fraction of the cost and the complexity of using multiple-radios. Such a network can be linked up to the Internet by using a single long range point-to-point (PtP) WiFi link as suggested by [4], or another suitable technology, to cover the distance to the closest wireline Internet access point. Using this architecture for providing voice services is exactly what is proposed by Village Telco [11]. However, they have not explored the possibility of having links longer than 1km. Increasing the distance among nodes, and its footprint, will reduce the number of nodes necessary to cover a given area and therefore reduce the capital expenditure required (which increases due to the lack of electricity in rural areas).

Most of the literature studying the effect of distance on WMN considers mainly the multiple radio mesh architecture described above [1-3, 12]. For that architecture, several optimisation proposals have been made available. In [12], a Medium Access Control (MAC) layer based on Time Division Multiple Access (TDMA) is proposed to optimise performance of PtP links. This work was further optimised by [1], adapted to point-to-multipoint links [2]. However, this alternative MAC does not provide support for single-radio mesh, e.g. multipoint-to-multipoint (MtM) and so, no results are available in this regard. Other authors, e.g. [3], have provided a model to analyse the performance of standard 802.11 MAC with long distances and proposed ways to optimise the performance by tweaking the values of some MAC parameters. However, in the simulations for MtM included by such work, only single hop communications were considered.

## III. ARCHITECTURE IN THEORY AND PRACTICE

### *A. Description of the architecture and the business case*

The mid-range single radio WMN presented herein aims primarily at reducing the cost of voice services in rural areas, acknowledged as the most important service for their dwellers [13]. In South Africa, for instance, mobile voice services, although available, remain unaffordable for many.

This architecture is not intended to reach regional scalability, but rather to provide sufficient fixed points of access (as many as nodes form the mesh backbone) to provide blanket coverage to rural villages that will then be linked up to the Internet with the most convenient technology available. Sharing an Internet connection is a solution that has been used successfully by community networks worldwide. The nodes forming the mesh backbone also provide the functionality of distributed wireless access points, thus allowing for Wi-Fi clients.

Production of Wi-Fi chipsets outgrew that of mobile phones in 2012 and by 2017 it will nearly double it according to [14] and [15]. WiFi is now present in every computer, tablet, all ranges of smart phones and in many feature phones, too. The current prediction for 2017 is that 74% of all manufactured mobile phones will be smart [14]. Considering their uptake in emerging markets, the architecture presented here will allow using either GSM or WiFi for voice communications on a single device. To do so, the nodes should have a VoIP server responsible of routing the calls from SIP clients configured in the WiFi -enabled phones. Additionally, they should have an Analogue Telephone Adapter (ATA), so an analogue phone can be directly connected allowing those not owning a mobile phone, or owning a non WiFi-enabled one, to benefit from this architecture too. Hence, calls with no operating costs among phones within the footprint of the network. Furthermore, with the right institutional structure in place and partnering with VoIP providers, calls can be made from these phones to any phone outside the network at a fraction of the cost they currently pay. Therefore, the created institution will have enough margin to keep the network operational while obtaining some surplus.

Although it might be considered that this architecture will only benefit VoIP providers and the community members, we see other stakeholders benefiting too. Using data from a survey conducted in December 2013 in 255 households from the community in the Eastern Cape Province of South Africa where we work, following a stratified random sample, only 26,7% of the individuals have reported making all or most of their calls to people staying in the community. This correlates well with the high emigration rates in the area, resulting in families split in urban and rural dwellings. This means that greater proportion of the calls will still terminate in the incumbents' networks, thus entitled to receive the Mobile Termination Rates (MTR). It is expected that with cheaper rates people will communicate more, and so the money received from the MTR will eventually surpass the meagre profit obtained from the few calls made in the current scenario. That would be the case if all calls were made from the WMN architecture described. However, a more realistic scenario is a blend of pre-existing mobile calls and VoIP using the WMN.

If the number of expected simultaneous calls is not very high (i.e. calls from analogue phones only), the Internet requirements for VoIP provision could be covered by existing 3G links, where available. As this number grows, other gateway options should be considered. Satellite technology, traditionally considered for Internet connectivity in rural areas, is not well suited for simultaneous voice services. The increased capillarity of fibre optics, especially in South Africa, allows, following the model from [4], using PtP WiFi link to connect the WMN to high capacity technologies. Although it falls outside the scope of this paper, this opens up the possibility to increase the portfolio of services provided and so to strengthen the business case, while at the same time provides a business case for those operating the fibre. We see this as a win-win situation for all involved.

*B. Description of the network in the field*

In the case study described below, only calls from analogue phones are allowed. Both calls to other analogue phones in the network, and to any phone outside of it are allowed. After conversations with different providers, calls from the fixed phones outside the network are made at a third of the price that villagers are currently paying to use their mobile phones. The initial investment to provide this service has been provided by charging villagers a fee to charge their mobile phones using the spare solar power generated in the node sites [16].

The minimal presence of WiFi-enabled phones or computers in the area has prevented the additional functionality of allowing these devices to make calls at the moment; something that, as argued above, will become a possibility in the long run.

The mesh network consists of 12 nodes scattered around 30km$^2$. The connectivity graph of the network is shown in Figure 1. The colours correspond as follows: green for links with reported RSSI (received signal strength indication) above 12 dBm, orange for links with one or both sides with values between 9 and 12 dBm, and red where one or both are below 9 dBm. Yellow links are so weak that they disappeared when setting the operating mode to 802.11g from 802.11bg.

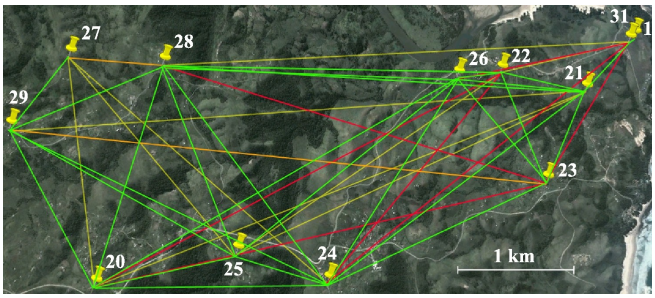



**Figure 1: Network links and distances.**

Ten of the 12 nodes are located in private houses chosen by the local authority. One of the other two nodes provides access to the server (1) and the last one is a repeater connecting the latter to the rest of the network (31). The server is located in the headquarters of a local NGO (non-governmental organisation), which has access to grid electricity and a backup battery system during power outages, but is located behind a hill so a repeater is needed. One of the design criteria to select hosting homesteads was seeing at least three other houses [17]. Thus, most of the chosen houses were those on top of hills and provide clear LoS with multiple neighbour nodes.

**Table 1: Comparison between hardware used.**

| **Hardware features** | **NS2** | **MP01** |
|---|---|---|
| Weight | 0.4 Kg | 1.1 Kg |
| Consumption | 4W | 2.5W |
| Chipset | Atheros AR2315 | Atheros AR2317 |
| Operation Modes | 802.11b/g | 802.11b/g |
| Max Tx Power | 26 | 20 |
| Antenna Gain | 10dBi panel | 8 dBi omni |

The 10 nodes in private houses are first generation Mesh Potato (MP01) with an external antenna attached; the repeater is an off-the-shelf Ubiquiti NanoStation (NS2), and the remaining node is an off-the-shelf MP01. Their technical specifications can be found in Table 1.

The NS2 utilises power over Ethernet (PoE) and the MP01 runs with an innovative power over telephone line (PoTL), a la PoE, but rather connecting an analogue telephone to the router along with power, which greatly simplifies the installation. The light weight of both antenna (MP01) and router allow them to be mounted on a standard bent aluminum TV pole attached to a house, thus avoiding the need for a tower and reducing deployment costs. Low power consumption further helps to reduce the cost of operation. The MP01 also includes a Foreign eXchange Station (FXS) port that allows a direct connection of an analogue two-wire telephone.

All nodes have been flashed with version 1.1 of the Village Telco Small Enterprise and Campus Network (VT-SECN) firmware, which is a customised version of OpenWRT Kamikaze. SECN includes madwifi-trunk-r3314 and batman-adv version 2011.2.0 and configures the wireless card to operate in ahdemo and ap modes simultaneously, utilising madwifi virtual interface functionality. The firmware for the MP01 also runs an Asterisk server that allows direct voice communication with any other MP01 flashed with VT-SECN as long as they have different IP addresses. A call can be setup by dialing the last octet of the IP address on an analogue phone, and the device itself can be configured following voice prompts, so avoiding the intervention of an expert. The firmware also automatically configures batman-adv over the interface operating in ahdemo mode, and the resulting bat0 interface created to run the protocol is bridged to the Ethernet and the Access Point interfaces. This allows end-devices to communicate transparently through the nodes without any additional configuration other than setting up an IP address within the range of the network (and selecting its SSID).

## IV. MATERIAL AND METHODS

*A. Theoretical and practical background*

In 802.11, access to the wireless channel is based on Carrier Sense Multiple Access/Collision Avoidance (CSMA/CA). Simo-Reigadas et al. provide an in-depth explanation of the parameters that influence the behaviour of CSMA/CA when long distances are considered [3]. These parameters are different for each of the 802.11 PHY layers. The standard values for those PHY layers available in the devices used are shown in Table 2. The presence of two alternate values in HR-DSSS (High Rate / Direct Sequence Spread Spectrum) is due to the compatibility mode of 802.11g: the second value is selected only if there is not an 802.11b-only station in the network.

**Table 2: CSMA/CA parameters per PHY.**

| PHY layers in 802.11b/g | DSSS | HR/DSSS | OFDM |
|---|---|---|---|
| SIFS | 10 µs | 10 µs | 16 µs |
| SlotTime | 20 µs | 20/9 µs | 9 µs |

The ACKTimeout is not given a closed value in the standard, but it is defined as ACKTimeout = aSIFSTime + aSlotTime + aPHY-RX-START-Delay; with aSIFSTime and aSlotTime, the SIFS and SlotTime included in Table 2. The third value is the time required for the PHY layer to realise it is receiving a frame and to generate an alarm to get ready to process it. Its value is PHY-dependent, and practical implementations, like the one made for madwifi, account for this parameter internally. Tests done using

athctrl for different PHYs set the same value for ACKTimeout. Then when madwifi refers to ACKTimeout, it does not refer to its complete definition in the standard, but only to the sum of the first two parameters shown above. In this paper, we refer to the ACKTimeout similarly.

The timing structure in the standard is built considering that the propagation delay is always shorter than 1 µs. For distances longer than 300m this is no longer true[1], and then the standard CSMA/CA mechanism does not operate as expected and collisions may appear with higher probability. Furthermore, if the ACKTimeout is not adjusted accordingly, the available throughput gets severely reduced since the ACK may arrive but get discarded since it does so late. Then the same packet is transmitted until the maximum number of retransmissions is reached. The madwifi driver allows the modification of the ACKTimeout and the SlotTime either individually or using the athctrl tool that sets them automatically when a given distance (in meters) is provided as an argument. When the second option is used, the SlotTime value is fixed to 9 plus the propagation delay for that distance, and ACKTimeout to SlotTime * 2 + 3.

Detailed descriptions of other standard mechanisms in the operation of mesh networks, like determining the IBSS and the time synchronisation, are skipped for simplicity since the use of *ahdemo* mode bypasses these mechanisms. A*hdemo* is a non-standard implementation that does not transmit beacons to form the mesh network. While this solves the Cell-Split and 'stuck beacon' problems, it requires nodes to individually configure the IBSS and the internal timers instead of those provided by the network.

The driver allows using different transmission rates with each neighbour (unicast rate) when configuring the transmission rate to auto. Several algorithms are allowed, but, again, one is considered to be standard: Minstrel. It records statistics of all packets transmitted (successfully or not) towards each neighbour and the rate used for each one. This should be enough for selecting the optimum for static channels. However, the wireless channel changes; so it is required to try other rates. Minstrel uses a percentage (10% by default) of the unicast packets that are sent at a rate other than optimal to adapt to the channel changes.

The broadcast rate is fixed using the parameter mcast_rate, which by default is 1 Mbps. The standard suggests that this rate should be the minimum rate supported for all neighbours, but tests have shown that configuring all nodes to 802.11g with a minimum operating rate of 6 Mbps, sets mcast_rate to 1 Mbps and so requires manual setting.

### *B. Optimisation methodology*

Drawing on the background provided above, when possible, a fixed value will be provided for the described parameters. For the SlotTime, where no previous work exists on how it affects the performance in a mid-range mesh network, traffic was generated on the network to assess the effect of varying its value. The relevant parameters were configured as follows:

- Radios used 802.11g PHY.
- Broadcast messages sent at 6 Mbps. This value was used only for the last set, as explained below; in the two other ones broadcast messages were sent at 1 Mbps.
- Unicast messages sent at rate chosen by Minstrel.
- Radios operated in basic mode, not using RTS/CTS.
- ACKTimeout fixed to 57 µs.

In [3] it is recommended that the value of ACKTimeout be set to the default used on the chipset plus round trip propagation delay for the longest link in the network, for all nodes. Thus, considering a link in which two nodes receive signal from each other, the longest link that is present in our mesh network is 5.05km (in between nodes 21 and 29). For madwifi the default value for the ACKTimeout is 23 µs, to which we have added 34 µs of the round trip delay on the 5.05km link, for a final value of 57 µs. Further analysis, has shown that this link is never used by batman-adv for unicast traffic due to its low RSSI. The longest active link in the network is 4.65km (in between the nodes 23 and 29). Thus, an ACKTimeout of 54 µs could have been used for further optimisation. However, the difference would have been negligible (3 µs less waiting for the ACK when collision occurs), and we preferred to cover all potential single-hops in the network.

To find the optimum SlotTime, we saturated the links to get the maximum achievable throughput in between every pair of nodes. To do so, we generated UDP traffic using iperf for 20 seconds in between two pairs of nodes consecutively. UDP traffic was chosen since the network is intended to be used mainly for VoIP traffic. The amount of traffic that saturated every pair of nodes was not straightforward. Initially 20 Mbps were generated in each direction, but iperf failed to handle such an amount of traffic and some links that allowed real calls presented 0 bps capacity. The maximum value of total throughput (sum of both directions) obtained in each per link, after having tested all the SlotTimes in the first set (from 9 µs to 54 µs), was multiplied by 0.75 and injected in each direction instead of 20 Mbps.

The results were far more 'real' with this setup, with almost all links providing some bps for all the SlotTime tested values (second set, from 24 µs to 119 µs). In this set values of SlotTime from 39 µs to 84 µs were not valid since one node shut down during the night thus changing the topology of the network. Again, some of the links were able to handle all the traffic injected for some of the SlotTime values tested. So for those links a similar procedure like the one described above was repeated. In addition, during this second set, we realised the broadcast rate was not being updated as mandated in the standard. So for the next set it was manually configured to 6 Mbps. With these two changes, a third set of results was produced with different SlotTime values ranging from 9 µs to 999 µs. In this third set, a total of 49 values were tested, and it is for this set these values that results are discussed below.

## V. RESULTS FROM TRAFFIC GENERATION

### *A. Modifying SlotTime*

In this section we show the results obtained from the third set of traffic generation. Figure 2 shows the aggregated end-to-end traffic between each pair of nodes in the network per value of SlotTime.

[1]The AirPropagationDelay parameter in the standard is said to be << 1 µs but, in fact, in most 802.11 hardware implementations, the electronics are faster than required by the standard and it works properly for longer distances; up to almost 1.4 km in 802.11a/g and up to almost 3 km in 802.11b.

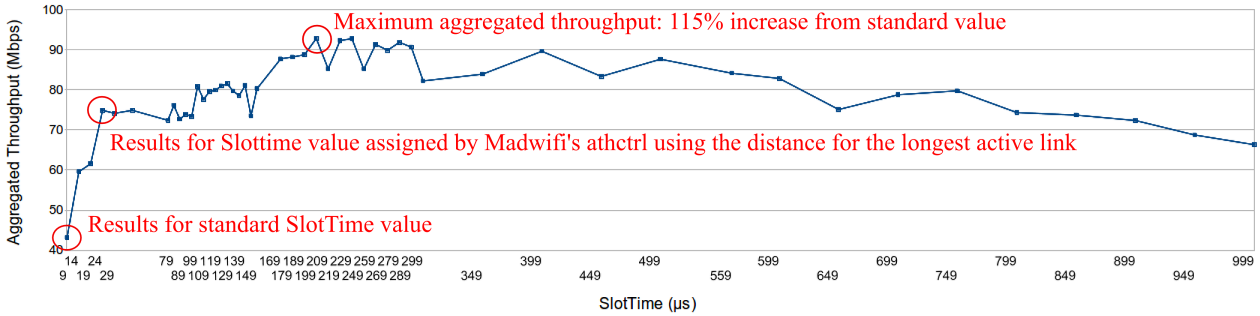



**Figure 2: Variation of the aggregated throughput with SlotTime.**

We acknowledge that this indicator does not show the maximum traffic the network can handle simultaneously, since several internal links are counted more than once as they are part of longer routes and the effect of the exposed node is neglected. However, it does offer a single value that can be compared among different SlotTime values.

As Figure 2 shows, aggregated traffic peaks for SlotTime at 199 µs, for which 92.75 Mbps are obtained. Although a single repetition was carried out for every SlotTime value, a similar trend is confirmed by the curve shown: other values around the peak exhibit similar performance. These values contrast heavily with the standard value for SlotTime in 802.11g (9 µs) for which 43.11 Mbps are obtained. Thus, using Slotime values around 199 µs provides an increase of up to 115% (92.75 Mbps/43.11 Mbps) over the standard value. Using Madwifi's *athctrl* for the longest active link in the network (4.65km) sets the the SlotTime to 24 µs. For this value, 74.87 Mbps are obtained, showing a considerable improvement (57%) over the standard value (74.87 Mbps/ 43.11 Mbps), but significantly below the ones obtained around SlotTime 199 µs. For this value of SlotTime, end-to-end traffic has on average 1.40 Mbps, with the route between 22 and 26 handling the most traffic, 4.49 Mbps, and the route between 1 and 27 carrying the least, 0.142 Mbps.

### *B. Considering the number of hops*

Due to using batman-adv, routes are likely to have changed from test to test, and even within tests, which may explain the variation of throughput shown in Figure 2. As a way of capturing the length of every route, Table 3 shows the most probable number of hops between every pair of nodes. It was obtained by using the next best hop to reach the other nodes reported by batman-adv at the end of every test. Using this table, the total throughput per number of hops was obtained for each of the values of SlotTime tested. Results were normalised in order to compare the results for each number of hops, and these are shown in Figure 3.

It can be seen that the total throughput increases with the SlotTime value in two phases. First, it gets a local maximum for all cases at 24 µs, a value that considers the one-way propagation delay in the network. Secondly, the total throughput for each case continues to increase up to a global maximum that happens in a strict order of number of hops. Then, oscillations occur around the maximum value, which peaks at different individual experiments. This shows that the optimal SlotTime value must be chosen according to the entire network topology and not only based on the one-hop propagation delays.

**Table 3: Number of hops for potential routes among the 12 nodes in the network.**

| | 1 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 31 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | 4 | 2 | 3 | 3 | 3 | 4 | 3 | 5 | 3 | 4 | 1 |
| 20 | | | 3 | 4 | 2 | 2 | 2 | 3 | 2 | 2 | 1 | 4 |
| 21 | | | | 1 | 1 | 1 | 2 | 1 | 3 | 1 | 2 | 1 |
| 22 | | | | | 1 | 3 | 2 | 1 | 4 | 2 | 3 | 3 |
| 23 | | | | | | 1 | 2 | 1 | 2 | 2 | 1 | 2 |
| 24 | | | | | | | 2 | 1 | 2 | 1 | 1 | 3 |
| 25 | | | | | | | | 1 | 2 | 1 | 1 | 3 |
| 26 | | | | | | | | | 3 | 1 | 2 | 2 |
| 27 | | | | | | | | | | 2 | 1 | 4 |
| 28 | | | | | | | | | | | 1 | 2 |
| 29 | | | | | | | | | | | | 3 |
| 31 | | | | | | | | | | | | |

In the values obtained per number of hops, there seems to be a very strong correlation between: a) the final throughput in the whole route; and, b) the throughput in the slowest hop in the route divided by the number of hops.

## VI. Discussion Of The Optimisation Results

There are two factors that appear to have an influence on increasing the throughput. Firstly, increasing the SlotTime to 24 µs, I.e taking into account the propagation delay of the longest active link in the network (4.65km). A noticeable increase can be seen when using 24 µs instead of the standard 9 µs. Secondly, for values much higher than 24 µs, the throughput for routes longer than one hop also experience a noticeable increase following the trend: the longer the route, the bigger the SlotTime value that produces such an increase. A tentative explanation for this may arise from the exposed node problem, i.e. the

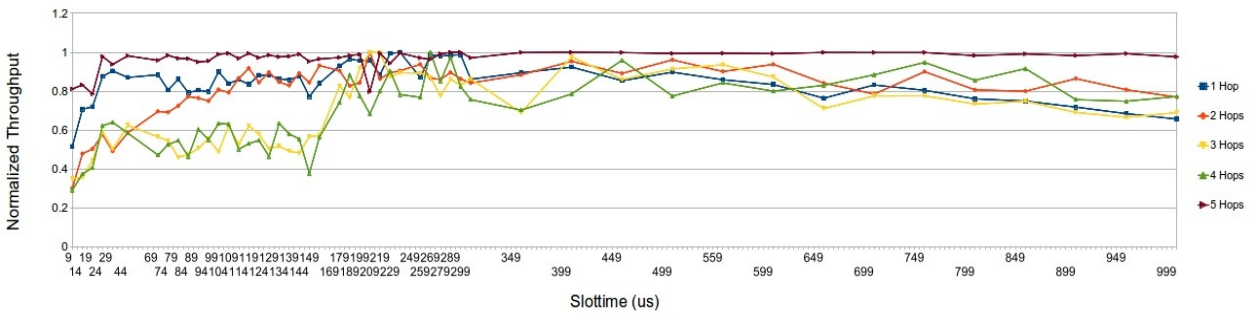



**Figure 3: Variation of the normalised aggregated throughput with the SlotTime per number of hops.**

interference range for a node is much larger than its transmission/reception range. Possibly, the longer the SlotTime is, the smaller the impact of the exposed node problem on the system. However, the distance between the furthest nodes in the network able to cause this problem is not enough to justify the SlotTime values that maximises traffic. Cross-interference between non-contiguous links could also play a role when the number of hops is bigger than two, but would not explain the steady increase in available throughput for two-hop routes up to SlotTime 114 µs. Thus, more research is needed to justify these positive empirical results.

## VII. CONCLUSION

In this paper, an architecture for a single-radio mid-range multi-hop WMN for the provision of VoIP services in rural areas of Southern Africa has been introduced. This architecture, which is cheaper, and easier to install and maintain, than multi-radio systems with directive antennas, has been argued to create a win-win business case to all parties involved, from rural villagers to service providers. For this network, using a methodology that would produce comparable results in similar scenarios, an increase of 115% of the aggregated throughput has been obtained by optimizing the value of the SlotTime. Further improvements are expected by optimising the CSMA/CA parameters as well as by using other available techniques. Still without them, the throughput measurements obtained for our multi-hop network show that values are clearly sufficient for carrying out VoIP traffic. Although analysis of other Quality of Service (QoS) parameters has not been included, user experiences validate the results obtained. In order to create a more generalizable model, a whole set of QoS parameters will be studied in different topologies via simulation once the appropriate tools have been developed.

## ACKNOWLEDGEMENTS

We thank Jay, Nic, Thomas, Nancy, Jor, Carlos the Village Telco community and Transcape. We also thank Telkom, Cisco, Aria Technologies and THRIP (Technology and Human Resources for Industry Partnership) for their financial support via the Telkom Center of Excellence (CoE) at the University of the Western Cape (UWC). This work is based on the research supported in part by the National Research Foundation of South Africa (Grant number (UID) 75191). Any opinion findings and conclusion or recommendations expressed in this material are those of the authors and therefore the NFR does not accept any liability in this regard.

## REFERENCES

[1] R. Patra, S. Nedevschi, S. Surana, A. Sheth, L. Subramanian, and E. Brewer. WiLDNet: Design and Implementation of High Performance WiFi Based Long Distance Networks. In Proc. 4th USENIX Symposium on Networked Systems Design and Implementation, 2007.

[2] J. Simo-Reigadas, A. Martinez-Fernandez, F.-J. Ramos-Lopez, and J. Seoane-Pascual. Modeling and Optimizing IEEE 802.11 DCF for Long-Distance Links. IEEE Transactions on Mobile Computing, 9(6): 881-896, 2010.

[3] R. Patra, S. Surana, S. Nedevschi, and E. Brewer. Optimal Scheduling and Power Control for TDMA Based Point to Multipoint Wireless Networks. In Proc. 2nd ACM Workshop on Networking Systems for Developing Regions, 2008.

[4] L. Subramanian, S. Surana, R. Patra, S. Nedevschi, M. Ho, E. Brewer, and A. Seth. Rethinking Wireless in the Developing World. In Proc. 5th Workshop on Hot Topics in Networks, 2006.

[5] G. Bernardi, P. Buneman, and M. Marina. Tegola Tiered Mesh Network Testbed in Rural Scotland. In Proc. ACM Workshop on Wireless Networks and Systems for Developing Regions, 1-8, 2008.

[6] A. Arjona, C. Westphal, J. Manner, A. Ylä-Jääski, and S. Takala. Can the current generation of wireless mesh networks compete with cellular voice? Computer Communications, 31(8): 1564-1578, 2008.

[7] J. Bicket, D. Aguayo, S. Biswas, and R. Morris. Architecture and Evaluation of an Unplanned 802.11b Mesh Network. In Proc. 11th International Conference on Mobile Computing and Networking, 31-42, 2005.

[8] D. Johnson. Evaluation of a single radio rural mesh network in South Africa. In Proc. 2nd IEEE/ACM International Conference on Information and Communication Technologies and Development, 2007.

[9] C. Rey-Moreno, W. D. Tucker, and J. Simo-Reigadas. Tuning a mid-range rural WiFi-based mesh network. In Proc. 4rd ACM Symposium on Computing for Development. Cape Town, South Africa. 2013.

[10] C. Rey-Moreno, Z. Roro, M. J. Siya, J. Simo-Reigadas, N. J. Bidwell, and W. D. Tucker. Towards a Sustainable Business Model for Rural Telephony. In Proc. of the III International Workshop on Research on ICT for Human Development, 2012.

[11] Village Telco. Retrieved May 4, 2013 from www.villagetelco.org.

[12] B. Raman and K. Chebrolu. Design and Evaluation of a new MAC Protocol for Long-Distance 802.11 Mesh Networks. In Proc. Annual International Conference on Mobile Computing and Networking, 2005.

[13] F. J. Proenza. The road to broadband development in developing countries is through competition driven by Wireless and VoIP. In Wireless Communication and Development: A Global Perspective, 2005.

[14] IDC. IDC-Press Release. Retrieved Nov 4, 2013, from http://www.idc.com/getdoc.jsp?containerId=prUS24302813.

[15] IHS. Small Cells with Wi-Fi Set to Reshape Wireless Communications Market. Retrieved Nov 4, 2013 from http://press.ihs.com/press-release/design-supply-chain/small-cells-wi-fi-set-reshape-wireless-communications-market.

[16] C. Rey-Moreno, Z. Roro, M.J. Siya and W.D. Tucker. Community-based solar power revenue alternative to improve sustainability of a rural wireless mesh network. In Proc. 6th International Conference on Information and Communication Technologies and Development. Cape Town, South Africa. 2013.

[17] C. Rey-Moreno, Z. Roro, W.D. Tucker, M.J. Siya, N.J. Bidwell, and J. Simo-Reigadas. Experiences, challenges and lessons from rolling out a rural WiFi mesh network. In Proc. 3rd ACM Symposium on Computing for Development. Bangalore, India. 2013.

**Carlos Rey-Moreno** received a Telecommunications Engineering degree at University Carlos III of Madrid and two Master degrees: Development and International Relations at Aalborg University and Telecommunications Networks for Developing Countries at University Rey Juan Carlos. Since 2012 he is an affiliated PhD student at UWC, where he researches the operation of a sustainable community network in rural South Africa.

# A Hybrid Fuzzy Logic-Based Call Admission Control in LTE Networks

Christophe B. Tokpo Ovengalt, Karim Djouani and Anish M. Kurien
Department of Electrical Engineering/ F'SATI
Tshwane University of Technology, Private Bag x680, Pretoria 0001
Tel: +27 12 3825911, Fax: +27 12 3825114
Emails: {tokpocb, djouani}@gmail.com, kurienam@tut.ac.za

**Abstract - The inaccurate measurement of key Call Admission Control (CAC) factors such as latency and packet loss, as well as a limited understanding of their respective influence on the end-user's data throughput, have caused mobile networks to underperform in terms of Quality of Service (QoS) provisioning. QoS factors include call rejection rates, call dropping and throughput. These QoS parameters are themselves subject to other factors such as user mobility, multi-antenna configurations and the state of channels across the network. In a wireless environment which deals with several service classes of different QoS needs, there is a growing need for the development of a robust admission scheme that reduces the negative impact of misestimating delays and losses. This paper is an improvement on the work presented in [1] where uncertainty is managed at the antecedents of the Fuzzy Logic Controller (FLC). The results obtained highlight the benefits of using type-2 fuzzy sets at the input of the controller.**

**Index Terms—Hybrid FLC, type-1 FLC, type-2 FLC, Channel Aggregation, CAC, and zSlice**

## I. INTRODUCTION

In wireless networks across the world today, a lot of research focus is on deploying 4G technology which promises to provide high user data rates of more than 1Gbps and lower probability of call dropping. Higher data rates have contributed to the increase in demand for real-time data applications such as videoconferencing, media streaming and mobile gaming. There are several challenges to maintain the QoS of ongoing real-time traffic while reducing the call rejection rate because of the growing demand for connections. QoS parameters include throughput, latency and delay. Wireless channels - also referred as wireless links in this paper - are assigned to users depending on availability. However, when demand exceeds the available network resources, there is a risk for congestion and a degradation of QoS for ongoing users [2][3]. Call Admission Control (CAC) is the technique commonly used to prevent congestion in a network. This paper focuses on data services as it is assumed that voice calls and data traffic are dealt with separately. Admission decisions are based on the channel quality and the expected throughput rate of the requested service. The accurate determination of the channel quality becomes very crucial when it comes to servicing real-time applications. Furthermore, factors such as packet loss and transmission delay are time-varying in a wireless network. The next section is a review of previous works. Then, a proposed model, a hybrid FLC and the results obtained are presented in sections III, IV and V respectively.

## II. REVIEW OF PREVIOUS WORKS

A number of CAC algorithms have been developed in wireless networking over the years for QoS provisioning, prioritization of services, fair resource allocation among subscribers and congestion control. Most specialists agree that the key factors in the determination of a channel's quality are packet loss and latency. It is therefore crucial to establish a relationship between these parameters and the link's throughput.

Mathis *et al.* [4] developed a model which predicts the data throughput of a TCP connection. They find that the bandwidth *B* that is available to the end-user is given by the relationship $B = (MSS \times C)/(RTT \times \sqrt{P})$; where *MSS* is the maximum segment size, *C* is the constant of proportionality, ideally $C = 1$, *RTT* is the round trip delay and P is the random packet loss of constant probability. Their findings are very useful in understanding the behavior of a network connection given the estimated packet loss and latency. However, they fail to consider some of the limitation associated with real-time services which tolerate various levels of packet loss [5]. Furthermore, the relationship between bandwidth, loss and delay collapses when the loss value approaches zero.

Wireless channels are more unstable and more subject to changes over time compared to wired links. This is due to a variety of factors such as thermal noise, interference, etc. The imprecise and time-varying nature of some admission parameters which define the QoS has led to the necessity to properly manage the changes that wireless channels are subject to over time. Probabilistic approaches have been developed to deal with loss uncertainty by predicting the behavior of channels [6][7] but the high number of uncertainty present in the wireless link causes very high levels of computational complexity. Moreover, probabilistic approaches focus more on accurately predicting the occurrence of an event, whereas fuzzy logic associates meaning and degrees of truth to a particular event. Fuzzy

logic allows for a more human-like evaluation of the level of influence each key factor has on the QoS. In [2], type-1 fuzzy logic controllers are used in a dynamic pricing scheme aimed at influencing the level of demand the CAC has to deal with as well as increasing revenue for mobile operators. The results presented show a more balanced demand over a 24 hour period as opposed to the high discrepancies observed when flat tariffs apply. However, a number of uncertainties still exist and critics of type-1 fuzzy logic often argue that its membership functions (MFs) are themselves not fuzzy in nature.

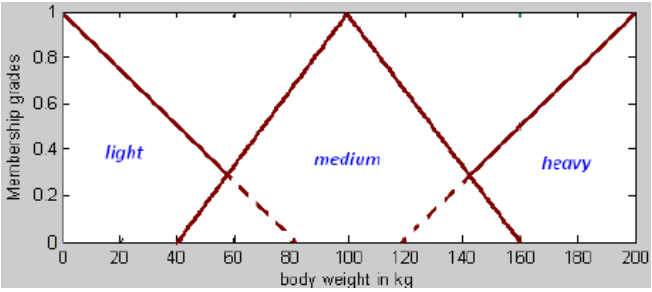



**Figure 1: Type-1 Membership Functions**

*Figure 1* above illustrates type-1 fuzzy logic membership functions (MFs) which are often referred to as primary membership functions. These MFs are two-dimensional with crisp values such as mass and length on the *x*-axis and membership grades $\mu_x \epsilon [0;1]$ on the *y*-axis. As opposed to 'true/false' or 'yes/no' statements used in Boolean algebra, fuzzy logic uses membership grades to determine the degrees of truth for every crisp quantity [9][10].

Boumella *et al.* [3] investigate uncertainty in the network traffic and attempts to address this unpredictability using Type-II FL; they conclude that *"the use of T1 FLC and T2 FLC as congestion controllers allows to alleviate congestion... T2 FLC gives better performance, in terms of QoS meeting and operator's revenue, than the T1 FLC use."* This indicates that the deployment of higher-level FL could yield better results in any attempt to achieve optimum utilization of the available bandwidth.

Higher order type-2 fuzzy logic was developed to cope with the imprecise nature of variable parameters. This is made possible by defining a region called footprint of uncertainty (FOU) which is delimited by an upper and a lower limit as illustrated in *Figure 4 & 5*. However, concerns were raised regarding the suitability of type-2 FLCs for real-time processing. Type-2 FLCs are more complex and require more computational effort than type-1 FLCs. The zSlice method was introduced to make type-2 FLCs suitable to real-time applications. It is an improvement on the general and interval type-2 techniques [8].

Interval type-2 fuzzy sets were introduced to considerably reduce the computational complexity associated with general type-2 systems that were deemed unsuitable for most real-time applications. Interval type-2 fuzzy sets are general type-2 fuzzy sets which have secondary membership equal to 1 for all primary memberships [8][9][10]. The zSlice method uses the same principle as interval type-2 sets with the exception that the secondary membership grades are not fixed to 1.

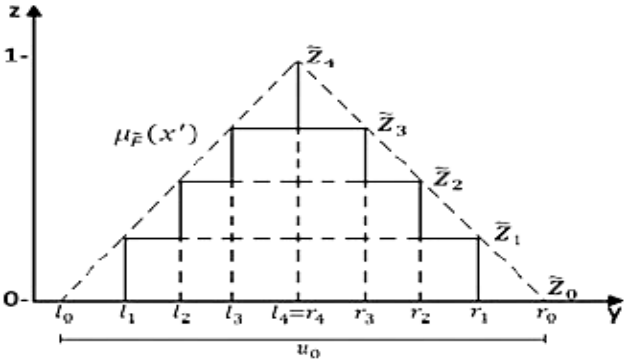



**Figure 2: zSlice Method**

With zSlices, every interval $u_i = [l_i; r_i]$ has a corresponding height $z_i$; where $0 \leq z_i \leq 1$. The values for these heights depend on the number of slicing $I$ which is user-defined. The higher the value of $I$, the greater computational effort, it is common practice for developers to keep $I \leq 5$. The third dimension (z-plane) membership functions of two type-2 fuzzy logic approaches are shown in *Figure 2*. The most popular defuzzification scheme is the Centroid Method defined below [9][10]:

$$x = \frac{\int \mu(x) x dx}{\int \mu(x) dx}$$

Where *x* is the crisp output of the FLC based on the centroid of the areas of non-zero MFs considered.

## III. TRAFFIC MODEL

The proposed model is a based on a 7-cell system which deals with call requests of different classes, with various QoS requirements and possible handoffs [1]. All cells have hexagonal shapes and are the same size. Calls arrive randomly based on a Poisson distribution algorithm. There are seven different carrier sizes in every cell: 180 kHz, 1.4 MHz, 3 MHz, 5 MHz, 10 MHz, 15 MHz and 20 MHz channels; Different types of calls with different QoS needs arrive randomly over time, with an average of 3600 calls during the busiest hours. The high demand of increased throughput for services such as data download necessitates the aggregation of 2 or more channels with a maximum of 5 aggregated channels for each call.

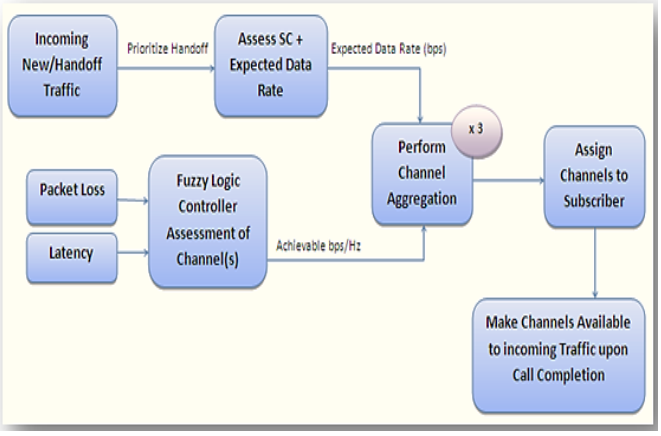



**Figure 3: Block diagram of proposed model**

Upon arrival, a service request is assigned bandwidth depending on its expected minimum end-user data rate; the data rates are assigned randomly but with respect to the percentage of traffic for each Service Class. The channel condition is evaluated by the FLC given its estimated packet loss and latency. The channel condition and the end-user's data rate requirements dictate the bandwidth that should be assigned to the user. This bandwidth is allocated after a carrier aggregation of available channels is performed. Upon completion of the call, the channels are made available to service incoming traffic. In the unlikely event of insufficient bandwidth, the call is rejected or dropped.

Although there are more and more real-time data applications available today, four (4) service classes are investigated. Each service class refers to applications or services with similar data rate requirements. Some other 4G services may or may not fit into one of the above classes. Only typical speed values are considered in this work in order to reduce complexity. Voice calls are not considered because most mobile operators manage their voice traffic using 2G or 3G technologies in heterogeneous systems.

| | Service Description (RT or NRT) | Expected Throughput |
|---|---|---|
| **Class 1** | Video calls, HD Audio/music streaming (RT) | 192 kbps up to 512 kbps |
| **Class 2** | SD video streaming, videoconferencing (RT), Mobile Internet | 1 up to 3 Mbps |
| **Class 3** | HD video streaming, Multi-player Online Gaming, 3D TV, Multiuser videoconferencing (RT) | 3 up to 10 Mbps |
| **Class 4** | High speed Media Download (NRT) | 100 Mbps Up to 1Gbps |

**Table 1: Service Classes**

## IV. HYBRID FUZZY LOGIC CONTROL

Fuzzy logic (FL) was initiated in 1965 by LA Zadeh of the University of California to model problems which dealt with imprecise data. FL techniques are either deterministic (ignores uncertainties) or probabilistic (all uncertainties represented as a probability distribution) [9]. In fields such as electronics and robotics [10], fuzzy logic (FL) has been very successful in dealing with uncertainties and they have consequently been considered in cellular networking as a promising approach to improve CAC processes in cellular networks.

The term 'hybrid FLC' refers to a fuzzy controller that has type-1 fuzzy sets (FSs) at the consequents and type-2 fuzzy sets at the antecedents. It was shown that no matter the efforts made to improve the accuracy of measurements of loss and latency, there still exist some imprecisions [5]. The FLC therefore needs to be able to capture and manage these uncertainties by further fuzzifying the MFs at the input of the fuzzy controller. It was not possible to capture and deal with the inaccuracies of measurements using a type-1 FLC. While type-1 associates meaning and membership grades to specific values of packet loss and latency, type-2 does also capture the imprecisions and variations in meaning of these parameters. The type-2 FSs for packet loss and latency are shown in the tables below.

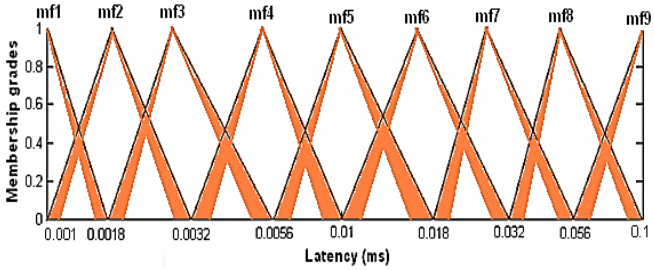



**Figure 4: Type-2 MFs for Latency**

The nine MFs for latency are: very low, low, slightly low, lower medium, medium, upper medium, slightly high, high, and very high.

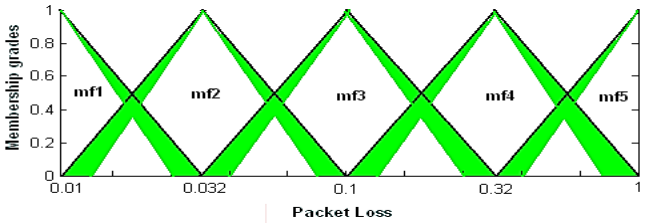



**Figure 5: Type-2 MFs for Packet Loss**

The five MFs for packet loss are: very low, low, medium, high, and very high.

The system's output, spectral efficiency (SE) in bps/Hz, is determined by a set of rules. The lower the packet loss and latency, the greater the throughput capability of the channel. The first 9 inference rules are defined in the table below:

| Rule | Loss | Latency | SE (bps/Hz) |
|---|---|---|---|
| 1 | Very low | Very low | 30 |
| 2 | Very low | Low | 27.57 |
| 3 | Very low | Slightly low | 25.13 |
| 4 | Very low | Lower medium | 22.7 |
| 5 | Very low | Medium | 20.27 |
| 6 | Very low | Upper medium | 17.83 |
| 7 | Very low | Slightly high | 15.4 |
| 8 | Very low | High | 12.97 |
| 9 | Very low | Very high | 10.53 |

**Table 2: First 9 inference rules**

## V. RESULTS

The set of results obtained below highlights the benefits

of properly dealing with the variations of the key admission factors. Calls are admitted to the network based on the quality of the available channels and when they do not risk degrading the quality of ongoing calls.

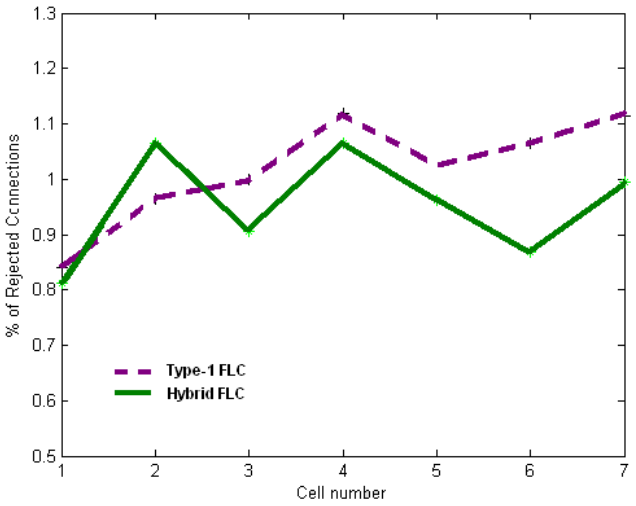



**Figure 6: Type-1 FLC vs. Hybrid FLC Performance**

The dotted curve depicts the results obtained using a type-1 controller, and the Hybrid controller's output is represented with the solid lines. A summary of the performance of type-1 and Hybrid FLCs is shown in the table below:

| CAC Technique | % of Rejected Calls | % of Rejected Handoffs |
|---|---|---|
| Type-1 FLC | 0.9626% | 0.0350% |
| Hybrid FLC | 0.9532% | 0.0262% |

**Table 3: Summary of Results**

Both techniques show that the call rejection rate is kept below 1%. However, it is desirable to nullify the call dropping rate because of the sensitivity of ongoing calls from the user's perspective [3]. The Hybrid FLC reduces the call dropping rate by almost 25%. This results in higher levels of Quality of User's Experience (QoE).

## VI. CONCLUSION

Type-1 fuzzy sets interpret the meaning of a particular event *x* in the admission process by giving it a degree of truth such that $\mu_x \in [0;1]$. However, they do not capture the possible variations in the meaning of these parameters. The introduction of type-2 fuzzy sets enables the system to better manage the imprecisions associated with the key factors of the connection admission process. The use of hybrid FLC yields a reduced call dropping probability and a lower call blocking rate. Future works include a deeper analysis of the use of type-2 fuzzy sets at the antecedents and the consequents of the controller.

## VII. REFERENCES


[1] CB Tokpo Ovengalt, K Djouani and AM Kurien, "A fuzzy approach for call admission control in LTE networks", 5th International Conference on Ambient Systems, Networks and Technologies (ANT-2014), published by Elsevier B. V., June 2014

[2] P Aloo, K Djouani, B Van Wyk and MO Odhiambo, "Fuzzy Logic Based Dynamic Pricing Scheme for Provision of QoS in Cellular Networks", pp1-8, Wireless Information Networks and Systems, July 2010

[3] N Boumella and K Djouani, "A Type-II Fuzzy Logic Decision System for Call Admission Control in Next Generation Mobile Networks", pp1-6, Global Telecommunications Conferences (Globecom 2010), IEEE 2010

[4] M Mathis, J Semke and J Mahdavi, "The Macroscopic Behavior of the TCP Congestion Avoidance Algorithm", pp67-82, ACM SIGCOMM Computer Communication Review, Vol. 27, Issue 3, July 1997

[5] 3GPP TS 23.203 V11.6.0-Technical Specification, "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Policy and charging control architecture", Release 11, pp38-40, June 2012

[6] J Sommers, P Barford, N Duffield and A Ron, "Improving Accuracy in End-to-end Packet Los Measurement", pp157-168, Conference on Applications, technologies, architectures, and protocols for computer communications, SIGCOMM '05, 2005

[7] RJ Gibbens, FP Kelly and PB Key, "A Decision-Theoretic Approach to Call Admission Control in ATM Networks", pp1101-13, IEEE Journal on Selected Areas in Communications, vol. 13, No. 6, August 1995

[8] A Sadeghian, JM Mendel, H Tahayori, "Advances in Type-2 Fuzzy Sets and Systems- Theory And Applications" pp68-73 ,Springer, 2013

[9] H Hagras, "General Type-2 Fuzzy Logic Systems to enable Better Uncertainty Handling for Real World Application", The University of Essex, England, UK, 2011

[10] C Wagner and H Hagras, "Toward General Type-2 Fuzzy Logic Systems Based on zSlices", pp637-660, IEEE Transactions on Fuzzy Systems, Vol. 18, NO. 4, August 2010



**Christophe B. Tokpo Ovengalt** received his BTECH degree in 2011 from the Tshwane University of Technology (TUT) and is presently studying towards an MTech degree in Electrical Engineering at TUT and jointly with an MSc degree through F'SATI at TUT. His research interests include fuzzy logic controllers, and wireless networks.

# Rainfall Cell Estimation and Attenuation Studies for Radio links at Subtropical Africa

Akintunde A. Alonge[1] (*Grad Student Member, IEEE*) and Thomas J. Afullo[2] (*Senior Member, SAIEE*)
Department of Electrical, Electronics and Computer Engineering
University of KwaZulu-Natal, Private Bag X54001, Durban
email: dtanthony740@gmail.com[1], Afullot@ukzn.ac.za[2]

**Abstract—The determination of path attenuation due to precipitation – is an important factor in the design of wireless networks – operating at microwave and millimeter bands. The rain cell area plays an important role in the estimation of path attenuation over radio links. In this study, rainfall events are retrieved from distrometer measurements installed at subtropical Durban, South Africa. The advection velocities of moving rain clouds undergoing Birth-Death (BD) scenarios are applied to translate rainfall time series into Rain Cell Diameter (RCD). An empirical length factor model is derived from the collected data, and then applied to predict rainfall attenuation. The proposed model is compared with terrestrial microwave link measurements from 19.5 GHz, 6.73 km horizontally-polarized system in Durban. It is found that the model compares well with link measurements, as well as, the International Telecommunication Union (ITU) estimates.**

**Index Terms—Rain rate; rain cell diameter; rainfall path attenuation; microwave links**

## I. INTRODUCTION

Rain fade is a contemporary problem associated with wireless communication networks (both terrestrial and satellite) as they evolve to higher frequency spectrum to enable the provisioning of robust services to end-users at over larger bandwidth performance. While signal degradation of microwave and millimeter wave become noticeable at 10 GHz especially during rainfall, rain fades are found to account for a large percentage of time-domain path losses in radio links [1], [2]. For a radio link budget designer of microwave or millimeter wave systems, the main objective is often to enhance the efficiency and effectiveness of allocated radio resources to reduce and/or eliminate outage times. Pertaining to this, the ITU-R has specified logical steps necessary to estimate rain attenuation for terrestrial and earth-space communication [3], [4]. These methods give consideration to rainfall rate, and hence, rain attenuation occurring at 99.99% probability over the year known as $R_{0.01}$. There are studies indicating the insufficiency of ITU-R techniques, especially at tropical and subtropical areas. At these locations, the rainfall process may be sudden, unpredictable and disadvantageous to wireless networks as they increase their outage probability.

An important aspect of rainfall attenuation studies is the presence of rain cells over radio links. Rain cell is a term generally describing the planar area occupied by rain droplets, within the transmission path of a radio link. The determination of rain cells have been researched in a number of studies [5]-[9]. In this current approach, a simplistic approach involving Birth-Death (BD) technique of rainfall spikes is applied. For simplicity, the arrival and departure of rain spikes are associated with rain cell motion over a radio link. The approximate cell shape assumed in this paper is circular, with the diameter as the main parameter. This was initiated by experimenting with rainfall rate data obtained from the JW distrometer collected from Durban, South Africa over two years.

## II. OVERVIEW OF RAIN CELL THEORY

It is a common knowledge that rain cells are formed when passing rain clouds produce rainfall (in droplrt form) which often intercept radio links [5]-[9]. When a communication link is intercepted, especially at microwave and millimeter wave, there is always an outage probability due to rainfall attenuation [5]. Hence, rain cells play a big role in the determination of rain attenuation. Mathematically, the Rain Cell Diameter (RCD) is given as, $D_c$, at any given location is represented by a negative power-law function given by:

$$D_c = \varphi R^{-\beta} \quad [km] \quad \text{for} \quad \alpha, \beta > 0 \qquad (1)$$

where value $\varphi$ and β are power-law coefficients related to the (1) with rainfall rate of $R$.

In a typical event, rain clouds are propelled by advection velocities which allow them to move from one point to another within a communication link. When seen from space borne radar images, rain cells usually appear as shadowed circular patterns defining radar signatures. As explained by [8], different signatures and resolutions of rain cells symmetry can be obtained by using different target frequency bands. The general consensus of rain cell shape approximation as being circular is noticeable when viewed from the radar, and is accepted by a number of researchers [9],[11],[12]. Equally, researchers have proposed other complex shape approximations with improved rain cell geometries, such as EXCELL and HYCELL [9], [10]. While space-bourne radar provide good image correlation, the use of ground rainfall data may be useful in ascertaining time constraints and progression associated with moving rainfall cells [13]. In this case, a rain gauge, distrometer or ground radar may be required [11], [14]. Radars can also be space borne as an embedded part of a satellite framework; this scenario allows for better correlation of the underlying shapes from space.

In this study, a ground instrument such as the impact distrometer is utilized. Thus, the generation of the cell diameter can be derived from the rainfall time series data. The distance gained by arriving rain rates, within the travelling rain clouds, towards the observation point is equivalent to the RCD. This diameter, $D_c$, is estimated from the advection velocity, $V_a$, assigned to the moving cloud by translation [6], [7], [14], [15]:

$$D_c = 0.06 V_a t \ [km] \quad \text{for } t \in T \qquad (2)$$

where $t$ is the successive point along each spike as they appear in the time series.

The eventual diameter corresponding to the rainfall rate in the time domain is estimated as a reverse cumulative sum. The advection velocity is assigned in (2) based on the Stratiform-Convective (S-C) rain rate bound, $R_{th}$, identified at Durban:

$$V_a = \begin{cases} 6\,m/s & for\ R \le R_{th} \\ 10\,m/s & for\ R > R_{th} \end{cases} \quad (3)$$

The typical value of $R_{th}$ can be found using radar reflectivity algorithms of Gamache and Houze [16]. This algorithm ensures that the S-C bound at any region, from rainfall measurements, is exactly at 38 dBZ. At Durban, $R_{th}$ is found to correspond to 11.34 mm/h over the entire rainfall period [17].

## III. MEASUREMENTS AND DATA PROCESSING

Rainfall Samples were taken from Joss-Waldvogel RD-80 disdrometer measurements installed at the University of KwaZulu-Natal, Howard Campus, Durban. The distrometer is an impact-actuated equipment with an outdoor and indoor unit, all synchronized with an archival system. The system has a sampling time of one minute and instantly logs the collected data at this sampling instance. The computational error associated with this device is about ±5%, with the output aggregated over twenty diameter channels. The collected samples consist of number rainfall sampled over a period of two years (January 2009 and December 2012). Only rain events with maximum rain rate of 10 mm/h are considered in this study. The highest rain rate for any event recorded within this period was 117.15 mm/h. The collected data are reclassified into two distinct groups: shower (10 mm/h $\le R <$ 40 mmh) and thunderstorm ($R \ge$ 40 mm/h) rainfall classes. These two classes are considered because they both represent the largest spectrum of rain rates available during intense rainfall. A total of 229 and 230 samples were eventually processed from the various shower and thunderstorm rain events at Durban.

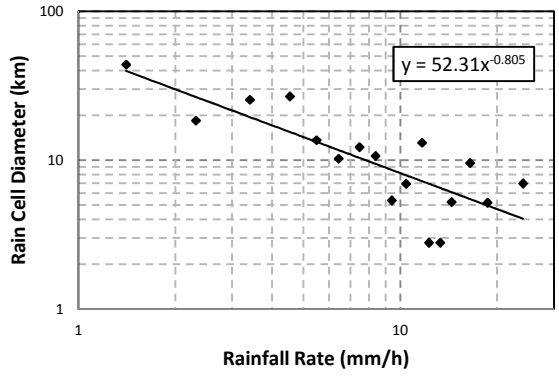
## IV. RESULTS AND DISCUSSIONS

The measurements from the disdrometer database are used to generate the RCDs corresponding to each rain event, and thereafter, the length factor is computed to estimate the rain attenuation. By applying the equations given in (2) and (3), the RCD at Durban for shower and thunderstorm regimes are estimated. On computing the RCDs, a data sorting is undertaken in a descending order according to their rainfall rates. The grouped data is averaged over similar rain rates, and the mean RCD is obtained from the data. In this way, similar mean rain rates and related RCDs are retrieved from the data. This procedure is repeated for samples according to the proposed rainfall rate classes. A relationship between rainfall rate and the RCD is established for data categories by applying regression technique. The results from the regression analysis are shown in Figs. 1*a* and 1*b*. It is found that slightly different relationships exist for the RCD relationships under these two scenarios. These relationships are given by:

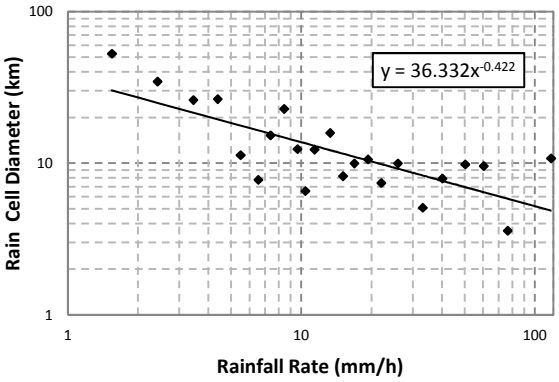
$$D_c = 52.31R^{-0.8}\ [km]\ \text{ for } 10 \le R < 40 \quad (4)$$

$$D_c = 36.33R^{-0.42}\ [km]\ \text{for } R \ge 40 \quad (5)$$

The fitting errors accompanying the regression technique is computed by applying the Root-Mean-Square Error (RMSE)




(*a*)




(*b*)

Figure 1. Regression fitting technique of collected samples for RCD dependence on rainfall rate at Durban: (a) Shower events (b) Thunderstorm events.

statistic. This is given by:

$$RMSE = \left(\frac{1}{n}\sum_{i=1}^{n}(o_i - m_i)^2\right)^{1/2} \quad (6)$$

where $n$ is the total number of samples, $o_i$ and $m_i$ are the observed data and model respectively.

From the comparison between the model and the measurements, RMS errors of 4.49 and 6.64 are computed for shower and thunderstorm regimes respectively. The RCD models at both sites obviously have different power-law coefficients. For instance, the RCD for shower events is observed to decline rapidly with rain rate. Thus, it is expected that smaller mean cell diameters occur during shower events compared to thunderstorm events, as the mean rain rate approaches maximum bound i.e. 40 mm/h. As seen in Fig. 2, the proposed models are compared with the Bryant *et. al* model [5] for tropical regions. It is observed that the thunderstorm RCD circular model gives the closest approximation compared to the tropical model. The RCD for shower events, on the other hand, indeed decays rapidly than the thunderstorm event.

Since the thunderstorm model compares well with rainfall rate spectrum categorized above 100 mm/h, it is desirable to adopt it as a general function for RCD at Durban. However, the thunderstorm model is seen to intercept the Bryant *et. al* tropical model close to 20 mm/h. Thus, in comparison with the Bryant *et al.* model, the thunderstorm model has lower RCD for rain rates lower than 20 mm/h and vice-versa. In an earlier investigation of RCDs over Durban, the general model obtained by Akuon and Afullo [6] is closer

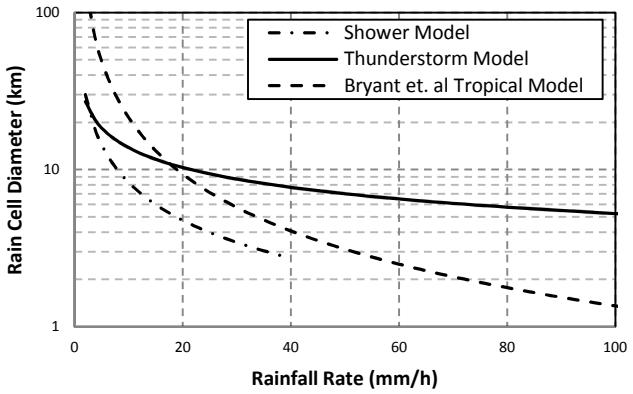



Figure 2. Comparison of the proposed RCD at Durban for different events at Durban with the Bryant *et. al* model.

to the thunderstorm model. This implies that the thunderstorm model is the best representation of the RCD in Durban.

### A. Computation of Length Factor

The *reduction factor*, now known as the *length factor,* in the ITU-R P.530-15 recommendation is an important parameter utilized for estimating path attenuation due to rain [3]. In this recent recommendation of ITU, the length factor is dependent on four parameters namely: frequency, rainfall rate, radio link distance and specific attenuation coefficient. This function is given by:

$$r = \left[0.477 d^{0.633} {R_{0.01}}^{0.073\alpha} f^{0.123} - 10.579(1 - exp(-0.024d))\right]^{-1} \quad (7)$$

where $d$ is the radio link distance, $R_{0.01}$ is the rain rate exceeded at 0.01% of the time, $f$ is the transmission frequency of the link and α is the exponent of the specific attenuation function.

In link budget planning, distance limitations are often encountered in microwave link systems for line-of-sight designs. Thus, it is usual to put into cognizance, the degrading Fresnel conditions, in the estimation of the length factor. It follows from the derivation of rain cell sizes that the length factor can be implemented. Based on this study, the empirical length factor is dependent on the rain cell size and link distance. This is given by:

$$r = \frac{\eta D_c}{L + D_c} \quad (8a)$$

where,

$$\eta = \xi_1 L^{-\xi_2} \quad \text{for} \quad L < 50\,km \quad (8b)$$

$$\xi_1 = a_1 R^{-b_2} \quad \text{and} \quad \xi_2 = a_2 R^{-b_2} \quad (8c)$$

where $L$ and $R$ are the link path length and rainfall rate respectively.

Unlike the current ITU-R length factor function in (7), the proposed length factor from our study is independent of transmission frequency. These results were fitted for link distances up to 50 km, since this is the upper bound distance required for optimal microwave link performance, in the Fresnel region. The obtained results for each classified regimes in Durban fitted for rain rates above 2 mm/h and link distances greater than 5 km are presented in Table I.

### B. Determination of Path Attenuation due to Rain

The path attenuation over a radio link is given as:

$$A_p = r\gamma L \quad [dB] \quad (9)$$

where the parameter γ is the specific attenuation due to rain in dB/km. Other parameters related to (9) have been defined accordingly.

The specific attenuation computed at Durban has been undertaken for four regimes in an earlier study [18] using the rain Drop Size Distribution (DSD) measurements. For this study, the modified gamma DSD model is applied, so that:

$$N(D_i) = N_m {D_i}^{\mu} exp(-\Lambda D_i) \quad [m^{-3}mm^{-1}] \quad (10)$$

where $N_m$ is the droplet concentration, $D_i$ is the diameter of the rain drops – ranging from 0.35 mm to 5.3 mm and Λ is the slope parameter.

TABLE I. PARAMETERS OF THE PATH REDUCTION FACTOR FOR SHOWER AND THUNDERSTORM REGIMES IN DURBAN [18]

| **RAIN REGIMES** | $r = \xi_1 L^{-\xi_2}$ **for 5 km ≤ $L$ ≤ 50 km** | |
|---|---|---|
| SHOWER | $\xi_1 = 10.89R^{-0.49}$ | $\xi_2 = 0.59R^{-0.33}$ |
| THUNDERSTORM | $\xi_1 = 5.65R^{-0.15}$ | $\xi_2 = 0.45R^{-0.12}$ |

Since the behaviour of RCDs under thunderstorm conditions is of major concern here, only DSD parameters for this regime is given from [16] as μ = 2, $N_m = 5233R^{0.0681}$ and $\Lambda = 4.242R^{0.141}$. By applying Mie scattering technique, the specific attenuation due to precipitation, over the entire diametric spectrum of rain drops is computed from [13] as:

$$\gamma = 4.343 \times 10^{-3} \sum_{i=1}^{20} N(D_I) Q_{ext}(D_i) \Delta D_i \quad [dB/km] \quad [11]$$

where $N(D_i)$ is the rain DSD, $Q_{ext}(D_i)$ is the extinction cross section in mm$^2$ and $\Delta D_i$ is the diameter interval in mm. The Mie technique is dependent on the ambient temperature at Durban which is assumed as 20°C in this study [19].

The predictions are compared with the terrestrial link measurements undertaken at Durban in 2004. The link transmits over a path length of 6.73 km, while operating at a transmission frequency of 19.5 GHz. Details and other necessary information related to the measurement are available in [20]. The plots for this comparison are presented in Figure 3 and Table II. The prediction from the proposed model is found to majorly approximate rainfall path attenuation over the average value of the link measurements. In addition, it is observed at this frequency that path attenuations over the entire rainfall rates are higher than those predicted by the ITU model. Broadly speaking, the generated results from the proposed length factor are projected to increase as higher microwave frequencies and intense rainfall rates are encountered along any hypothetical microwave link. This is mainly as a result of the presence of multiple rain cells along the link especially as rainfall gets intense during a typical rainfall event. It is envisaged this problem can be resolved by using the BD characteristics to

TABLE II. COMPARISON OF PROPOSED MODEL WITH LINK MEASUREMENTS AND ITU-R MODEL AT DURBAN

| RAINFALL RATE (mm/h) | LENGTH FACTOR (proposed) | LENGTH FACTOR (ITU-R) | PREDICTED ATTENUATION (dB) | ITU-R ATTENUATION (dB) | LINK MEASUREMENT (dB) | | |
|---|---|---|---|---|---|---|---|
| | | | | | Min | Ave | Max |
| 1.00 | 2.04 | 1.39 | 0.87 | 0.81 | 0.38 | 1.20 | 3.40 |
| 2.60 | 1.80 | 1.11 | 2.17 | 1.79 | 0.80 | 2.50 | 5.40 |
| 3.00 | 1.76 | 1.08 | 2.48 | 2.02 | 0.90 | 2.80 | 5.60 |
| 4.20 | 1.67 | 1.01 | 3.38 | 2.69 | 1.40 | 4.00 | 6.10 |
| 5.80 | 1.58 | 0.95 | 4.53 | 3.56 | 2.50 | 5.10 | 6.30 |
| 10.00 | 1.42 | 0.86 | 7.34 | 5.74 | 4.90 | 8.60 | 16.30 |
| 15.00 | 1.29 | 0.80 | 10.39 | 8.21 | 5.90 | 13.00 | 23.00 |
| 21.00 | 1.19 | 0.75 | 13.75 | 11.08 | 6.80 | 16.20 | 27.20 |
| 30.00 | 1.09 | 0.71 | 18.37 | 15.26 | 7.00 | 16.40 | 28.50 |
| 51.00 | 0.93 | 0.65 | 27.79 | 24.63 | 7.10 | 19.20 | 32.00 |
| 79.00 | 0.81 | 0.61 | 38.46 | 36.65 | 12.00 | 20.00 | 34.00 |

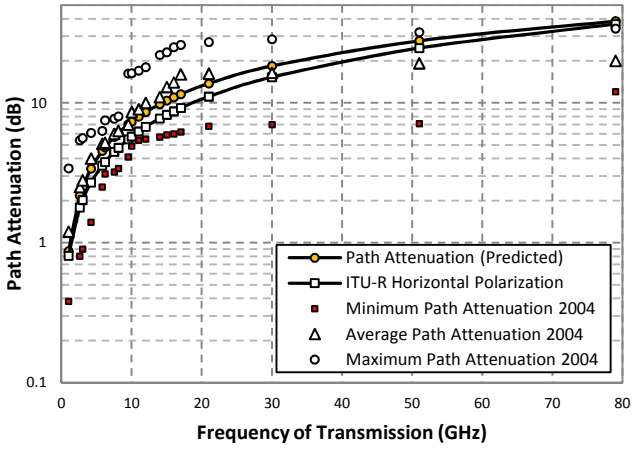



Figure 3. Comparison of predicted attenuation with measurements at 19.5 GHz over a 6.73 horizontally-polarized microwave link.

estimate the maximum factor required to compensate for the actual number of multiple rain cells existing within a typical link. The overall path length of a radio link is important in the determination of the optimal number of spikes, and hence rain appropriate length factor, at any location. Thus, it is expected in future studies that the existence of multiple cells be incorporated in the proposed model

## V. CONCLUSION

The estimation of RCD from ground measurements is an interesting approach to understanding the dynamics of rainfall attenuation. Although, there are certain limitations due to tracking and detection of multiple cells, it is obvious that the proposed circular approximation is encouraging. The effect of the multiple cells, over wireless links, using BD technique is a subject of future research. It is noteworthy that the approach applied in this study can be applied at other locations and compared with the ITU-R standards for validation. This study will assist in the planning of radio links, more so, as the demand for higher carrier frequencies vis-à-vis achieving high data-rate efficiencies is critical to service delivery.

## REFERENCES


[1] F.J.A. Andrade, L.A.R. da Silva Mello and M.S. Pontes, "Statistical Modeling of Rain Attenuation on Tropical Terrestrial Links," *J. of Microw, Opt. and Electromagnetic App*., Vol. 11, No. 2, pp. 296 – 303, 2012.

[2] J.S. Mandeep and J.E. Allnut, "Rain Attenuation Predictions at KU-Band in South East Asia Countries," *Progress in Electromagnetic Research*, PIER 76, PP. 65 – 74, 2007.

[3] ITU-R, *Propagation Data and Prediction Methods for the Design of Terrestrial Line-of-Sight Systems*, ITU-R Rec. P.530-15, Geneva, 2013.

[4] ITU-R, *Specific Attenuation Model for Rain for use in Prediction Methods*, ITU-R Rec. P.838-3, Geneva, 2005.

[5] G.H. Bryant, I. Adimula, C. Riva and G. Brussard, "Rain Attenuation Statistics from Rain Cell Diameters and Heights," *Int. J. Sat Commun.*, Vol. 19, pp. 263-283, 2001.

[6] P.O. Akuon and T.J.O Afullo, "Rain Cell Sizing for the Design of High Capacity Radio Link Systems in South Africa," *Progress in Electromagnetic Research B*, Vol. 35, pp. 263 - 285, 2011.

[7] P.O. Akuon and T.J.O. Afullo, "Rain Cell Size Statistics from Rain Gauge Data for Site Diversity Planning and Attenuation Prediction," *Proceedings of SATNAC 2011 Conference*, East London Convention Centre, East London, 4-7, pp. 213-216, September 2011.

[8] C. Melsheimer, W. Alpers and M. Gade, "Simultaeneous Observation of Rain Cells over the Ocean by the Synthetic Aperture Radar aboard the ERS Satellites and by Surface-based Weather Radars," *J. of Geophysical Research*, Vol. 106, No C3, pp. 4665 – 4677, March 2001.

[9] F. Féral, H. Sauvageot, L. Castanet and J. Lemorton, "HYCELL—A new hybrid model of the rain horizontal distribution for propagation studies: 1. Modeling of the rain cell," *Radio Sci.*, Vol 38, No 3, pp.1056, 2003.

[10] C. Capsoni, F. Fedi, C. Magistroni, A. Paraboni and A. Pawlina, "Data and theory for a new model of the horizontal structure of rain cells for propagation applications," *Radio Sci.*, Vol. 22, No 3, pp. 395-404, 1987.

[11] M. Karklinsky and E. Morin, "Spatial characteristics of radar-derived convective rain cells over Southern Israel," *Meteorologische Zeitschrift*, Vol. 15, No. 5, pp. 513 – 520, October 2006.

[12] M.S. Pontes, L.A.R. Da Silva Mello and R.S.L. de Souza, "Modelling of Effective Rainfall Rate and Rain Attenuation in Terrestrial Links in the Tropics," *6th International Conference on Information, Communication and Signal Processing*, Singapore, December 2007, pp. 1 – 4.

[13] G.O. Ajayi, S. Feng, S.M. Radicella, B.M. Reddy, *Handbook on Radiopropagation Related to Satellite Communications in Tropical and Subtropical Countries*, ICTP, 1996, pp. 7–14.

[14] S. Begum, C. Nagaraja, I. Otung, "Analysis of Rain Cell Size Distribution for Application in site Diversity," *IEEE Trans. Antennas Propag.*, Oct. 2006, pp. 1-5.

[15] A. Pawlina, "No rain intervals within rain events: some statistics based on Milano radar and rain-gauge data," *COST Action 280*, 1st International Workshop, July 2002.

[16] J.F. Gamache and R.A. Houze, "Mesocale Air Motions associated with Tropical Squall Line," *Monthly Weather Review*, 110, pp. 118-135, 1982.

[17] A.A. Alonge and Afullo, T.J., "Rainfall Microstructural Analysis for Microwave Networks: Comparison at Equatorial and Subtropical Africa", *Progress in Electromagnetic Research B*, Vol. 59, 2014.

[18] A.A. Alonge and Afullo, T.J., "Regime Analysis of Rainfall Drop-size Distribution Models for Microwave Terrestrial Network", *IET Microwave Antennas Propag.*, Vol. 6, issue 4, pp. 393-403, 2012.

[19] A.A. Alonge, *Correlation of Raindrop Size Distribution with Rain Rate Derived from Disdrometers and Rain Gauge Networks in Southern Africa*, M.Sc dissertation submitted to the University of KwaZulu-Natal, Durban, December 2011.

[20] K. Naicker and S.H. Mneney, "Propagation measurements and modeling for terrestrial line-of-sight links at 19.5 GHz," *Proc. of IEEE AFRICON conference*, vol. 01, pp. 95−100, Sept. 2004.

**Akintunde A. Alonge** received his M.Sc [Eng] degree from the University of KwaZulu-Natal, Durban, South Africa (2012) and B.Eng (Hons) at the Federal University of Technology, Akure, Nigeria (2007). He is currently working towards his Ph.D at UKZN. His research interests include wireless communication systems, system design for Satellite & Terrestrial networks and signal processing techniques.

**Thomas J. Afullo** received the B.Sc (Hon) Electrical Engineering from the University of Nairobi, Kenya (1979), the MSEE from West Virginia University, USA (1983), and the Bijzondre License in Technology and Ph.D in Electrical Engineering from the Vrije Universiteit Brussel (VUB), Belgium (1989). He is currently Professor of Microwave Engineering at the Department of Electrical, Electronic & Computer Engineering, University of KwaZulu-Natal (UKZN), Durban.

# Prediction of Time-series Rain Attenuation based on Rain Rate using Synthetic Storm Techniques over a Subtropical Region

Joseph S. Ojo[1] and Pius A. Owolawi[2]
Department of Electrical Engineering
Mangosuthu University of Technology, P. O. Box 12363, Jacobs 4026 Durban
Tel: +27 616507924, Fax: (031) 90 728932
email: *josnno@yahoo.com[1]; owolawi@mut.co.za[2]

**Abstract-** Research into fade mitigation techniques continue to be of interest globally as more about wireless millimeter wave communication services operating at high frequency bands expanded daily. At such higher microwave frequencies (> 10 GHz), rain attenuation is the dominant propagation impairment in the quality of radio communication signals. In this work, we present results of time-series rain-induced attenuation simulated based on ten years of concurrent time-series rain rate measurement using a synthetic storm techniques (SST) in the Ku and Ka-frequency bands in Durban, a subtropical location in South Africa. Results are presented based on event-wise, cumulative distribution of one minute rain rate, time-series seasonal variation of rain rate, diurnal fades margin and diurnal variation of rain attenuation in order to obtain the optimum thresholds. Comparison of syntheisized rain attenuation results with experimental data on slant path Earth-satellite links shows that the SST model provides similar results close to the measured attenuation and can thus be applied to estimate rain attenuation in this region. The result further shows that the fade margins to be implemented vary from season to season. Therefore, applying the same fade margin during each of the seasons may either underestimate or overestimate the required values needed in this region.

**Index Terms—** Synthetic storm techniques, Time-series rain attenuation, Subtropical region

## I. INTRODUCTION

In the recent time, satellite services have become the order of the day in our daily activities in the area of direct to home services (dth-services), tele-medicine, tele-education; Internet and banking services, telephony services, conferencing to mention but few. These services are only available at higher frequency bands such as Ku (12-18), Ka-band (20-30), and other higher frequency bands owing to the ability to provide larger transmission bandwidth and higher data rate. However, rain attenuation is the dominant propagation impairment in the quality of radio communication signals operating at frequencies higher than 10 GHz [1]. The severity of the impairment depends on the rainfall intensity that varies with regional locations among others and this grossly affects the quality of service (QoS) to be delivered by satellite operators [2]. The severities are even more pronounced at the tropical, subtropical and equatorial region, where intense rainfall events are common as compared with the temperate region as earlier reported by [1, 3 - 4]. A subtropical rainfall differs from tropical rainfall in terms of storm; the tropical storms generate more rain compared to subtropical storms with typically produced heavy thunderstorm activity in the subtropical when compared to a tropical storm. Rainfall is more of a stratiform type in the subtropical region when compared to the tropical region which is more of convective type. It is of key importance, therefore, to examine rain attenuation, not only to provide additional information for the improvement of rain attenuation models, but also to determine the level of degradation arises owing to the subtropical climates in terms of fade mitigation in EHF communication systems that may be useful for system engineers.

In the recent times and due to lack of experimental data for rain attenuation for most locations in the tropical and subtropical region, SST has been identified to be a powerful and accurate tool that can produce all the necessary statistics of rain attenuation based on the time series of rain rate $R$ (mm/h) at the site, at any frequency and polarization, and for any slant path above about 10° whenever a real radio link is not available [5]. It therefore shows that whenever there is information for time-series prediction of rain attenuation, fade countermeasure techniques such as adaptive control of signal power, coding and data rate may be implemented effectively.

In this paper, we use the rainfall rate time series measured in Durban, South Africa to provide reliable attenuation time-series of Ku- and Ka band signals on a hypothetical satellite link towards the Intelsat 17 (IS-17) Satellite (Geostationary at 66$^o$E) with one of its service footprint links in the study area using SST. An attempt is also made to test the validity of the result obtained based on the experimental data collected in the same region.

## II. STUDY LOCATION, DATA AND MEASUREMENT

The study location is Durban (altitude 8 m; 29$^o$ 97$^{'}$ S, 30$^o$ 95$^{'}$ E) with a mild and sub-tropical climate. The city belongs to the coastal Savannah region located along the Eastern frontiers of the Republic of Southern Africa. It has an average annual precipitation of over 1000 mm and is unarguably regarded as one of the wettest cities in South Africa. Durban experiences warm and wet summers and mild, moist to dry winters. The study, therefore attempts to establish local first order statistical rain attenuation using

SST to mitigate the severe fade experienced at higher frequency bands because of the huge presence of hydrometeors in this region.

Ten years (2002-2011) rain rate data of 5-minute integration were obtained from South Africa Weather Services (SAWS) for the study location. The most widely used method of data acquisition by SAWS is via a network of rain gauges. Rain gauges used are standard 127 mm in accordance with the World Meteorological Organization (WMO) standard. The 127 mm is the diameter of the rimmed circular funnel opening. The other kind of rain gauge used by SAWS is the automated rain gauge, which is a tipping bucket rain gauge with a 200 mm funnel opening. The number of drops collected every 10 seconds is electronically counted and is then averaged over 5-minute integration time. The Automatic Gain Control (AGC) voltage of each channel is sampled continuously and stored in digital mode, with the date and time of each tipping of the rain gauge. The rain gauge calibration is maintained by periodic cleaning of the capillary. Other details of measurement methodology in the present study are available in [6]. Since 1-minute integration, rain rate data is needed for radio propagation purposes, the 5-minute rain rate data obtained were converted using the Hybrid Polynomial-type model (HPT model) as detailed in [6]. HPT model is considered for conversion of rain rate from five-minute to one-minute equivalent of Durban with a general expression given as:

$$R_1(p_i) = aR_5(p_i)^2 + bR_5(p) + c \quad (1)$$

where $a$, $b$ and $c$ are constants with values $a$ = 0.0014, $b$ = 1.2021 and $c$ = -0.3543 for Durban.

### A. SYNTHETIC STORM TECHNOQUES

The synthetic storm techniques have as one of the major advantages of employing local time series with the assumption of dual-layer representation of precipitation to include the additional attenuation attributable to the melting layer. It may also reproduce dynamic characteristics of rain fade and power spectra. Studies show that SST has been applied to estimate multivariate probability distributions of rain attenuation simultaneously exceeded in distant sites, information that is useful to design satellite systems with a common on-board resources [7–8]. The value of rainfall as a function of the length where that rain moved on the line because of the wind with particular speed can as well be described using SST method. The application of the SST also depends on some assumptions taken in the formulation: the advection velocity of rain cell is assumed constant and in the direction of the projection of the link on the ground [9]. In other words, rain attenuation statistics on a hypothetical link passing through or near from a site can be determined by segmenting the links, with each segment to be equal to the travel distance from the rain structure as it is burnt-out by the wind during one sampling time of the measurement of the rainfall rate. During every sampling period, the attenuation by rain is estimated as the summation of specific attenuation (dB/km) multiplied by the length of the segment. This can be expressed as:

$$A_m = \sum_{j=0}^{n=1} kR_{m-j}^{\alpha} \Delta L_j \text{ dB}) \quad (2)$$

$$\Delta L = V_t \times t \quad \text{(km)} \quad (3)$$

where k and α are coefficients relating rainfall rate to the specific rain attenuation (dB/km) as estimated by [10]; R is the m-th sample of rain rate measurement, m is the number of segments making up the link; $\Delta L$ is the length of the m-th segment of the link, $V_t$ is the wind on a line and $t$ is the integration time, 1min. Detailed physical and mathematical fundamental are described at length by [7] and [8] and therefore not repeated here.

In order to estimate our time series rain attenuation using SST, we have used storm speeds, $v$= 12.5 m/s as obtained from meteorological measurements in order to show the sensitivity of this parameter to the SST model. The vertical structure of the precipitation medium has been modeled with two layers of different depths, layer A starts from the ground with the existence of rain (raindrops with a water temperature of 20°C) and layer B with melting hydrometeors at 0°C [9] as depicted in Fig. 1. $R_A$ is the homogeneously falling rain at layer A, while $R_B$ is the apparent rain rate of layer B dominated by ice, where

$$R_A = rR_B \text{ with } r = 3.134 \quad (4)$$

The parameters, necessary to relate the rainfall rate to the specific rain attenuation (dB/km) are as calculated from ITU 838-6 [11]. The recent ITU-R rain height has also been taken into account [12].

The input parameters needed by the SST model in our region are considered as follows. The altitude above sea level of the earth station is $H_S$= 0.008 km. According to [12] the height of the precipitation (rain and melting layer) above sea level, $H_B$ depends on the latitude (ϕ) of the Earth station and can be expressed as:

$$H_B = 5 \qquad \phi < 23^o$$
$$H_B = 5 - 0.075(\phi - 23^o) \qquad \phi \geq 23^o \quad (5)$$

From the simulation, our $H_B$ = 4.4825 km. Also, the thickness of the melting layer (h) is considered to be 0.4 km regardless of the latitude [5]. The Intelsat 17 (IS-17) Satellite (Geostationary at 66°E) with its service footprint links at an elevation angle of 38.4° is assumed.

Hence the specific attenuation at a given point is converted into signal attenuation, $A(x_0)$ for a satellite path using the following expression [7]:

$$A(x_0) = k_A \int_0^{L_A} R^{\alpha A}(x_0 + \Delta x_0, \xi) d\xi + k_B r^{\alpha B} \int_{L_A}^{L_B} R^{\alpha B}(x_0, \xi) d\xi \quad (6)$$

where ξ is the distance measured along the satellite path. According to [12] the height above sea level, $H_A$, of the upper limit of layer A is given by:

$H_A = H_B - h = 4.0825$ km. (7)

The radio path lengths are calculated using

$$L_A = \frac{H_A - H_S}{\sin(\theta)} = 6.5563 \text{ km} \quad (8)$$

and

$$L_B = \frac{H_B - H_S}{\sin(\theta)} = 7.2036 \text{ km} \quad (9)$$

where θ is the link elevation angle; $\Delta x_0$ is a shifting parameter that accounts for the fact that the radio path exists layer B at $x_0 + \Delta x_0$ and can be expressed as [5]

$$\Delta x_0 = \Delta L \cos(\theta) = \frac{h}{\tan(\theta)} \quad (10)$$

and

$$\Delta L = L_B - L_A = \frac{h}{\sin(\theta)} \quad (11)$$

Hence, our SST's are derived based on the above expressions by applying the Fourier transforms theory and taking into consideration some of the assumptions.

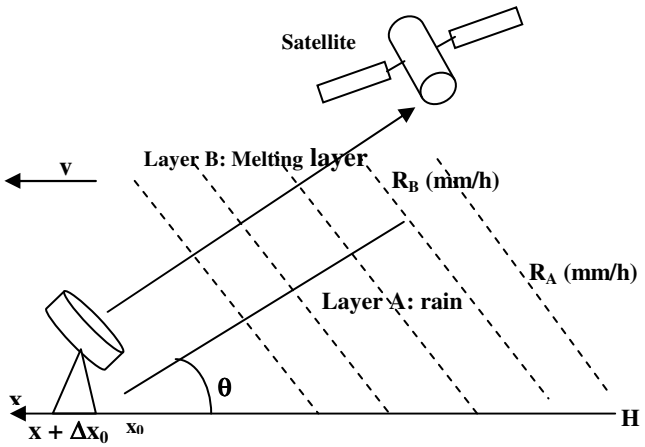



**Fig. 1**: Schematic diagram of rain structure for synthetic storm technique

### B. *PROPAGATION EXPERIMENTAL DATA*

Propagation measurements over an earth-space path have been data carried out in the year 2004 at the University of KwaZulu-Natal, Durban. The Ka-band modulated signal horizontally polarized at a frequency of 19.5 GHz level has been received by an offset feed valuline WR42/R220 parabolic antennae of diameter 0.6 m with an elevation angle of 30.980°. Detailed characteristics of the setup link are reported in the work of [13]. Simulated results were validated with the experimental data

## III. RESULTS AND DISCUSSION

Fig. 2 presents the cumulative distribution probability (CDP) of one-minute rain rate and ITU-R estimate values. A good fit could be observed at higher percentages of time with low rain rates (lower than 20 mm/h). However, considerable differences between the measured data and the ITU-R are obtained at lower percentages of time with large rain rates (greater than 40 mm/h). The results are presented to access the suitability of the ITU-R recommendations values of this region. Fig. 3 shows the mean monthly rainfall accumulation over the period of study. The location belongs to the coastal savannah region and experiences regular rainfalls all year long with the maximum rainfall accumulation during the summer (January) while the minimum average rainfall accumulation of about 24 mm was observed during the winter months. The variation in the monthly rainfall accumulation is as a result of the contrast in temperature between the warm south-flowing Mozambique-Augulhas current and the cold north-flowing Benguela. The analysis shows that the wettest months are between December and March. This is in agreement with the previous studies by [13-14] where they both showed that Durban (east coast) has higher rainfall than Lamberts Bay (west coast). This analysis is very important for estimating the link budgeting based on information on seasonal rain-pattern in this region.

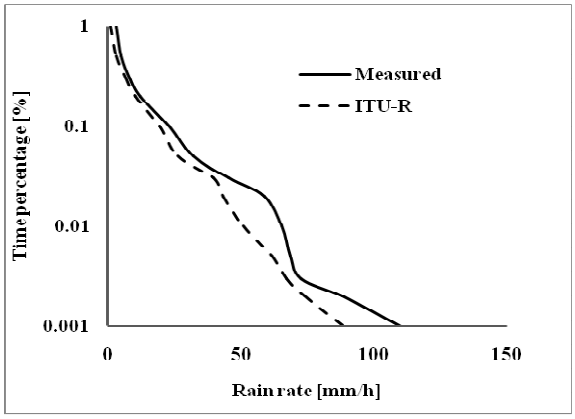



**Fig. 2**: CDP of one-minute rain rate

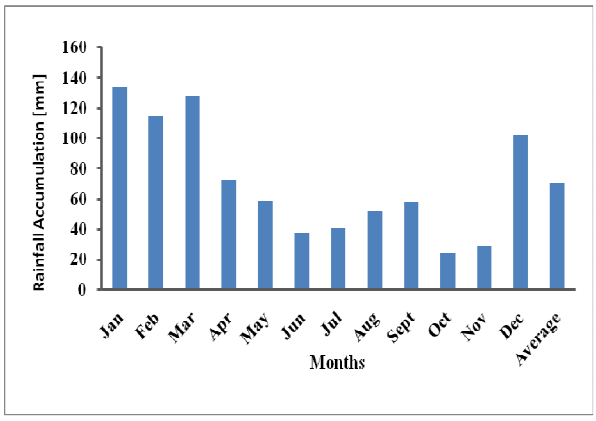



**Fig. 3**: Average monthly accumulation

We also made a comparison between the time series rain rate event with the synthesized attenuation (Ku and Ka-band frequencies) obtained from the rain event on 13th January, 2004 using an average storm speed of 12.5 m/s. The result is presented in Fig. 4. The measured rain rate time series and the rain attenuation time series using SST show strong correlation with similar pattern throughout the event.

Fig. 5 presents the comparison among the rain attenuation obtained using the SST model, measured data from our propagation campaign at Ka-band frequency of 19.5 GHz and the ITU-R 618-10 [15]. It could be observed that the SST model provides data very close to the measured attenuation, while the ITU recommendation underestimates the attenuation due to rain in this region. However, more year of propagation measurement is suggested to ascertain the level of the closeness of the predicted with measured values of attenuation using a statistical approach.

Fig. 6 also present the diurnal time-series rain attenuation (season-wise) at different exceedence probability. For each hour of the day, summer months suffered more of rain-

induced attenuation follows by autumn, winter and spring months present significantly less rain attenuation.

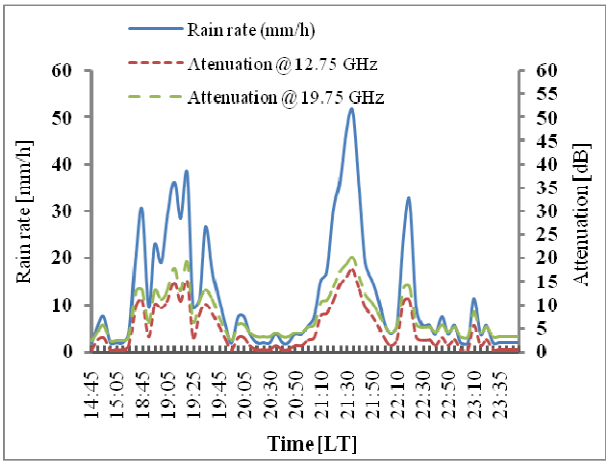



**Fig. 4:** Rain rate and predicted rain attenuation exceedence

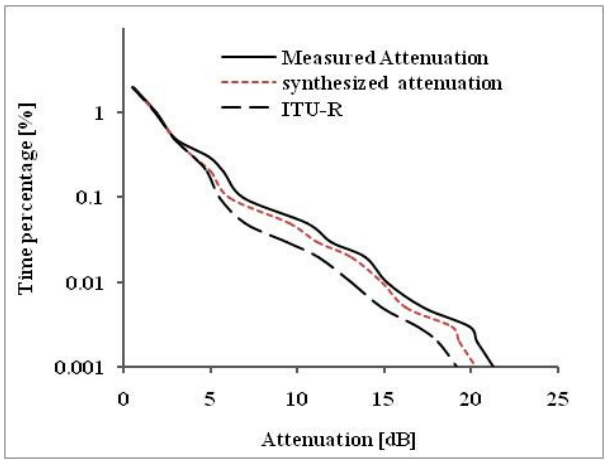



**Fig. 5:** Predicted slant path attenuation using SST and and measured during the year 2004 along Intelsat-IS-17 Line-of-sight at Ka-band frequency.

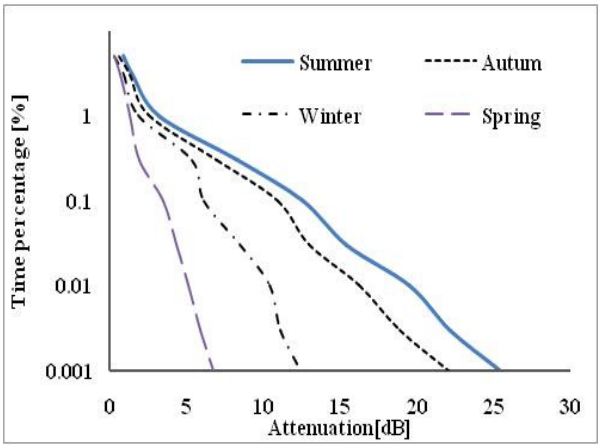



**Fig. 6**: Seasonal variation of synthesized rain attenuation

Table 1 summarizes the fade margins to be implemented as a function of the availability of time. We observe that the fade margin to be implemented vary from season to season. For instance, at 99.99% availability of time needed by most satellite transmission links as recommended by the ITU, the fade margin varies between 5.98 dB during the spring season to 25.40 dB in summer season with overall fade margins of 15.34 dB. Although, a fade margin of about 15.34 dB yields a link availability at 99.99% averaged over the season, the 99.9% link availability cannot be achieved during the summer season with only a provision of about 99.6%. That is, the 15.34 dB would have led to about 114 hrs of link outage during the summer seasons in the observed years rather than 29 hrs expected. This is a lot of large fade margin needed to be implemented at this high link availability hence, suitable fade mitigation techniques like site diversity, advanced forward error correction among others are needed to cater for larger than expected fading conditions [9].

**Table 1:** Summary of fade margin to be implemented for all the observed periods on a seasonal basis.

| Availability | Summer | Autumn | Winter | Spring |
|---|---|---|---|---|
| 99.99% | 25.40 | 18.78 | 11.20 | 5.98 |
| 99.95% | 22.15 | 17.90 | 11.10 | 5.91 |
| 99.90% | 12.50 | 10.90 | 6.10 | 3.45 |
| 99.50% | 8.25 | 6.98 | 5.34 | 1.90 |
| 99.00% | 3.20 | 2.50 | 1.80 | 1.30 |
| 95.00% | 1.50 | 1.30 | 0.90 | 0.65 |

## IV. CONCLUSION

Investigation of rain attenuation in subtropical region is very necessary, in order to provide additional information for the improvement of rain attenuation models and to determine the level of degradation arises due to the subtropical climates in terms of fade mitigation systems that can be useful by the system engineers. In this paper, rain attenuation time-series has been evaluated using synthetic storm based on long term measurement of the time-series rain rate. Comparison of measured rain rate data with recent ITU-R show considerable differences in lower percentages of time with large rain rates (greater than 40 mm/h). Fade margin to be implemented also vary from season to season. Therefore, applying same fade margin during each of the seasons may either underestimate or overestimate the required values. Application of our results to high link availability (99.99%) resulted into a large fade margin needed to be implemented; hence, a suitable fade mitigation techniques like site diversity, advanced forward error correction among others may be explored to cater for larger than expected fading conditions. We also observed that the SST model provides data very close to the measured attenuation and can therefore be used to estimate rain attenuation in this region

## REFERENCES

[1] Ojo, J.S., Ajewole, M.O. and Sarkar, S.K. (2008) "Rain Rate and Rain Attenuation Prediction for Satellite Communication in Ku and Ka Bands Over Nigeria", Progress in Electromagnetics Research B, Vol. 5, pp 207–223.

[2] Ajayi. G.O, Feng. S, Radicella, S. M, and Reddy, B. M (Ed), "Handbook on radio Propagation Related to Satellite Communication in Tropical and Subtropical Countries", ICTP, Trieste, Italy, pp 7-14, 1996

[3] Moupfounma F.: Model of Rainfall-rate Distribution for Radio System Design. IEEE Proceedings, Vol.132, Pt. H, No.1, pp. 39-43, Feb. 1985
[4] Ojo J.S and Omotosho T.V.: Comparison of 1-min rain rate derived from TRMM satellite data and rain gauge data for microwave applications in Nigeria" Journal of Atmospheric and Solar-Terrestrial Physics 102 pp. 17–25, 2013.
[5] Matricciani E. (2008) "Global formulation of the synthetic storm technique to calculate rain attenuation only from rain rate probability distributions", *Antenna and propagation symposium*, 2008.978-1-4244-2042-1/08/(c)2008 IEEE.
[6] Owolawi P.A (2011): Derivation of One-minute Rain Rate from Five-minute Equivalent for the Calculation of Rain Attenuation in South Africa. PIERS ONLINE, VOL. 7, NO. 6, 2011, pp. 524-535.
[7] Drufuca, G (1974) "Rain attenuation statistics for frequencies above 10 GHz from rain gauge observations", J. Resc Atmos., Vol. 1-2, pp 399-411.
[8] Matricciani E and Riva C. (2005): "The search for the most reliable long term rain attenuation CDF of a slant path and the impact on prediction models". *IEEE Trans Antennas Propag*, 53 3075–95, pp 207–223.
[9] Das D and Maitra A (2014): "Time series prediction of rain attenuation from rain rate measurement using the synthetic storm techniques for a tropical location", *Int. J. of Electronics and Communications,* 68, pp. 33-36.
[10] Maggiori D (1981): "Computed transmission through rain in the 1-400 GHz frequency range for spherical and elliptical drops and any polarization. Alta Freq (Italy), 50, 262.
[11] ITU-R P.838-3, "Specific attenuation for rain for use in prediction methods, ITU-R, Geneva, Switzerland, pp 6, 2005
[12] ITU-R P.839-4, "Characterizations of precipitation for propagation modeling height Model for prediction, ITU-R, Geneva, Switzerland, pp 4, 2013
[13] Fashuyi M.O and T.J Afullo (2007) "Rain Attenuation Prediction and Modeling for Line-of- Sight Links on Terrestrial Paths in south Africa," *American Geophysical Union: Radio Science*, Vol. 42, RS5006, doi: 1029/2007RS003618.
[14] Alonge A. A. and T. J. Afullo (2012): seasonal analysis and prediction of rainfall effects in eastern South Africa at microwave frequencies *Progress In Electromagnetics Research B,* Vol. 40, pp. 279-303.
[15] ITU-R P.618-10, "Propagation data and prediction methods required for the design of Earth- space Telecommunication systems, ITU-R Ser., Geneva, Switzerland, pp. 4, 2012.

**Joseph Sunday Ojo (*PhD*)** received his undergraduate degree in 1998 from Federal University of Technology, Akure, Nigeria and also bagged his Master's and *PhD* in Communication Physics from the same institution in 2004 and 2009 respectively. His research interests include radiowave propagation, wireless mobile and satellite communication networks among others.

**Pius Adewale Owolawi (*PhD*)** received his undergraduate degree in 2001 from the Federal University of Technology, Akure, Nigeria and also bagged his Master's and *PhD* in Electrical Engineering from University of Kwazulu Natal, South Africa in 2006 and 2010 respectively. His research interests include radiowave propagation, RF and microwave Systems.

# Multicast Group Flow Rate Scaling in WiMAX Networks

Didacienne Mukanyiligira, Alexandru Murgu
Department of Electrical Engineering
University of Cape Town, P. O. Box 7701, Rondebosch, Cape Town
Tel: +27 21 650 4801, Fax: +27 21 650 2465
email: mkndid001@myuct.ac.za, Alexandru.Murgu@uct.ac.za

**Abstract-Multicast broadcast services in wireless network, such as Worldwide Interoperability for Microwave Access (WiMAX), is an effective technique used for the transmission of data at high speed rate to high speed mobile users. WiMAX technology provides internet access to nomadic users. Advances in WiMAX technology allow scalability of quality of service classes differentiation. Multicast flow transmission however suffers from quality of services differentiation. Therefore multicast users are not able to receive multicast contents with their desired quality of service. To address this problem, we designed a multicast flow rate scaling algorithm. The algorithm allocates bandwidth to multicast flow by considering QoS and feedback channel state information constraints. The algorithm will be evaluated in ns3 simulator, while throughput and delay will be used to evaluate the performance. This research contribution considers the multicasting group rate scaling problem to be solved by QoS differentiation of multimedia multicast flows in order to optimize the multicast broadcast services (MCBCS)/WiMAX bandwidth shareability and throughput improvement.**

**Index Terms—WiMAX, MCBCS, QoS Scalability.**

## I. INTRODUCTION

WiMAX is a high speed wireless technology which supports different types of services such as data, voice telephony, video, Internet protocol television streams, and Internet access for mobile and fixed users. WiMAX provides mobility up to the speed of 120 kilometer per hour, and high speed data rate to nomadic and mobile users [1]. WiMAX offers multimedia services including voice which leads to specific demands on Quality of Service (QoS) differentiation, and thereby imposing a different QoS management of such services. WiMAX supports up to five QoS classes. These include Unsolicited Grant Service (UGS) for voice application, real time Polling Service (rtPS) for video application, extended rtPS (ertPS) for voice with activity detection, non-real time Polling Service (nrtPS) web transactions, and Best Effort (BE) service for file transfer [1], [2].

Additionally, WiMAX supports different networking paradigms including multicast services. Multicasting is a technique used to deliver the same content to subscriber stations simultaneously. Multicast groups must be formed and the contents are delivered at a single rate per multicast group. Different from unicast, in multicasting users in bad channel conditions will receive reduced QoS in multicast system. This will negatively affect network throughput and efficiency [3].

Although multicasting is an effective technique it suffers from bandwidth management problems in the mobile networks due to user mobility and handover. (MCBCS)/WiMAX framework is used for the simultaneous delivery of the same content to large multicast groups of users [4]. This means that the multicasting groups share the same channel and contents at the regional, national, and geographical level. Hence it increases the scalability of QoS classes, bandwidth utilization efficiency, and flexibility [4]. Multicast broadcast services (MCBCS)/WiMAX framework has been developed to solve bandwidth problems.

Although MCBCS addressed the bandwidth problems, however multicast group flow rate was not considered. Researchers in [5] and [11] showed that the minimum multicast transmission rate provides data rate to the users located at the edge of a base station with limited throughput to users in good channel condition. However, if the transmission rate is maximum, this has also a detrimental effect on users in bad channel condition. In [6] it is proposed an opportunistic multicast scheduling with resource fairness constraints which considers channel state condition, but this algorithm does not QoS constraint for the multicast members.

Song et al. [7] Investigated the traffic scheduling for multiple users. It considered QoS demands when multiple fading channels are shared among the users subject to efficiency and flexible resource allocation. Also, the utility function approach for waiting time management of queues was considered. This increases the efficiency of bandwidth use, and ensures reasonable delay of traffic delivery to individual users. The algorithm proposed in [7] considers both channel and queue state information for achieving the maximum aggregate utility. It also considers the optimization of bandwidth allocation for multicast traffic in order to enhance the QoS support of multicast services in WiMAX network by scheduling the downlink multicast flows. However they did not consider multicast group rate selection. Researchers in [8] considered an extension to cooperative multicast scheduling in order to improve the network throughput with no QoS constraints. In [9] and [10] the delay, throughput, and fairness for multi-state adaptation to WiMAX channel condition has been improved. In [11] a generalized multicasting scheduling technique was considered in order to determine the optimum transmission rate of a multicast group to optimize the overall throughput of the network.

The contributions of this paper are: 1) designing a new scheduling algorithm for the multicast flows which considers rate selection and QoS constraints; 2) designing a scheduling hierarchy of QoS classes (UGS, rtPS, nrtPS, BE).

The rest of this paper is organized as follows. Section II introduces the multicasting services in WiMAX networks and discusses the MCBCS/WiMAX framework and WiMAX Service Classes Scheduling. Section III presents the derivation of the multicast rate flow scaling model developed. Section IV discusses the software representation of the MCBCS/WIMAX framework by describing algorithm and pseudo code of the new designed multicast flow scheduler. And finally section V gives the future work and concludes the paper.

## II. Multicasting Services in WiMAX Networks

Multicasting is a technique used to transmit popular contents to a group of users. Users requesting the same content have to form a multicast group before receiving such contents. Group formation is done at the base station after receiving a request from a subscriber station. Then the base station transmits the differentiated multicast contents to groups, so that such groups share bandwidth.

### A. MCBCS/WiMAX Framework

The three major aspects of MCBCS/WiMAX framework are the Connectivity Service Network (CSN), Access Service Network (ASN), and the Subscriber Stations (SSs). SSs are grouped into three multicast groups such as Multicast Group 1 (MC G1), Multicast Group 2 (MC G2), and Multicast Group 3 (MC G3) as shown in figure 1. The differences between the WiMAX network architecture and MCBCS/WiMAX framework is that, the later has an MCBCS controller located in the CSN. This is responsible to manage the MCBCS services in the CSN. Base stations and subscriber stations have additional Multicast Broadcast Service (MBS) [12] functionalities. MCBCS/WiMAX ensures that the multicast contents can only be accessed by the authorized users of a multicast group [12].

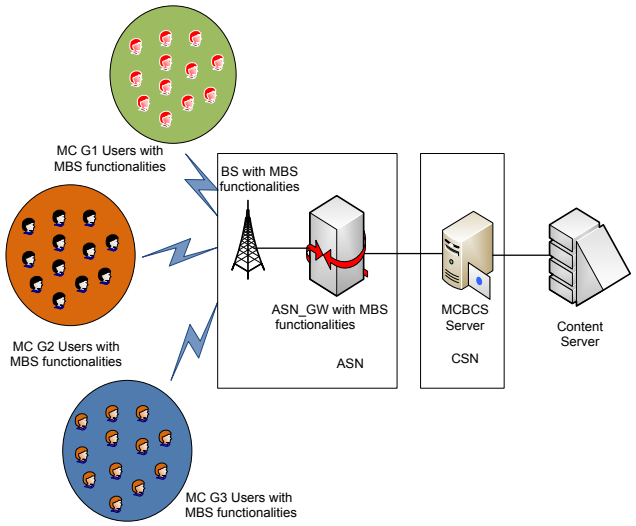



Figure 1. MCBCS/WiMAX framework

ASN is composed of base station and Access Service Network Gateway (ASN-GW). For MCBCS/WiMAX framework, the ASN-GW has MBS functionalities which control MBS proxy, Distribution Data path Flow (DPF), synchronous service Flow Authorization (SFA), and upper synchronization executer. The base station also has MBS functionalities which are responsible of MBS Data Path Flow (DPF), Service Flow Management (SFM), lower synchronization executer, synchronization controller, and upper synchronization executer [12].

MCBCS controller is owned by the Network Service Providers (NSP) whereas the Network Application Provider (NAP) owns the ASN. NSP and NAP are different business entities, but they may coexist in the same environment [12].

### B. Service Classes Scheduling in WiMAX

MCBCS/WiMAX has five different service classes, namely; UGS, ertPS, rtPS, nrtPS, and BE listed in decreasing order of their priority as it is shown in table 1 [13]. The first three service classes were designed for real time applications and the last two are for non-real time applications. UGS supports real-time service flows which are fixed-size data packets delivered at periodic intervals. ertPS is an addition by the 802.16e standard enforcing the efficiency of UGS and rtPS for real time applications based on variable bit rate. rtPS supports real-time service flows which are variable size data packets and delivered at periodic intervals. nrtPS has been created for supporting delay-tolerant service flows and it requires minimum data rate. BE supports service flows which demand minimum service guarantees, and thus can be delivered on the network bandwidth availability [13]. Table 1 shows the QoS attributes such as Minimum Reserved Traffic Rate (MRTR), Maximum Sustained Traffic Rate (MSTR), traffic priority, maximum latency, jitter tolerance, and their specific applications.

Table 1. Service classes characterization [13]

| QoS classes | MSTR | MRTR | Traffic Priority | Max Latency | Jitter tolerance | Applications |
|---|---|---|---|---|---|---|
| UGS | √ | | | √ | √ | VoIP |
| ertPS | √ | √ | √ | √ | √ | Voice with activity detection (VoIP) |
| rtPS | √ | √ | √ | √ | | Video or audio streaming |
| nrtPS | √ | √ | √ | | | FTP |
| BE | √ | | √ | | | Data transfer, Web browsing etc. |

### C. WiMAX MAC and Physical layer

In MCBCS/WiMAX, the media Access Control (MAC) layer works on connection-oriented principle and it is composed of three substrates: the MAC Convergence Sublayer (CS), MAC Common Part Sublayer (CPS) and MAC security sublayer. Convergence sublayer is an interface between the MAC layer and the upper layers. The Institute of Electrical and Electronic Engineers (IEEE) standard defined two CS such as, the Asynchronous Transfer Mode (ATM) and packet switching service. CS receives Service Data Units (SDUs) from the upper layers, suppress header of these packets and then it does addressing and mapping of SDUs. The CPS has functionalities such as packet fragmentation and concatenation of SDU into MAC Packet Data Units (PDUs), transmission of MAC PDUs, and assigns the connection unique Connection ID (CID) [14].

The connection queues are prioritized based on real time and non-real time demands. Then the scheduler allocates sub-channels and time slots to particular queues. Convergence Sublayer and Security Sublayer are not the concern of this paper, the work is only related to Common Part Sublayer where the scheduling of multicast flows takes place [14].

Figure 2 describes the MAC layer hierarchy and its functionalities. Communication between upper layers and the CS is through the CS Service Access Point (SAP). The communication from CS sublayer to CPS is through the MAC-SAP. The scheduling, queue prioritization, and bandwidth allocation are done at the CPS sublayer which will allocate bandwidth and time slots to the packets to be transmitted. The physical layer considers the Shannon theorem to predict maximum possible rate the for a channel state.

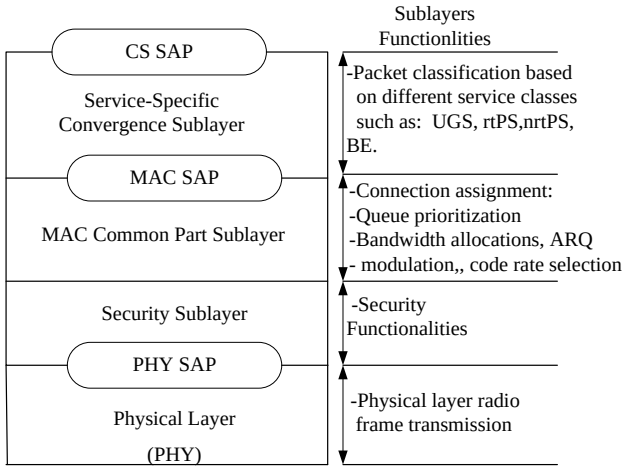



Figure 2. MAC sub-layers hierarchy including PHY layer [14].

WiMAX supports four types of modulations: Binary Phase Shift Keying (BPSK)**,** Quadrature Phase Shift Keying (QPSK), 16 Quadrature Amplitude Modulation (QAM), and 64 Quadrature Amplitude Modulation (QAM) [1]. The benefit of using more than one modulation scheme in WiMAX is that when the channel state condition changes, the modulation will also change adaptively. In efficient multicasting, the transmission rate changes to any of the rate produced from modulation and coding rate to optimize the network bandwidth utilization.

## III. MULTICAST FLOW RATE SCALING MODEL

### *A Problem Formulation*

Let us consider $Th_i$ to be the average throughput of user $i$, $G$, the total number of multicast groups, $N_g$ the total number of subscriber stations in multicast group $g$, $r_{ig}(t)$ channel state information rate for subscriber station $i$ in group $g$ at time $t$. Let $r_{eg}(t)$ be the estimated rate assigned to group $g$ at time $t$, and $\delta_{ig}(t)$ be the QoS rate needed to meet the QoS requirements for user $i$ in the multicast group $g$. $Th_{ig}(t)$ is the throughput calculated via statistical smoothing techniques such as the exponential moving average throughput for user $i$ in group $g$.

The problem is formulated for a single base station, which schedules multicast flows. The scheduler uses the information about channel state from each subscriber station in the multicast group and the QoS requirements from different multicast flows to determine the multicast group rate. The rate $r_{ig}(t)$ is computed as follows [15]

$$r_{ig}(t) = BW log_2(1 + SNR_i) \tag{1}$$

where $BW$ is the bandwidth, and $SNR_i$ is the Signal to Noise Ratio for user $i$ in group $g$.
The rate $\delta_{ig}(t)$ is calculated as [16]

$$\delta_{ig}(t) = min\{max\{\vartheta_i, r_i^{min}(t)\}, r_i^{max}\} \tag{2}$$

where $\vartheta_i$ is the minimum required average traffic rate for the real time users.
$r_i^{min}$ is the required data rate to ensure the QoS guarantees, and $r_i^{max}$ is the maximum possible data rate of the real time users in time slot t, which can send packets in the queue in one time slot.

Let's define the scheduler for allocating resources as follows [10].

$$1_g(t) := \begin{cases} 1, \text{if the scheduler allocate resources to group } g \text{ at time } t \\ 0, \text{otherwise} \end{cases} \tag{3}$$

then the updated (exponential/moving) average throughput will be

$$\begin{aligned} &Th_{ig}(t+1) \\ &= (1-\rho)Th_{ig}(t) \\ &+ \rho r_{eg}(t) 1_g(t) 1_{\{r_{ig}(t) \le r_{eg}(t) \le r_{ig}(t)\}} \end{aligned} \tag{4}$$

where $\rho$ is the latency time scale parameter representing the number of slots the subscriber station $i$ in group $g$. Group $g$ receives rate of $r_{eg}(t)$ only if it is selected and its channel state information rate $r_{ig}(t)$ is greater or equal to $r_{eg}(t)$, and channel state information rate $r_{ig}(t)$ is less or equal to the rate $\delta_{ig}(t)$demanded by QoS constraints.

### *B. Multicast Flow Scheduling*

Let us consider the aggregate throughput of group $g$ at time $t$ denoted by $Th_g(t)$ [10] and the aggregate rate of all subscriber stations in group $g$ at time $t$ by $\varphi_{g,t}(x)$ where the base station transmits at rate $x$.

$$\text{Then} \qquad Th_g(t) := \sum_{i=1}^{N_g} Th_{ig}(t) \tag{5}$$

$$\begin{aligned} \varphi_{gt}(x) &= \sum_{i=1}^{N_g} x \cdot 1_{\{x \le r_{ig}(t)\}} 1_{\{x \ge \delta_{ig}(t)\}} \\ &= \sum_{i=1}^{N_g} x \cdot 1_{\{\delta_{ig} \le x \le r_{ig}\}} \end{aligned} \tag{6}$$

this leads to

$$\varphi_{gt}(x) = \sum_{i=1}^{N_g} x \cdot \min\left\{1_{\{x \le r_{ig}(t)\}}, 1_{\{x \ge \delta_{ig}(t)\}}\right\} \quad (7)$$

where

$$1_{\{x \le r_{ig}(t)\}} = \begin{cases} 1, \text{if } x \le r_{ig}(t) \\ 0, \quad \text{otherwise} \end{cases} \quad (8)$$

and

$$1_{\{x \ge \delta_{ig}(t)\}} = \begin{cases} 1, \text{if } x \ge \delta_{ig}(t) \\ 0, \quad \text{otherwise} \end{cases} \quad (9)$$

Equations (4), (6), and (7) are the newly defined equations for multicast flow scheduler. This is an intra-class scheduler of traffic flows. In [10], the authors considered multicast group rate selection where the aggregate rate depends only on the maximum information rate of each subscriber station but no consideration was given to QoS demands in selecting the multicast flow rate.

The novelty claim of this paper is that we consider the multicast flow rate scaling taking into consideration both the channel state rate for each subscriber station and the QoS demands for multicast flows.

### *C. Group Rate Selection*

The estimated rate assigned to group $g$ at time $t$ to satisfy QoS and channel state constraints is given as;

$$r_{eg}(t) = arg \max_{x} \sum_{i=1}^{N_g} x \cdot \min\left\{1_{\{x \le r_{ig}(t)\}}, 1_{\{x \ge \delta_{ig}(t)\}}\right\} \quad (10)$$

Let us denote the value of $x$ which maximizes $\varphi_{gt}(x)$ as $x^*$. The base station chooses to serve the group $g(t)$ at rate $r_{eg}(t)$, given by

$$g(t) = arg \max_{1 \le g \le G} \frac{\varphi_g\left(r_{eg}(t)\right)}{Th_g(t)} \quad (11)$$

and we denote it $g^*$

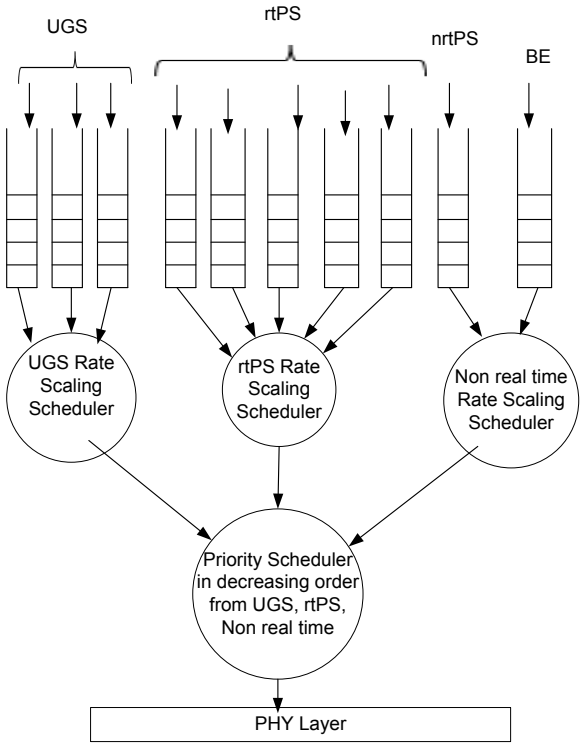



Figure 3. Scheduling hierarchy for flow rate scaling.

## IV. SOFTWARE REPRESENTATION OF MCBCS SCALING

In software representation of the multicast flow scalability we consider three aspects, namely, intra-class and interclass scheduler. We design the scheduling algorithm and its hierarchical architecture. The flow chart in figure 4 shows the logics of how this new flow scheduler will be implemented. The algorithm is being implemented in ns3 simulation software. The pseudo code of flow rate scaling, rate selection and how it will be used in scheduling multicast traffic is shown in figure 5.

In the proposed scheduler shown in figure 3, the inter class Priority-Based Scheduler is used to schedule multicast traffic. Priority based scheduler from UGS, rtPS and non-real time multicast connections which are nrtPS and BE are considered. The intra-class Proportional fair scheduler (PFS) is used in UGS multicast connections and also in rtPS multicast connections. The inter class Proportional Fair scheduler is used between nrtPS and BE effort multicast connections.

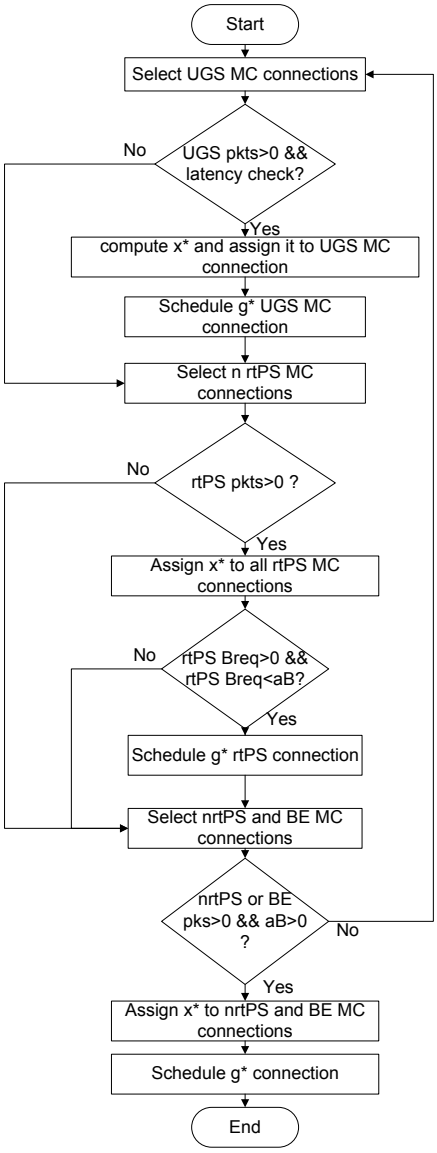



Figure 4. MCBCS bandwidth sharing flow chart.

The scheduler shown in figure 4 is being implemented in ns3. The scheduler starts by selecting UGS multicast connections. The scheduler then checks for connections which have packets and delay constraints as this type of class is delay sensitive. Then sets $x^*$ transmission rate to each multicast connection and choose the group to assign connection according to (11).

After scheduling all UGS connections, the scheduler will continue to rtPS flows. The scheduler will first check if the connections have packets before assigning the transmission rate $x^*$ to those connections. After this it checks for the bandwidth availability constrains and schedules all rtPS connections using (11).

Different from the previous steps, the scheduler selects two types of flows at the same time such as nrtpPS and BE connections. The scheduler starts by checking if the there is any packet in their connections and then maps the transmission rate $x^*$to the connections. After this it assigns $x^*$rate to all rtPS and BE effort connection and then checks if there is available bandwidth. Lastly it schedules $g^*$ connection according to (11).

1. Select UGS multicast connections
2. **while** UGS connections **Do**
3. **if** UGS connections has packets to transmit and latency check are satisfied
4. **for** modulation 1 to 7
5. **for** multicast group 1 to G
6. Compute φ according to (7);
7. Compute $r_{eg}$according to (10);
8. Assign the $x^*$ to each multicast group according to (10);
9. Compute g according to (11);
10. Schedule $g^*$
11. **end** For
12. **end** For
13. **else**
14. Select rtPS multicast connections
15. **while** rtPS connections **Do**
16. **if** rtPS connections has packets and bandwidth needs are satisfied
17. **for** modulation 1 to 7
18. **for** multicast group 1 to G
19. Compute φ according to (7); Compute $r_{eg}$according to (10);
20. Assign the $x^*$ value of each multicast group according to (10);
21. Compute g according to (11);
22. Schedule $g^*$
23. **end** For
24. **end** For
25. **else**
26. **while** nrtPS and BE connections **Do**
27. **if** nrtPS and BE connections has packets and bandwidth is available
28. **for** modulation 1 to 7
29. **for** multicast group 1 to G
30. Compute φ according to (7);
31. Compute $r_{eg}$according to (10);
32. Assign the $x^*$ value of each multicast group according to (10);
33. Compute g from Equation (11) ;
34. Schedule $g^*$
35. **end** For
36. **end** For

Figure 5. Multicast Scheduling Pseudo code

## V. CONCLUSION AND FUTURE WORK

In this paper we have proposed a new multicast flow scheduling algorithm which considers real time and non-real time traffic classes. This algorithm considers the channel state rate information of each subscriber station and the QoS requirement for each multicast flow. It computes the multicast transmission rate which improves the throughput of each subscriber station. The future work will consider the proposed algorithm for mapping the multicasting flows virtualization control environment in order to move MCBCS/WIMAX into highly scalable management framework for multicasting.

ACKNOLEDGMENTS

This work is supported in part by University of Cape Town, South Africa.

REFERENCES

[1] L. Nuaymi, *WiMAX: technology for broadband wireless access*. John Wiley & Sons Ltd, 2007.

[2] R. Prasad, F.J. Velez *WiMAX Networks: Techno-Economic Vision and Challenges*. London: Springer, p. 488, 2010.

[3] Afolabi, R.O., Dadlani A., Kiseon Kim "Multicast Scheduling and Resource Allocation Algorithms for OFDMA-Based Systems: A Survey," *IEEE Commun. Surv. Tutorials*, vol. 14, no. 1, pp. 240 – 254, 2013.

[4] Alvarion Ltd., Multicast and Broadcast Services (MCBCS), 2010.

[5] Wen-jun XU, Zhi-qiang HE, Kai NIU, Jia-ru LIN, Wei-ling WU "Multicast resource allocation with min-rate requirements in OFDM systems," *Journal of China Uniersities Posts and Telecommunications*, vol. 17, no. 3, 2010.

[6] W. Liu, L. Zhu, X. Wang, H. Yu, and M. Guizani, "Opportunistic Multicast Scheduling with Resource Fairness Constraints in Cellular Networks," *IEEE Int. Conf. Commun.*, pp. 1–5, Jun. 2011.

[7] G. Song , Ye (Geoffrey) Li , Leonard J. Cimini, Jr, and H. Zheng "Joint Channel aware and queue aware data scheduling in multiple shared wireless channels," *IEEE Wireless Communications and Networking Conference*, pp. 1939 – 1944, 2004.

[8] F. Hou, L. X. Cai, Pin-Han Ho, X. Shen, and J. Zhang "A Cooperative Multicast Scheduling Scheme for Multimedia Services in IEEE 802.16 Networks," *IEEE Trans on Wirel. Commun..*, vol. 8, no. 3, pp. 1508 – 1519, 2009.

[9] H. Du , J. Liu, and J. Liang, "Downlink Scheduling for multimedia Multicast/Broadcast over Mobile WiMAX: Connection-oriented Multi-state Adaptation," *IEEE Wirel. Commun. Mag.*, vol. 16, no. 4, pp. 72–79, 2009.

[10] H. Won, H. Cai, D. Y. Eun, K.. Guo, A. Netravali, I. Rhee, and K. Sabnani, "Multicast scheduling in cellular data network," *IEEE Trans. Wirel. Commun.*, vol. 8, no. 9, pp. 4540–4549, 2009.

[11] V. Vukadinovic and G. Karlsson, "Multicast scheduling with resource fairness constraints in cellular networks," *Springer wireless networks*, vol. 15, no. 5, 2009.

[12] WiMAX Forum, "Network Architecture System Requirements, Network Protocols and Architecture for Multi-cast Broad-cast Services," Tech. Rep. WMF-T33-112-R015v02, 2011.

[13] R. Muhaned, J. G. Andrews, A. Ghosh, *Fundamentals of WiMAX and understanding*. Pearson Education, Inc, p. 496,2007.

[14] J. Roberto B., de Marca. John, K.C. Chen, *Mobile WiMAX*. England: Wiley & Sons Ltd, 2008.

[15] L. W. Couch, *Digital and Analog Communication Systems*. Pearson Education, 2013.

[16] Y. Kim, K. Son, and S. Chong, "QoS Scheduling for Heterogeneous Traffic in OFDMA-Based Wireless Systems," *IEEE Global Communication Conf.*, pp. 1–6, Nov. 2009.

**Mukanyiligira Didacienne** received her undergraduate degree in 2005 from Kigali Institute of Science and Technology in Electromechanical Eng. and her Master's Degree from Huazhong University of Science and Technology, China in communication and Information systems. She is presently studying towards her PhD degree at the University of Cape Town. Her research interests include multicasting, resource sharing, WiMAX, and network virtualization.

# CONVERGED SERVICES

# Integration of Phonotactic Features for Language Identification on Code-Switched Speech

Koena Ronny Mabokela and Madimetja Jonas Manamela
Telkom Centre of Excellence for Speech Technology, Department of Computer Science
University of Limpopo (Turfloop Campus), Private Bag, X1106, Sovenga, 0727
Tel: +27 79 711 3298, Fax: +27 15 268 3487
email:krmabokela@gmail.com; jonas.manamela@ul.ac.za

**Abstract-This paper presents an incorporation of phoneme sequences as language information to perform language identification (LID) in code-switched speech. The one-pass recognition system converts the spoken utterances into an occurrence of phone sequences. We employed hidden Markov model (HMM) to build robust content-dependent acoustic models that can handle multiple languages within an utterance. We reported two phoneme mapping methods to determine the phoneme similarities among our target languages. A statistical phoneme-based bigram language model is incorporated for speech decoding to obviate possible phone mismatches. We supervised support vector machine (SVM) which learned the language transition of the phonotactic information in the mixed-language speech given the recognized phone sequences. The back-end decision is taken by an SVM which classifies language identity given the likelihood scores based on the monolingual phone occurrence segments. The experiments were performed with commonly mixed Northern Sotho and English speech corpora. We evaluate the system measuring the performance of the phone recognition and LID portions separately. We were able to obtain a phone recognition accuracy of 84.4% when using data-driven phoneme mapping approach modeled with 16 Gaussian mixtures per state. The proposed system achieved an acceptable LID accuracy of 89.6% and average of 81.4% on code-switched speech and monolingual speech segments respectively.**

**Index Terms— language identification; phonotactics; acoustics; language model; code-switch speech**

## I. INTRODUCTION

Most multilingual speakers have the ability and tendency for engaging in code-switching – a mixed-language phenomenon that is referred to as the usage of more than one language within an utterance. It appears to be preferred commonly in multilingual societies [1]. According to the constitution established by the national legislation of Republic South Africa, chapter 1 - of the Bill of Rights (1996) - section 6, it states that: *Pan South African Language Board (PanSALB) has the right to promote the use all eleven official languages*. This has resulted in South Africa being a multilingual society with eleven official languages.

Most native speakers have the tendency to use more than one official language in their daily spoken conversations. The general use of South African languages in daily conversations portrays a mixed language mode (e.g., in Radio and Television shows, News Broadcasting, religious worship services, interviews and presentations). Amongst 6909 natural languages spoken worldwide as revealed by Ethnologue database, English is generally used as a global language for communication [2]. However, the African reality in many communication episodes is that English is frequently mixed with indigenous under-resourced official languages. For most African indigenous speakers prefer to utter numerical digits, codes, and times in English. The relevance of the research stems from the fact that it is common for more than one language to be spoken in the same region in South Africa.

Usually, code-switched speech contains the phrases or words from the secondary language. It is known as the language embedded in the primary language. Code-switched speech is commonly just spoken but not formally written. For this reason, code-switched speech can be classified in the same category as under-resourced languages [1]. The tendency of mixing more than one official languages as an emerging modern-day style of communication pose a greater challenge to speech-enabled technologies. The state-of-the-art human language technology (HLT) has thus far reached an era where the main focus of its current research activities includes, the development of automated systems that enable multilingual individuals to interact easily with smart computing devices. This serves as a major goal of the HLT sector towards benefitting a wide range of multicultural societies - from the not-so-literate ruralites in the remote rural communities who want to obtain relevant lifesaving medical information over a cellphone line, to professional scientists in state-of-the-art industry laboratories who need to focus on commercial problem-solving with computer devices [3].

To date much research has been done on the state-of-the-art spoken language processing systems which rely profoundly on pattern recognition, one of the most challenging problems for computational systems. In a broader sense, the capability to recognize, identify or classify speech pattern forms the core of artificial intelligence [4, 6]. Spoken language identification (LID) refers to an automatic agent that can accurately determine which language is spoken in a given sample of speech utterance. An LID system is an enabling technology for a wide range of multilingual speech processing applications, such as multilingual information retrieval [5], spoken language translation [4], and telephone call routing system [6]. The most significant trends in the state-of-the-art speech

technologies is the ability to support multiple input and output languages, more especially if the applications of interest are intended for global markets and linguistically diverse end-user communities [4].

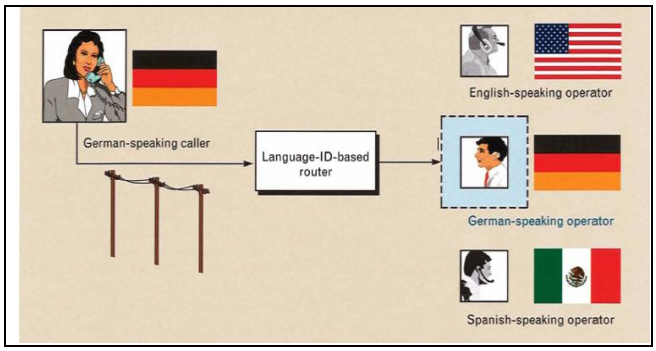



**Figure 1**: Example of an LID based call routing system [22]

In this paper, we propose a language identification system integrated with a single-pass phone recognition system to identify languages in a code-switched utterance. We proposed to build a single-pass phone recognition system to decode multiple languages within an utterance [1, 7]. Our experimental approach used context-dependent HMM-based acoustic models and phone-based language model. We adopted nearly the same technique as parallel phoneme recognition followed by language modeling (P-PRLM) to do language identification on code-switched speech. Although we report only on experiments conducted using two official languages of South Africa, the same procedure can be followed and extended to other under-resourced languages.

This paper is organized as follows: Section II discusses the background of language identification on code-switched speech and related work. Section III describes the experimental data used for the system. In section IV, the proposed integrated LID system explained in details. Section V presents the experimental setup and results. Finally, the conclusion and future directions are offered in Section VI.

## II. Background and Related Work

In Singapore, Mandarin and English are often mixed in spoken conversations [1], in Hong Kong a code-switching between Cantonese and English is used in many occasions [8] and in Taiwan, a Mandarin-Taiwanese code-switching speech was reported [9]. Others reported a mixed-language speech found in India between Hindi and English [10]. It is also observed in South Africa, where a code-switching between two South African indigenous languages such as Xhosa and Zulu was studied for LID and multilingual speech recognition [11]. Recently, Modipa *et al.* [12] reported a context-dependent modeling technique of English vowels in Sepedi code-switched speech where the process of obtaining a phone mapping from embedded language to the matrix language was investigated.

There are few reported approaches in the repertoire of code-switched speech. One approach used is to integrate multiples cues such as acoustics, prosodic, and phonetics to distinguish languages in a code-switch speech utterance [8]. A language boundary detection (LBD) method is applied to detect multiple languages within an utterance [9]. The second approach, such as the delta-Bayesian information criterion (Delta-BIC) and Latent semantics analysis (LSA) is used to separate English, Mandarin and Taiwanese in code-switched utterances [9]. Lastly, an approach that uses maximum a posteriori based estimation was used to jointly segment and identify utterances of a mixed language [13]. The above mentioned approaches use an LID module that incorporate LBD module. The LID systems that incorporate LBD module are usually not preferred due to incorrect assumptions that code-switched speech segments are independent of each other and as a result, errors in the LID module cannot be recovered [1]. Therefore, in this case, if the LBD module cannot achieve 100% then it will directly influence the performance of LID module, thereby limiting the performance of the speech recognition module [1, 10].

On the other hand, a multilingual speech recognition approach can be able to handle code-switched speech that comprises of multilingual acoustic model, multilingual pronunciation dictionary, and multilingual language model that allows the mixing or sharing of models across different language units [1, 10]. Nevertheless, a multilingual ASR approach does not need an additional LID module to identify speech segments since language information is incorporated directly into the system [1]. One technique is to use a linguistic knowledge-based method to establish a multilingual phone set mapping or clustering of similar phonetic features that share the training data [7].

The common examples are International Phonetic Alphabet (IPA), Assessment Methods Phonetic Alphabet (SAMPA) and Wordbet [15]. Another technique is to map language-dependent phones using a data-driven approach such as clustering of specific phones according to distance measured between similar acoustic models. Examples of data-driven based method are Confusion Matrix, Bhattacharyya Distances and Kullback-Leibler Divergent which takes spectral characteristics into consideration [15].

Lyu *et al.* [16] proposed the use of a word-based lexical model LID system which uses the lexicon information to distinguish code-switching speech within an utterance. A two-stage scheme system is used where a large vocabulary continuous speech recognition (LVCSR) system is employed. Then a trained word-based lexical model is applied to identify languages via recognized words sequences. The approaches such as P-PRLM, phoneme recognition followed by language modelling (PRLM) [4] and parallel phoneme recognition vector space modelling (PPR-VSM) [17] are some of the most popular approaches to LID system. The P-PRLM approach employs multiple phoneme recognizers that tokenize the speech waveform into sequences of phonemes. Then the resulting sequence of phonemes is passed to the n-gram language model which determines the most probable language from the target languages [6, 18]. The supervised SVM has proven to be the best classifier [18].

A similar approach is used to distinguish among eleven officially spoken languages of South Africa [18]. It was implemented using P-PRLM architecture and techniques such as phoneme frequency filtering - where SVM-based classifier is used to classify languages at the back-end. The

SVM classifier which was able to achieve the average LID rate of 71.78% on test samples of length 3-10 seconds and LID rate of 82.39% when clustering of similar language families was affected.

## III. Proposed Approach

Our integrated system is targeted to identify only two languages, namely, Northern Sotho and English, on code-switched speech utterances. Figure 2 shows the front-end of the phone recognition system designed to decode mixed-language speech utterances. A phone recognition system takes speech waveform and output the corresponding phone sequences. This is done when a phone recognition system estimates the likelihood score of the optimal phone sequences given the acoustic features extracted from the speech utterance waveform. We assume that the speech waveform can be segmented into a sequence of phones. To achieve this, a phone n-gram language model is employed to estimate the likelihood score of the $n^{th}$ phone given the (n-1) $^{th}$ of the preceding phones.

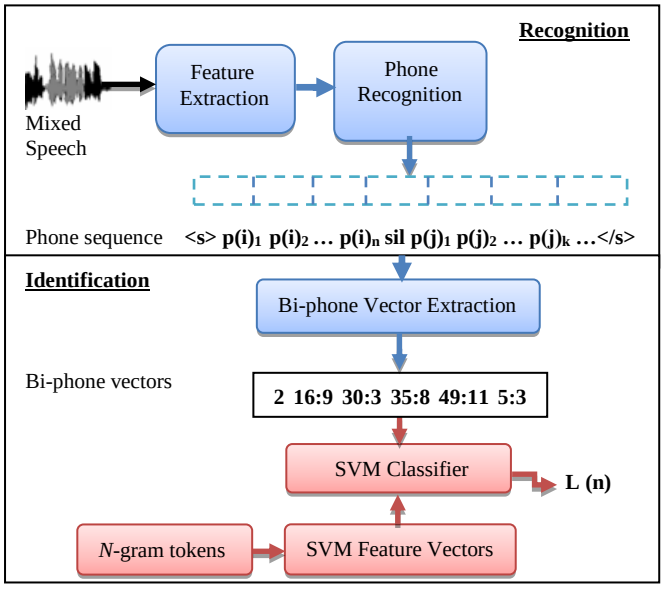



**Figure 2**: The scheme designed for language identification on code-switched speech.

A Baum-Welch iteration algorithm is used during training of acoustic models to perform HMM-based parameter re-estimation. For the recognition purpose, the acoustic features are compared with the HMM-based acoustic models as well as the phone language model. The sequences of phone strings are decoded by the Viterbi decoding algorithm which searches the optimal sequence of the phones using the combined likelihood scores from the acoustic model and phone language model.

The SVM-based classifier is used to identify only two class feature samples; languages outside the targeted range will not be classified. For each phone sequence generated from the phone recognition, the bi-phone occurrences are extracted from the phone sequences and converted into a suitable SVM format with a unique numerical representation. This approach is similar to vector space modeling [4]. Then LID is performed by using SVM-based classifier to score the phoneme sequence of a test utterance. The language model with the highest log likelihood score is chosen to be the most likely sample for classification.

The bi-phone frequencies are then used as an input to the back-end SVM-based classifier. The bi-phone feature vectors have the following numerical attributes, a label is the class label in a numerical representation, a feature index represents ordered feature indexes - that is, the location of that particular bi-phone feature, usually, integer representation, and in our case, a feature value represents the frequency count or occurrences of each bi-phone feature attribute. The SVM classification model is used to separate vectors in a binary classification and hypothesize the maximum likelihood score of the bi-phone frequencies of each language [6].

## IV. Experimental Data

This section describes the speech corpus used for training of robust acoustic model development. Furthermore, we discuss the selection of phone set and the creation of bilingual pronunciation dictionary.

### A. *Speech Corpus Development*

The state-of-the-art LID system requires a large amount of training speech data. Under this condition, a large portion of the mixed speech corpus was attained by combining two monolingual speech data corpus [8]. The corpus used for training of acoustic model included recorded speech data and their respective transcriptions of locally-produced Northern Sotho developed within Telkom Centre of Excellence for Speech Technology (TCoE4ST) and freely available LWAZI South African English speech data often used for speech technology experiments [18]. The speech corpus was divided into two components; training and testing data sets.

The TCoE4ST locally-produced Northern Sotho speech corpus had an amount of 3465 utterances. From the LWAZI English speech corpus, we selected about 1840 recorded speech data and their respective sentential form utterances that were used as training speech data set of the system. Each speaker produced approximately 30 utterances that were phonetically balanced. The speech data were recorded over a telephone channel at 8 kHz sampling rate. The two speech corpora were combined together to form a large vocabulary of sizable mixed-language speech corpus for training the overall integrated system. Table 1 shows the summarized amount of speech data of the mixed speech corpus with two sub-sets.

**Table 1**: *The overall statistics of the mixed speech corpus*

| | **Train set** | **Test set** | **Total** |
|---|---|---|---|
| # Speakers | 143 | 5 | 148 |
| Duration (hours) | 5.5 | 1.0 | 6.5 |
| # Utterances | 5305 | 660 | 5965 |

The speech data set used for testing was not part of training data set. As code-switched speech is generally spoken but not formally written, it is not easy to find code-switched speech data [1]. It is for this reason that a simple finite loop grammar was used to generate about 300 artificially code-switched texts that are semantically and syntactically correct. The generated texts were recorded and included in a test set. We manually reassured the quality of the utterances by removing disfluencies such as long pauses,

laughs and hiccups. Within the code-switched speech, the calculated percentage of Northern Sotho words is 74.2% and English words are 25.8% excluding silences. The average ratio of code-switches within each utterance was not more than 0.5 when counting only switches between Northern Sotho and English. We have extended the test data set by adding 360 quality monolingual utterances.

### B. Bilingual Dictionary and Phone Set

The experimental bilingual pronunciation dictionary used was achieved with combining several monolingual pronunciation word lexicons without retaining duplicate words. For the primary Northern Sotho language we used a limited vocabulary of Northern Sotho pronunciation dictionary that was locally-produced within the TCoE4ST and LWAZI, a freely available Northern Sotho pronunciation dictionary. For the secondary English language, we used freely available LWAZI English pronunciation dictionary often used for speech technology research tasks [18]. All the words in the pronunciation dictionary were manually verified and correctly checked for redundant phone representation.

The compiled bilingual dictionary contained 85891 unique word tokens. The representation used in the bilingual pronunciation dictionary followed the Speech Assessment Method Phonetic Alphabet (SAMPA) notations based on International Phonetic Association (IPA) rules and also taking into consideration of the pronunciation rules [7, 10]. In this case, phones with similar phonetic features were mapped into one best phone candidate representation to lower confusion within the combined phone set. Some English vowel phones were left unmapped since they did not match any of Northern Sotho vowel phones.

## V. Phoneme Mappings Techniques

In this paper, we adopted two different phoneme mapping strategies to determine the phone similarities among the target languages. The first mapping technique is linguistically motivated phoneme mapping which requires a linguistic expert while the other technique is a data-driven phoneme mapping.

### A. Linguistically-Motivated mappings

Our target languages in this experiment have similar phonemes. For this reason, we adopted the same procedure of using linguistic knowledge-based method to construct a combined phone set for the matrix and embedded languages [7, 9]. To achieve this, the language-dependent speech units are defined based on the characteristics of their phonemic properties as represented on the IPA-based scheme [7]. We used knowledge-based IPA method to create linguistically motivated phonetic pairwise mappings. We relied on extensive language documentations where necessary.

The multilingual acoustic model is built by mapping the English phonemes to the Northern Sotho phonemes. This approach is motivated by the occurrence of similar phonemes in our target languages and we also aim to reduce a larger number of phonemes. The criterion to construct linguistically-motivated mappings is obtained as follows [7]: If the IPA classification is similar to the one of the Northern Sotho phoneme then the English phonemes are mapped directly. We define a mapping that maps each English phoneme to its closest matching Northern Sotho phoneme based on the IPA. If there is no close match to be found, then an English phoneme that occurs most frequently, the phoneme inventory is extended with that English phoneme. If none from the above criterion is applicable, then each phoneme is mapped to Northern Sotho phoneme that is mostly confused with as according to a confusion matrix. The diphthongs of English were separated into vowels when applying an IPA-based method. Each phonemic vowel was then mapped to its equivalent phoneme of the target language.

### B. Data-driven mappings

We performed the same procedure which is followed in section A but this time; we defined the phoneme mapping of English to Northern Sotho using the confusion matrix that was attained when the speech recognition system was trained with directly combined phonemes of the target languages. The data-driven mapping which is based on the confusion matrix is built by recognizing the speech utterances of the target language with source language acoustic models [7, 10]. This mapping method consists of the counts of the confusion pairs existing when aligning the speech recognition output and transcriptions of the speech data. The advantage of this approach is that it is fully data-driven and does not require a linguistic expert [14]. For each phoneme of the English language, the most often confusable phoneme to the matrix language is selected for mapping. The phoneme mapping is obtained as follows: For each phoneme $Ø_{L1}$ from the target language, the respective best source candidate phoneme $Ø_{L2}$ is matched. We measure the similarity by selecting the number of phoneme confusions as $\mathbf{c}(Ø_{L1,} Ø_{L2})$. The target phoneme is matched as follows:

$$Ø_{L1} = \boldsymbol{max}\ \mathbf{c}(Ø_{L1,} Ø_{L2}) \quad (1)$$

Thus, for each target phoneme $Ø_{L1}$ source candidate phoneme with the highest number of confusions is determined. If the same number of confusions occurs on two or more source candidate phonemes, the decision on the choice of the target phoneme $Ø_{L2}$ is made by a linguistic expert. The same procedure is employed even when there are no confusions between target and source candidate phonemes.

## VI. Experiments and Results

This section describes the experimental setup, the tools used to develop the system and the speech data used for testing of phone recognition system. Lastly, the experimental results are presented and analyzed.

### A. Experiments Setup

We applied a Hamming window of 25ms length with an overlapping window frame length of 10ms and the pre-emphasis coefficient of 0.97. Acoustic features are obtained using 39-dimensional static Mel-frequency Cepstral Coefficients (MFCCs) with 13 deltas and 13 acceleration coefficients. The Cepstral Mean and Variance Normalization (CMVN) preprocessing and semi-tied transformations are applied to the HMMs. The CMVN is used to overcome the undesired variations across the channels and distortion [9].

The HMM-based phone recognition system was created

with a widely used standard Hidden Markov Model Toolkit (HTK) [19]. The acoustic model uses a three state left-to-right HMM. The HMM-based consist of the tied-state triphones clustered by a decision tree technique. Each HMM state distribution is modeled by 8 Gaussian mixture models (GMM) with a diagonal covariance matrix. Furthermore, the optimal phone insertion penalties and language scaling factors were properly tuned to balance the number of inserted and deleted phone during speech decoding.

A phone language model was incorporated in the phone recognizer for the purpose of speech decoding. The training transcriptions together with the generated code-switched texts were formatted into phone transcriptions and were used to develop the phone language model. A suitable bigram phone language model with discount interpolation was independently trained using a freely available Stanford Research Institute language model (SRILM) toolkit [20]. The resultant best phone language model had a perplexity of 13.8 without reporting out-of-vocabulary (OOV) rate.

The SVM-based classifier was implemented using a freely downloadable library for SVM (LIBSVM) toolkit - an integrated package for training SVM classifier [21]. This SVM program is a suitable package for classifying numerical attributes. The phone sequences used to train the classifier resulted with 3201 support vectors from models. The training process was also aimed at maximizing the margin as well as minimizing the training errors. The bi-phone vector attributes for both testing and training were scaled in the range of [0, 1]. The benefit of scaling data sets is to speed up training and classification process in order to obtain the best model performance and to avoid numerical differences that could lead to over-fitting if the training data attributes are in a large range [16]. A grid search is a simple search technique which was used to estimate the SVM parameters such as C, gamma, margin-error trade-off parameter and kernel width before training the classifier [4, 7]. The Radial Basis Function (RBF) kernel was used for training the classifier. We obtained the optimal parameter for RBF kernel and applied 5-fold cross-validation on the training set as well as estimating each grid point for the accuracy of the classifier.

### *B. Experimental Results and Analysis*

We initiated our experiments with a baseline system. In the following experiments we first evaluated the experimental results of the baseline system and compared them with the results of the integrated LID system applied with two phoneme similarity techniques for mapping the phoneme of the target languages. The baseline LID system was developed using directly combined phoneme set. We did not perform any specific phoneme mapping in the phoneme set. The phoneme set size was large with 67 phonemes. The results obtained from the mixed speech test set are shown in Table 2. We evaluated the experiments on a domain limited test data set.

On the initial phoneme recognition experiment, context-dependent HMM-based acoustic models with 8 Gaussian probability density functions per state were engaged. In an IPA-based phoneme recognition system, the phoneme (such as /**@**/) without specific mapping was approximated to have similar acoustic features as /**a**/. A phoneme such as /**{**/ of English was mapped to /**E**/ using IPA-based scheme and to /**a**/ in data-driven based scheme. Only 45% source candidate phonemes were mapped to the target candidate phonemes in an IPA-based phoneme recognition system and 48% source candidate phonemes were mapped to their highest confusable target phonemes when using a data-driven based phoneme recognition system. These methods allow sharing of the parameters in the HMM-based acoustic models of the target languages. We report the phone error rate (PER) and LID accuracy. Table 2 shows the experimental results of the integrated LID systems and the size of the phoneme set. The results shows that PER and LID accuracy improves when the phoneme mapping is applied.

**Table 2**: *The experimental results of the integrated LID systems and the phoneme set size.*

| | Phone Set size | PER (%) | LID (%) |
|---|---|---|---|
| Direct Mixed | 67 | 33.7 | 85.0 |
| IPA-based | 38 | 28.7 | 85.8 |
| Data-driven | **38** | **19.2** | **87.3** |

The baseline SVM-based classifier was trained using a 5-fold cross-validation which yielded an cross-validation accuracy of 97.5% on trained classification models and has predicted the best parameter value of C =0.5 and γ=0.5. The experimental results of the SVM-based LID classifier were also obtained using RBF kernel. Both phoneme mapping approaches achieve a significant improvement over the baseline results. The data-driven approach was able to outperform the baseline system and the IPA-based approach. The IPA-based approach was able to perform better with the PER of 5% and LID accuracy of 0.8%. The data-driven approach was able to better the performance with the PER of 14.5% as well as the LID accuracy of 2.3%. The SVM-based classifier was trained using a 5-fold cross-validation and RBF kernel which yielded an cross-validation accuracy of 99.75% on trained classification models and has predicted the best parameter value of C =2 and γ=0.5. The evaluated monolingual LID accuracy was as follows: we were able to attain LID accuracy of 81.7% and 81.1% for Northern Sotho and English respectively.

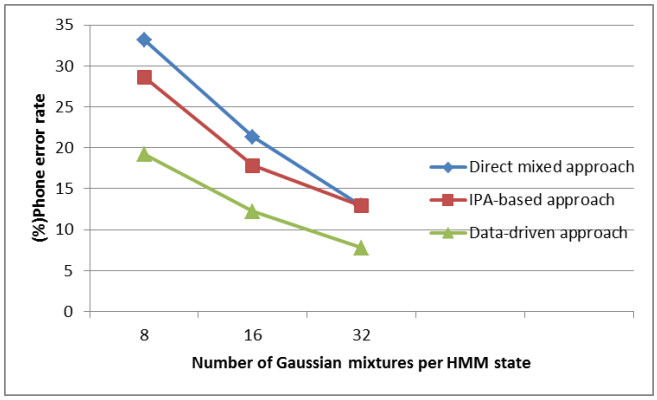



**Figure 3:** The PER of the directly mixed LID, IPA-based and data-driven LID system using 8, 16 and 32 Gaussian mixtures per HMM state.

Figure 3 represents the behavior of the PER with an increasing Gaussian mixtures per HMM state from 8 mixtures up to 32 Gaussian mixtures on each phoneme recognition system. The triphone models were then

improved by gradually increasing the number of Gaussian mixtures, and performing four iterations of embedded re-estimation after each increase. This procedure was continuously until the models had 32 mixtures per state, after which the phoneme recognition results no longer improved significantly on the test data set. We further observed that our trained context-dependent acoustic models with 16 and 32 Gaussian mixtures per state as they tend to better the performance.

The results show that PER improves when context-dependent HMM-based acoustic models with 16 and 32 Gaussian probability density functions per state are engaged. As expected, the data-driven approach performed better even when the mixtures were increased. Both phoneme mapping approaches give better results as compared to the baseline results. The highest performance was observed when our context-dependent acoustic models with 32 Gaussian mixture probability density functions per state are engaged. The data-driven approach was able to achieve the PER of 7.7% outperforming even the IPA-based approach. We also observed that not much significant differences from the back-end LID classifier has been achieved, since the LID accuracies of our three approaches were found to be within the range of 83.7% and 89.6% when the 16 mixtures per HMM state were engaged. The better LID accuracy was achieved by data-driven approach. Our three approaches were able to obtain an improved LID accuracy that ranges from 83.8% to 96.7% when 32 Gaussian mixtures were applied. The data-driven approach was able to achieve the best LID accuracy of 96.7%. The overall LID system was able to perform better due to domain-limited testing speech corpus.

## VII. CONCLUSION

This paper presents an incorporation of phonotactic information to perform language identification (LID) on mixed-language speech. We proposed two phoneme mapping techniques to deal with code-switched utterances at the pronunciation level. The IPA-based approach is derived on linguistic knowledge while the data-driven approach is based on the confusion matrix. Moreover, we also investigated the performance of the PER on increased number of Gaussian mixtures per state. Our proposed IPA-based and data-driven approaches have shown a significant improvement on both PER and LID accuracies. We observed that the data-driven method outperforms the IPA-based approach. We achieved a better PER of 7.7% with a data-driven approach when the context-dependent acoustic models with 32 Gaussian mixtures per state were engaged. In future, we hope to train our system with more robust context-dependent code-switched acoustic models for further evaluation and performance analysis. We hope to increase our speech corpus for further evaluation. Further evaluation of false reject (FR) and false accept (FA) rates on per-utterance basis will be explored.

## VIII. REFERENCES


[1] J. Weiner, N. T. Vu, D. Telaar, F. Metze, T. Schultz, D. –C. Lyu, E. Chng, and H. Li. "Integration of Language Identification into a Recognition System for Spoken Conversations Containing Code-Switches," In Proc. of *SLTU 2012*, Cape Town, South Africa, pp.1-4, May 2012.

[2] M. P. Lewis, G. F. Simons, and Charles D. Fenning (Eds). 2013. Ethonologue Languages of the World, Seventeenth edtion. Dallas, Taxas SIL Internationa. Online version: http://www.ethonologue.com/world, Accessed on: 12 August 2013.

[3] HLT, http://www.meraka.org.za/humanLanguage.htm, Accessed on: 06 June 2013.

[4] H. Li, K. A. Lee and B. Ma, "Spoken Language Recognition: From fundamentals to practice", In Proceedings of IEE, Vol.101 (5), pp.1136-1159, December 2013.

[5] R. Tong, B. Ma, D. Zhu, H. Li, and E. S. Chng, "Integrating acoustic, prosodic and phonotactic features for spoken language identification," In *Proc. ICASSP*, pp. 205–208, 2006.

[6] E. Ambikairajah, H. Li, L. Wang, B. Yin, and V. Sethu, "Language Identification: A Tutorial", *In IEEE Circuits and Systems Magazine*, Volume: 11, Issue: 2, pp.82-108, 2011.

[7] W. Zhirong, U. Topkara, T. Schultz and A. Waibel, "Towards Universal Speech Recognition," *In Proc. ICMI 2002*, Pittsburgh, 2002.

[8] D. -C. Lyu and R. -Y. Lyu, "Language Identification on Code-Switching Utterances Using Multiple Cues," *In Proceedings of INTERSPEECH*, Brisbane, Australia, pp. 711-714, September 2008.

[9] C. -H. Wu, Y. -H. Chiu, C.-J. Shia and C. -Y. Lin, "Automatic Segmentation and Identification of Mixed-Language Speech Using Delta-BIC and LSA-Based GMMs," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 266-276, January 2006.

[10] K. Bhuvanagiri and S. K. Kopparapu, "Mixed Language Speech Recognition without Explicit Identification", *American Journal of Signal Processing 2012*, Vol. 2, Issue 5, pp. 92-97, 2012

[11] T. Niesler and D. Willett, "Lanaguge identification and multilingual speech recognition using discriminatory trained acoustic models," In Multiligual Speech and Language Processing, 2006

[12] T. I. Modipa, M. H. Davel and F. de Wet, Context-dependent modelling of English vowels in Sepedi code-switched speech, In Proc. of *PRASA 2012*, Pretoria, South Africa, November 2012

[13] C. J. Shia, Y. H. Chiu, J. H. Hsieh and C. H. Wu, "Language Boundary Detection and Identification of Mixed-Language Speech Based on MAP Estimation," In Proc of ICASSP, pp. 381-384, 2004.

[14] C. –H Wu, H. –P. Shen and Y. –T. Yang, "Phone set contruction based on context-sensitive articulatory attributes for code-swiching recognition," In proc. ICASSP, pp.4865-4868, 2012

[15] D. -C. Lyu, R. -Y. Lyu, C. -L. Zhu and M.-T. Ko, "Language Identification In Code-switching Speech Using Words-based Lexical Model"Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium, Tainan, pp. 460-464, December 2010.

[16] W. M Campbell, JP Campbell, DA Reynolds, E Singer, "Support vector machines for speaker and language recognition", Computer Speech & Language, 20, pp.210-229, 2006.

[17] M. Peche, M. Davel, and E. Barnard, "Development of a spoken language identification system for South African languages," SAIEE Africa Research Journal, Vol. 100(4), pp. 97-105, December 2009.

[18] LWAZI, http://www.meraka.org.za/lwazi/, Accessed on: 06 June 2013.

[19] S. J. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, Cambridge University,2002. (For HTK Version 3.2.1), http://htk.eng.cam.ac.uk, Accessed on: 05 May 2013.

[20] Stolcke, "SRILM - An extensible language modeling toolkit", In Proc. ICSLP, Denver, CO, pp. 901-904, November 2002

[21] C, -C. Chang and C. -J. Lin, LIBSVM-A library foe support vector machine, 2001, http://www.csie.ntu.edu.tw/~cjlin/libsvm, Accessed on: 29 April 2013.

[22] M. A. Zissman., "Automatic language identication of telephone speech". Lincoln Laboratory Journal, 8(2):115-144, 1995.



**Koena Ronny Mabokela** received his Honors in Computer Science in 2011 from University of Limpopo. He is completing his MSc Computer Science degree at the same institution. He is now working at Telkom SA in NIP: Technical Product Development section as an Operational Specialist. He has a keen research interests to Broadband and Network Services and Speech Technologies.

# Rendering South African Sign Language sentences from SignWriting notation

Kenzo Abrahams[1], Mehrdad Ghaziasgar[1], James Connan[2] and Reg Dodds[1]
Department of Computer Science
University of Western Cape[1], Private Bag X17 Bellville, 7535, South Africa
Tel: +27 21 959-3010, Fax: +27 21 959-3006
and Department of Computer Science
Rhodes University[2], PO Box 94 Grahamstown, 6140
Tel: +27 46 603-8291, Fax: +27 46 636-1915
email: 2831775@myuwc.ac.za, mghaziasgar@uwc.ac.za[1]; j.connan@ru.ac.za[2]; rdodds@uwc.ac.za[1]

**Abstract— The 2011 national census statistic suggests that an estimated 520 000 Deaf people use South African Sign Language as their main communication medium. The SASL group of the University of the Western Cape developed a system that animates a human avatar model to sign out simple phrases expressed in SignWriting notation. This research proposes an extension to the existing system by extending it to render entire sentences. Sign Writing Markup Language (SWML), an XML-based notation for writing SignWriting, is provided as input. Animation parameters used to animate the avatar are generated from the SWML. Experimentation on 7 sentences shows that the resulting system achieved 97% accuracy for hand shapes. The facial expressions rendered achieved 95% accuracy. Movements performed by the avatar achieved 91% accuracy.**

**Index Terms—Avatar, Deaf, Sign Language animation, Animation parameters**

## I. INTRODUCTION

The 2011 national census of South Africa estimated that there are nearly 5 million people who are Deaf or hard of hearing, with at least 520 000 of these people being profoundly deaf and using South African Sign Language (SASL) as their main communication medium [1]. Sign Languages make use of manual, facial and other body movements to communicate in a visual-gestural modality, rather than orally. The difference in modality between Sign Languages and spoken languages often creates a communication barrier between Deaf and hearing individuals. This becomes much more of an issue when one realizes that in South Africa there are no public services, such as medical assistance, offered in SASL.

The South African Sign Language (SASL) research group at the University of the Western Cape (UWC) has made significant contributions towards the innovation in technologies to aid translation between SASL and English. The group is in the process of developing a machine translation system in an attempt to allow the Deaf community to access information in their native language [2]. The system involves two distinct procedures: translating from SASL to English; and translating from English to SASL.

The procedure relevant to this research is the translation of English to SASL. Translation of English to SASL involves the extraction of semantic information from English speech to visualize SASL phrases. To visualize SASL phrases, van Wyk created a 3-dimensional humanoid avatar that can perform physically realistic movements [3]. Moemedi used van Wyk's avatar to create the current sign language animation system used by the SASL group [4].

The system receives input in the form of SignWriting, a Sign Language transcription notation. The avatar is animated to express SASL phrases represented in SignWriting notation. The system can currently process basic SignWriting movement symbols such that only simple SASL phrases can be animated.

This paper discusses the extension of Moemedi's system to allow full sentences to be rendered from the SignWriting notation. With this extension, the avatar is able to render phrases with multiple complex motion paths.

The rest of the paper is organized as follows: Section II discusses related work; Section III gives an overview of the SignWriting Markup Language and its structure; Section IV discusses how the animation parameters are generated from the SignWriting Markup Language; Section V discusses how the animations are produced; Section VI discusses the experimental setup used to test the system, the results of which are presented in Section VII; and Section VIII concludes the paper.

## II. RELATED WORK

This section first discusses the Sign Language animation system developed by Moemedi, which is improved upon in this research. The second sub-section discusses a similar system which animates sign language phrases from SignWriting notation.

### *A. SignWriting animation for South African Sign Language*

Moemedi created a system which can animate Sign Language phrases using the avatar created in Blender by van Wyk [4]. The system uses SignWriting as input. SignWriting is a notation created by Valerie Sutton in 1974 for the Centre of the Sutton Writing [5]. The symbols which are used to make up signs are abstractions of a signer's hands, face and body movements as he/she is performing Sign Language [5]. These symbols are converted to SignWriting Markup Language (SWML), an XML-based

format developed for the storage, indexing and processing of the SignWriting notation. The input process is handled by SignText editor, an interface for creating SignWriting notation pictographs.

The XML file is parsed and information pertaining to the structure of the sign is extracted using Python's SAX API. Each symbol is represented by a unique 6-tuple ID which represents all the transformations the symbol went through at the time that it was created. The SWML is analysed and categorized into different kinematic problems dependent on the symbols present in the sign.

The avatar is compliant with the humanoid animation (H-anim) standard, a specification for defining human modeling and animation. The avatar's face is encoded using the MPEG-4 facial definition parameters (FDP) to allow facial expressions to be rendered [4].

To animate a sign, the system defines key frames from the SWML. These key frames include both body animation parameters (BAPs) and face animation parameters (FAPs). The BAPs contain information about how the avatar's joints should be rotated in order to reproduce the SASL represented by the input. The FAPs define the position of feature points on the face which allow different facial expressions to be rendered. Once all the key frames have been generated, Blender's quaternion interpolation method is used to produce a smooth animation of the sign from the key frames.

Ten evaluators knowledgeable in both SignWriting and American Sign Language (ASL) were invited to test the system via the internet. Eight ASL phrases were produced by the system using SignWriting notation representations of the phrases. These phrases were: 'thank you', 'hello', 'now', 'same', 'food', 'house', 'understand' and 'time'. A high recognition accuracy of 92% was achieved across all the phrases when evaluators were asked to indicate whether the animation was a correct representation of the SignWriting.

This system will be extended using a similar implementation to allow signs to be strung together creating sentences.

### *B. Synthesis of Virtual Reality from SWML using MPEG-4 Body Animation Parameters*

Papadorgiorgaki *et al.* developed a system that synthesizes Greek Sign Language from SignWriting notation at the Informatics and Telematics Institute in Greece [6]. The system only processes hand shape, hand motion and facial expression symbols. Only simple phrases can be rendered with this system.

The signs are converted into SWML and analyzed by the system. Information relating to supported symbols are used to calculate MPEG-4 body and face animation parameters.

The system determines the total number of key frames that need to be produced, based on the number and nature of the available movement symbols. A single key frame is produced if no movement symbols are present in the sign. On the other hand, multiple animation key frames are produced if symbols describing dynamic information exist.

Once all the key frames have been computed, interpolation is used to increase the frame rate of the key frames sequence. This results in a smooth transition between key frames. This system uses linear interpolation, a method of curve fitting which makes use of linear polynomials to create curves. Linear interpolation dictates a constant rate of change from one key frame to the next.

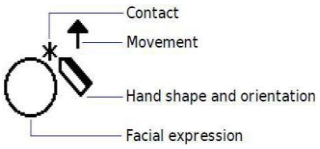
To visualize the animation of the sign, the system makes use of EPLFBody BAP player for the avatars body and a Miraface FAP player to render facial expressions. The EPLFBody BAP player developed by the École Polytechnique Fédérale Lausanne (EPFL) for the Synthetic and Natural Hybrid Coding (SNHC) subgroup of MPEG-4 is used to animate the body of H-Anim compliant avatars using the generated key frame sequence [6]. To animate the face of the avatar, the Miraface FAP player, also developed for the MPEG-4 SNHC, was used.

No experimental procedure or results are documented for the system. The system is related to this research since it produces animations from SignWriting focusing on hand shape, hand movement and facial expressions. It also uses the key frame animation technique to render the resulting animations. However, the system can only render simple phrases from SignWriting. The animations are also not as smooth as Moemedi's system due to the interpolation method.

## III. SIGNWRITING MARK-UP LANGUAGE

In order to fully understand how the proposed system generates the animation key frames, the structure of SignWriting and SWML are described. The design concept for SignWriting was to represent movement as they are visually perceived, rather than representing the meaning that these movements convey. Almost all international Sign Languages, including American Sign Language (ASL) and SASL can be represented in SignWriting.

The symbols in SignWriting are constructed in a pictograph called a signbox. The symbols are place into the pictograph to construct the sign. The top of Figure 1 depicts the SignWriting for the SASL translation of the phrase "hello".




```
<?xml version="1.0"?>
<!DOCTYPE swml SYSTEM "http://www.signpuddle.com/swml/swml
  <swml dialect="S" version="1.1" lang="sgn" glosslang="">
    <sign lane="0">
      <gloss>sign-1</gloss>
      <symbol x="90" y="68">03-03-002-01-01-01</symbol>
      <symbol x="130" y="51">02-05-001-01-01-01</symbol>
      <symbol x="137" y="73">01-05-012-01-02-02</symbol>
      <symbol x="123" y="64">02-01-001-01-01-01</symbol>
    </sign>
  </swml>
```

Figure 1: SASL translation of "hello" along with the resulting SWML

Each symbol represents a specific aspect relating to the structure of sign language. The symbols are divided into 7 categories. These are: hand shapes, movements, facial expressions, trunk and limbs, dynamics, punctuation and location sorting [5].

The graphical nature of SignWriting makes it unsuitable for computer processing. For SignWriting to be an efficient computer notation system, it should facilitate tasks such as storage, processing and even the indexing of signs. For this

purpose, SWML, an XML-based format for representing signs in SignWriting, has been proposed [5]. Due to its XML nature it is application-and computer platform-independent and can be used in language and document processing such as translation, animation, etc [6]. The bottom of Figure 1 depicts the SWML equivalent of the SignWriting phrase "hello".

Each symbol can be identified by a unique 6-tuple identifier in SWML. This identifier is used to determine which aspect of sign language a symbol relates to. The 6-tuple symbol identifier is defined as:

$$S_{id} = (c, g, b, v, f, r)$$

where $c$ is the category that the symbol belongs to, $g$ is the group that represents the basic shape of the symbol within the category, $b$ is the base and specifies which symbol within a group is being referred to, $v$ is the variation parameter specifying possible variations of the symbol, $f$ is the fill parameter specifying how the symbol is filled and $r$ specifies the rotation applied to the symbol in steps of 45°. Note that the values for these parameters are integers. $S_x$ and $S_y$, respectively, denote the $x$ and $y$ position of the symbol in the signbox. $S_{length}$ and $S_{width}$ denote the length and width of a sign, respectively. Thus, the symbol's full definition is:

$$S = (S_{id}, S_x, S_y, S_{length}, S_{width})$$

Signs are composed of a set of symbols as follows:

$$Sign = \{S_1, S_2, \ldots, S_n\}$$

By merging signs and animating them consecutively, sentences can be formed.

$$Sentence = \{sign_1, sign_2, \ldots, sign_n\}$$

The symbol identifiers are used to generate animation parameters which are used to animate the avatar. The following section explains how the symbol identifier is used to generate the body and face animation parameters required to perform the sign.

## IV. Converting SWML into Animation Parameters

This section describes how the symbol identifier is used to generate animation parameters. A parser is implemented to handle this procedure. The parser serves two purposes. First, it is responsible for checking whether the symbols are used correctly in the sign and if the sign is valid. For example if a sign contains a movement symbol relating to the left hand and there is no left hand present in the sign, an error occurs and that sign cannot be processed.

Second, the parser is tasked with extracting information from the symbol identifiers to generate the resulting body and face animation parameters. In order to do so the parser needs to interpret the SWML input. Information provided by each field in the symbol identifier is used to specify the animation parameters of the corresponding joints. The improved system can interpret symbols from three categories namely; hand shapes, facial expressions and movements. The following subsections describe the generation of animation parameters for each of the three symbol categories by the parser.

### A. Hand Category Symbols

The basic symbols in SignWriting can go through multiple transformations when added to a sign pictograph. These transformations include rotation, change in orientation, as well as mirroring. All the basic hand shape symbols are right handed by default, with no rotation applied to them. Mirroring the right hand symbol results in the left hand version of that symbol. The *rotation* (r) field indicates whether or not a symbol is mirrored. The rotation field can have a value from 1 to 16, where a symbol refers to the right hand if $1 \leq r \leq 8$ and refers to the left hand if $9 \leq r \leq 16$ as seen in Figure 2.

| Rotation | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Right Hand |  |  |  |  |  |  |  |  |
| Rotation | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| Left Hand |  |  |  |  |  |  |  |  |

**Figure 2:** The angles of rotation for both right and left hand.

The *rotation* field also indicates the degree to which the hand is rotated. This rotation occurs in steps of 45°. The right hand is rotated anti-clockwise and the left hand is rotated clockwise. Figure 2 depicts all possible rotations for both the left and right hands.

If the system identifies a hand symbol in a sign, it needs to determine animation parameters for the finger joints to produce the hand shape. To determine the actual hand shape of a symbol, the first 4 parameters of the symbol identifier are used. These specify the rotations of the finger joints to produce the hand shape.

A predefined lookup table defines animation parameters for the finger joints. The combination of the first 4 parameters of the symbol identifier are used as keys for the lookup table. The avatar allows both forward and inverse kinetics to be used such that bones can be rotated to create poses [7]. Rotational constraints are applied to the bones of the avatar to ensure that only humanly possible poses can be achieved [3].

Once the animation parameters for the finger joints are determined, the system then needs to determine the orientation of the hand. There are two types of hand orientations in SignWriting. The first is concerned with the direction in which the palm of the signer is facing. This is depicted in Figure 3. This orientation is viewed from the signer's point of view. The *fill* field in the symbol id identifies the orientation and can be an integer value between 1 and 6. If $f = 1$ or $f = 4$ then the palm is facing the signer. If $f = 2$ or $f = 5$ the signer sees the side of his hand. If $f = 3$ or $f = 6$ the back of the hand is visible to the signer.

**Figure 3:** The possible directions the signer's palm can face.

The second type of orientation is concerned with the plane to which the hand is parallel. The hand can be parallel to either the floor or the wall plane as can be seen in Figure 4. This orientation is determined by the *fill* field of the symbol id as well. If $1 \leq f \leq 3$, the hand is parallel to the wall plane and if $4 \leq f \leq 6$, the hand is parallel to the floor plane.

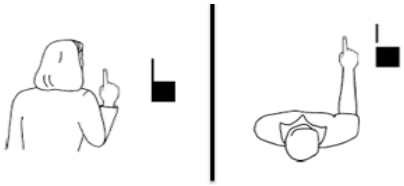


**Figure 4:** Orientation of the hand along the wall plane and along the floor plane

The *fill* and *rotation* fields are used to determine the animation parameters for the elbow and shoulders to achieve the desired orientation and rotation of the hand. The system processes all hand symbols using this approach and stores the resulting animation parameters.

### B. *Facial Expression Category Symbols*

Facial expressions are a very important aspect in sign language because they express grammar and portray feeling and interest towards the subject being communicated [8]. Examples of facial expressions in sign language include raising of eyebrows, blinking eyes and the shape of the mouth. Changing the facial expression of a sign can potentially change the meaning of a sign.

The symbols relating to facial expressions encode information about the parts of the face that move and how this movement should occur. The system uses an approach similar to the previous category symbols to determine the FAPs. The face symbol is identified using the combination of the first 4 parameters of the symbol identifier. The *rotation* field indicates in which direction the part of the face needs to move.

In cases where the facial expression refers to eyes, cheeks, eyebrows and ears, the *fill* field is used to determine if the left, right or both parts must move. If $f = 1$ both parts are included in the facial expression. If $f = 2$ only the right part is included in the facial expression. If $f = 3$ only the left part is included in the facial expression.

Additional bones are added to feature points such as the eyes and eyebrows on the avatars face as shown in Figure 5 [3]. These bones have no limitations such that facial expressions can be easily performed. Figure 6 depicts two facial expressions produced by the avatar.

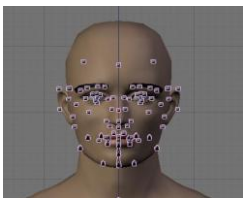


**Figure 5:** The "skeleton" of the avatar's face

Once the face is identified, along with the movements associated with it, the system generates the FAPs. A predefined lookup table is used which defines FAPs for the corresponding facial parts. After the FAPs are generated they are stored.

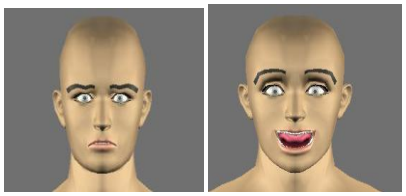


**Figure 6:** Example facial expressions performed by the avatar.

### C. *Movement Category Symbols*

Arrows are used in SignWriting to indicate movement paths. There are a variety of arrows indicating movements associated with the left hand, right hand or both hands moving simultaneously. Figure 7 depicts different types of movements for the right hand, left hand and both hands along the different planes.

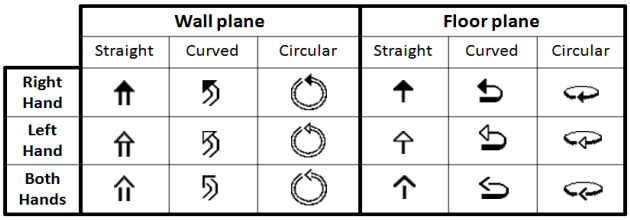



**Figure 7:** Different type of movements in SignWriting

The different types of movements are identified using the combination of the first 4 parameters of the symbol identifier. Using this combination, it is determined if the movement in question is a straight movement, a curve, or circular movement. The first 2 parameters of the symbol id indicate whether the movement is along the wall or floor plane.

The *fill* field indicates to which hand the movement is associated. If $f = 1$, the movement is associated with the right hand. If $f = 2$, the movement is associated with the left hand. If $f = 3$, the movement is associated with both the right and left hands. As with the hand symbols, the movement symbols can also be rotated in $45^{\circ}$ increments. This determines the direction of the movement and is indictated by the *rotation* field of the symbol id. The movements can be of varying lengths. To distinguish between shorter and longer movements, the *variation* field in the symbol id is used.

To produce the desired movement, the system makes use of key framing [9]. In this technique, the starting point and the end point of the movement are defined. Interpolation is then used to produce the in-between frames resulting in a smooth transition from the start to end points. Different movement types require varying numbers of key frames to be defined.

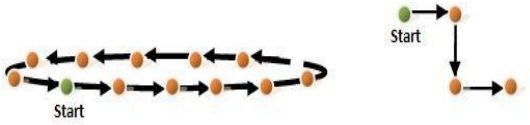



**Figure 8:** Circular movements require more key frames to be defined to produce the movement.

A straight movement requires two key frames, the starting point and the end point. However, a circular movement requires more key frames to produce a circular motion animation as shown in Figure 8. The movement category also contains contact symbols which are processed using this approach. Once all the movement symbols have been processed, the BAPs are generated and stored.

## V. ANIMATION

All the symbols provided as input to the system in SWML format are converted to animation parameters. The animation is represented as a series of key frame poses, facial expressions and hand shapes. The system uses quaternion interpolation to acquire a smooth transition between the frames. This results in the avatar performing the sign that corresponds to the input which is displayed on the screen. The system starts in a neutral position. Only once all the signs have been performed does the avatar return to the neutral position.

## VI. EXPERIMENTAL SETUP

The experiment carried out on the system intended to determine the accuracy of the sentences produced by the avatar. In terms of this research, accuracy is defined as whether or not animations produced from SWML resemble SASL videos. Note that accuracy is regardless of the underlying meaning of the sentence. Based on this, animations were produced from SignWriting notation input to represent corresponding SASL videos received from the Linguistics Department at Rhodes University. These videos contain proficient signers signing various sentences of varying lengths. The sentences chosen for testing emphasize hand shapes, facial expressions as well as different types of movement that can be rendered by the avatar.

The SASL videos were then displayed, at half speed alongside the rendered videos, to evaluators that were completely unknowledgeable in Sign Language to determine whether or not the two videos matched. The reason for selecting evaluators that do not understand sign language was to determine if the rendered animations match the videos of the real signers, without worrying about the underlying linguistic aspects. Using unknowledgeable evaluators also results in a more stringent test, since small variations that could be considered acceptable in Sign Language could be seen as a mismatch. In future, a test will be carried out to evaluate understandably of rendered Sign Language using Deaf individuals.

Each evaluator was shown each SASL video alongside the rendered animation side-by-side three times. The first time, only the hand shape of the video was evaluated. The second and third times evaluated the facial expression and hand motions in the animation respectively. In each case, the entire animation was marked as either being a "Match" or "Mismatch".

Therefore, this resulting in a total of 8 evaluations per sentence for each parameter, and $7 \times 8 = 56$ evaluations per parameter. An accuracy per sentence was determined by computing the total number of matches for that sentence as a percentage of the total number of evaluations for each parameter.

The results of the hand shape, facial expression and hand motion accuracy are provided and discussed in the next section.

TABLE I: DETAILS OF EVALUATORS

| Evaluator | Age | Gender |
|---|---|---|
| **1** | 21 | Male |
| **2** | 21 | Male |
| **3** | 46 | Female |
| **4** | 25 | Male |
| **5** | 49 | Male |
| **6** | 32 | Female |
| **7** | 24 | Male |
| **8** | 16 | Female |

Seven sentences provided by Rhodes University were used. Table II summarizes the sentences used.

TABLE II: SUMMARY OF SENTENCES USED TO TEST THE SYSTEM

| Sentences | Number of hand symbols | Number of face symbols | Number of motions |
|---|---|---|---|
| All students will pass | 4 | 2 | 4 |
| He teaches the child | 3 | 1 | 3 |
| I love you | 1 | 1 | 1 |
| I want to get a balloon | 4 | 3 | 4 |
| I will give you a balloon | 5 | 3 | 4 |
| The cow is being milked | 2 | 1 | 2 |
| The woman milks the cow | 4 | 2 | 4 |
| Total | 23 | 13 | 22 |

## VII. RESULTS AND ANALYSIS

### A. *Hand shape accuracy*

TABLE II: ACCURACY ACHIEVED FOR HAND SHAPE FOR EACH SENTENCE

| Sentence | Hand shape accuracy (%) |
|---|---|
| All students will pass | 91 |
| He teaches the child | 100 |
| I love you | 100 |
| I want to get a balloon | 94 |
| I will give you a balloon | 100 |
| The cow is being milked | 100 |
| The woman milks the cow | 97 |
| Average | 97 |

As can be seen in Table III, the system renders hand shapes accurately. Four sentences achieved a hand shape accuracy of 100%, with the rest achieving an accuracy of no less than 90%. The first sign of the sentence "all students will pass" contains a contact. The hand shapes used in this sentence cause the contact to appear as though one hand is behind the other. Some of the evaluators noticed this

discrepancy and marked the hand shape as incorrect. Regardless, this result is very encouraging because it suggests that the system is capable of reproducing multiple hand shapes accurately in one sentence.

### B. *Facial expression accuracy*

TABLE IIV: ACCURACY ACHIEVED FOR FACIAL EXPRESSIONS FOR EACH SENTENCE

| Sentence | Face accuracy (%) |
|---|---|
| All students will pass | 94 |
| He teaches the child | 96 |
| I love you | 100 |
| I want to get a balloon | 100 |
| I will give you a balloon | 100 |
| The cow is being milked | 75 |
| The woman milks the cow | 97 |
| Average | 95 |

As can be seen in Table IV, the facial expressions produced by the avatar are highly accurate. Six sentences achieved an accuracy higher than 90%. Three out of the six sentences achieved a 100% accuracy. Only one sentence achieved an accuracy lower than 90%. For the sentence “The cow is being milked” the signer says the word cow while performing the sign. Some of the evaluators looked for the shape of the mouth during this sign as it was being performed by the avatar. The facial expression used in the animation does not resemble a person saying the word cow, thus the facial expression was marked as a mismatch in such cases.

### C. *Movement accuracy*

TABLE III: ACCURACY ACHIEVED FOR MOVEMENTS FOR EACH SENTENCE

| Sentence | Movement accuracy (%) |
|---|---|
| All students will pass | 94 |
| He teaches the child | 83 |
| I love you | 100 |
| I want to get a balloon | 78 |
| I will give you a balloon | 91 |
| The cow is being milked | 93 |
| The woman milks the cow | 97 |
| Average | 91 |

As seen in Table V, the movements produced by the avatar are of an accurate nature, with 5 out of the 7 sentences receiving an accuracy of higher than 90%. One sentence achieved a 100% accuracy. The sentence "I want to get a balloon" contains a change in orientation of the right hand which makes the movement appear to be distorted. This results in an unnatural movement being produced by the avatar. The evaluators noticed this unnatural movement and marked the sentence as incorrect. Nevertheless, it achieves a high accuracy of 78%.

## VIII. CONCLUSION

This paper presents an extension to Moemedi's sign language animation system which allows full sentences to be rendered by the avatar from SignWriting notation input. The system renders sentences by allowing more than one sign to be animated at a once. Animation parameters are generated from the SWML of the SignWriting notation input. Experimentation shows that the system is very capable of rendering sentences that look as natural as a human signer.

## IX. REFERENCES


[1] Stats SA Library Cataloguing-in-Publication (CIP) Data, Census 2011 Statistical release – P0301.4 / Statistics South Africa. Pretoria: Statistics South Africa, 2012

[2] N. L. Naidoo, and J. Connan "Gesture recognition Using Feature Vectors ", in *the Southern Africa Telecommunication Networks and Applications Conference*, Royal Swazi Spa, Swaziland, 2008.

[3] D. E. van Wyk, Virtual Human Modeling and Animation for Sign Language Visualization, MSc thesis, University of the Western Cape

[4] K. A. Moemedi,. (2010). *Rendering an Avatar from Sign Writing Notation for Sign Language Animation*, MSc Thesis, University of the Western Cape.

[5] A. C. da Rocha Costa, & G. P. Dimuro, (2001). A SignWriting-based approach to sign language processing. In *Proceedings from GW 2001: Gesture Workshop* (pp. 202-212).

[6] M. Papadogiorgaki, N. Grammalidis, N. Sarris, & M.G. Strintzis, (2004, May). Synthesis of virtual reality animations from SWML using MPEG-4 body animation parameters. In *Workshop on the Representation and Processing of Sign Languages, 4th International Conference on Language Resources and Evaluation LREC 2004*.

[7] D. van Wyk, & J Connan, High Quality Flexible H-Anim Hands for Sign Language Visualization.

[8] V. Sutton (2000, January). Sutton Movement Writing & Shorthand: A global writing system for a global age. [Online]. Available: http://www.valeriesutton.org/

[9] N. Burtnyk and M. Wein, Interactive skeleton techniques for enhancing motion dynamics in keyframe animation. Communications of the ACM, 19, 564-584.



**Kenzo Abrahams** is an M.Sc student at the University of the Western Cape Department of Computer Science. He is currently doing research on sign language synthesis and novel communication applications on mobile interfaces for the Deaf and hearing impaired.

**Mehrdad Ghaziasgar** is the project manager of the South African Sign Language (SASL) research group. He has a wide range of interests that include internet programming, mobile computing, computer vision and machine learning.

**James Connan** heads up the South African Sign Language (SASL) research group. His interests are computer vision and machine learning.

**Reg Dodds** is a lecturer at the University of the Western Cape. His interests are include machine learning, algorithms and complexity theory.

# Graphemes and Phonemes as Acoustic Sub-word Units for Continuous Speech Recognition of Under-resourced Languages

Mabu Manaileng[1], Jonas Manamela[2]
Department of Computer Science
University of Limpopo, Private Bag X1106, Sovenga 0727
Tel: +27 15 2682155, Fax: +27 15 2683487
email: ManailengMJ@gmail.com[1]; Jonas.Manamela@ul.ac.za[2]

**Abstract-This paper presents comparative results from using both graphemes and phonemes as acoustic sub-word units for automatic speech recognition (ASR) experiments on three official and under-resourced languages of South Africa, namely, Sepedi, IsiNdebele and Tshivenda. We compare the performance of grapheme-based recognition systems to that of phoneme-based recognition systems for monolingual and cross-lingual speech recognition. For each language, a hidden Markov model (HMM) trained ASR system was developed using both graphemes and phonemes as sub-word units. The employed framework models emission distributions by eight Gaussian Mixture Models (GMMs) with two mixture increments. Identical speech datasets were used for each experiment per language. The need to use graphemes was influenced by the under-resourced nature of the languages in question. When little or no expert knowledge about the language(s) of interest is available or is unaffordable, it becomes important to use graphemes, instead of phonemes as sub-word acoustic language modeling units. The performance of all systems was evaluated on the word error rate (WER). The grapheme-based approach showed superiority on the phoneme-based approach by attaining a WER reduction of 0.16% for IsiNdebele, and 2.35% for Tshivenda. However, the grapheme-based approach suffered a WER increment of 0.73% for Sepedi and 2.37% in the cross-lingual system.**

**Index Terms—ASR, grapheme, phoneme, WER, under-resourced language.**

## I. INTRODUCTION

The performance of ASR systems is heavily influenced by the comprehensiveness of a pronunciation dictionary used during training and in the decoding process [1]. A pronunciation dictionary provides a mapping of words to a sequence of sub-word units for each entry in the vocabulary. The sub-word units are then used to model the acoustic realization of the words. Phonemes are the most commonly used sub-word units and have shown notable successes in the development of ASR systems [2, 3]. However, the use of graphemes as sub-word units achieves comparable recognition results [1-3].

The use of hand-crafted pronunciation dictionaries raises problems when dealing with rare and under-resourced languages [1]. Many of these languages are only spoken by few people (although they may be their first languages) and thus it becomes nearly impossible to afford the creation of hand-crafted dictionaries. Furthermore, there are two kinds of problems that can be introduced by a pronunciation dictionary. The first one can be introduced during training through a false mapping between a word and its modeling units, which as a result, will contaminate the acoustic model set. The models will not describe the actual or consistent acoustics that they ought to represent. Secondly, the incorrect mapping will falsify the scoring of hypotheses by applying the wrong models (i.e., the reference tokens in the models) to the score calculation.

In cases were no linguistic expert knowledge is available or affordable for hand-crafting pronunciation dictionaries, new methods are needed to solve this problem by automating the mapping process. However, even the automatic tools often require hand-labeled training materials and rely on manual revision and verification. Several different methods to automate the process have been introduced in the past [4-7]. Most of these methods are based on finding rules for the conversion of the written form of a word to a phonetic transcription, either by applying rules [4] or by statistical approaches [5].

Recently, the use of graphemes as acoustic modeling units – instead of phonemes, has been increasingly studied. Graphemes have the advantage over phonemes in that they make the creation of a pronunciation dictionary a trivial task that does not require any linguistic knowledge [1]. However, the generally looser relation of graphemes to pronunciation than that of phonemes necessitates the use of context-dependent modeling techniques and the sharing of parameters for different models [1, 7].

Prior ASR experiments have shown that the quality of grapheme-based recognizers is highly dependent on the nature of the grapheme-to-phoneme relation of a specific language. Schukat-Talamazzini et al. [8] and Kanthak [7] were some of the first persons to present results for speech recognition systems based on the orthography of a word. Black and Llitjos [9] successfully relied on graphemes for text-to-speech synthesis systems in minority languages. As reported in [2, 3], both studies investigated the use of

graphemes for languages with phoneme-grapheme relationships of differing closeness in the context of cross-lingual speech recognition. All these experiments have shown that for certain languages graphemes are suitable modeling units for speech recognition.

Our study is aimed at investigating the potential of using graphemes, instead of phonemes, as sub-word units for automatic speech recognition using selected under-resourced languages of Limpopo province in South Africa. Using graphemes instead of phonemes as sub-word units for automatic speech recognition is likely to reduce the cost and time needed for the development of satisfactory ASR systems for our targeted languages. We attempted to address the problem of creating pronunciation dictionaries in a non-optimal manner with respect to time and cost. The approach of developing ASR systems that rely solely on graphemes rather than phonemes as sub-word units is thus adopted. The mapping in the dictionary now becomes completely trivial, since every word is simply segmented into its constituent alphabetic letters. An expert knowledge of sub-word units of words is therefore no longer needed. The quality of grapheme-based ASR systems depends significantly on the grapheme-to-phoneme relationship of the language, that is, the degree of relatedness between *how words are pronounced* (articulation) and *how they are written* (orthography) [2, 3]. Therefore, our study also investigates the grapheme-to-phoneme relationship of our targeted under-resourced language(s).

Firstly, three different monolingual ASR systems are presented, each with two distinct experiments. The systems are differentiated according to their language. Each language is experimented upon using both phonemes and graphemes as sub-word units. The speech corpora used are identical per language (i.e. system) but each experiment has a unique pronunciation dictionary. The recognition results of each experiment per ASR system are compared and analyzed. Secondly, a cross-lingual ASR system with two distinct experiments is presented. The only difference between the two experiments is the pronunciation dictionaries, which are differentiated by their acoustic sub-word units. The results of the two cross-lingual ASR experiments are also analyzed and compared.

The remainder of this paper is structured as follows: Section II discusses the concept of under-resourced languages. Section III discusses some of the works related to what we are presenting. Section IV presents the data sets used, focusing primarily on the construction of the pronunciation dictionaries and speech databases. Experiments and results are described and presented in Section V. Finally, the paper is concluded with a summary of the main observations and future work in Section VI.

## II. UNDER-RESOURCED LANGUAGES

Under-resourced languages are characterized as languages with some of (if not all) the following attributes: lack of a unique writing system or stable orthography, limited presence on the web, lack of linguistic expertise, lack of electronic resources for speech and language processing – such as monolingual corpora, bilingual electronic dictionaries, transcribed speech data, pronunciation dictionaries and vocabulary lists [10]. The term is often used interchangeably with: resource-poor languages, less-resourced languages, low-data languages or low-density languages.

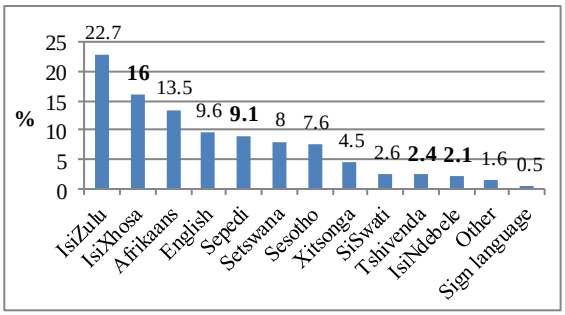
The concept of under-resourced languages should not be confused with that of minority languages – which are languages spoken by a minority of the population in a territory [10]. Some under-resourced languages are actually official languages of their countries and at times spoken by a very large population. However, some minority languages can often be considered as rather well-resourced. Consequently, under-resourced languages are not necessarily endangered whereas minority languages may be endangered [10].

### A. *Languages of South Africa*

South Africa is a highly diversified country with eleven official languages and very wide social and cultural disparities. Table 1 shows the distribution of languages across the population, the bolded languages in Table 1 form the core research focus in the Telkom Centre of Excellence for Speech Technology (TCoE4ST) at the University of Limpopo.

Various speech corpora for South African languages have been created and released in recent years, including the LWAZI telephone speech corpora [11] and National Centre for Human Language Technologies (NCHLT) [12] speech corpora – substantial collection of large broadband speech corpora. These corpora are all focused on the eleven official languages in South Africa. In recent years, speech technology projects that bridge language barriers while also addressing social issues have achieved substantial attention in South Africa [13].

TABLE 1. THE USE OF SOUTH AFRICAN LANGUAGES AND FOCUS AREA OF THE STUDY[14]




### B. *Languages of South Africa*

All eleven official languages of South Africa can be found in the Limpopo province [14]. Most official languages in Limpopo are considered under-resourced, due to their lack of speech processing tools and resources such as speech corpora, linguistic experts and pronunciation dictionaries – in most cases. Most of these languages are spoken by

comparatively few people. Only 2% of the population in the province speaks IsiNdebele as their first language, 16.7% speaks Tshivenda, 17% speaks Xitsonga and a staggering 52.9% speaks Sepedi. The other languages are shared across the remaining 11.4% of the population [14]. Although an account for all eleven languages does exist [11, 12], speech technology researchers often encounter problems regarding representation of various languages and their dialects together with perfectly hand-crafted pronunciation dictionaries.

## III. Multilingual Grapheme-based Acoustic Models

Research in the domain of ASR for under-resourced languages has focused on the efficient development of multilingual and cross-lingual grapheme-based ASR approaches that can make use of resources available in other languages. The use of multilingual grapheme models for rapid bootstrapping of acoustic models to new languages was investigated in [15, 16]. Data driven mapping of grapheme sub-word units across languages was studied in [15]. In [16], polyphone decision tree based tying for porting decision trees to a new language was applied for grapheme models. The study focused specifically on porting multilingual grapheme models to German and it was found to be beneficial compared to monolingual grapheme models when limited adaptation speech data for training is available.

As demonstrated in [7], grapheme-based acoustic units in combination with decision tree state tying may reach the performance of phoneme-based units for at least a couple of European languages. The approach is completely driven by the acoustic data and does not require any linguistic or phonetic knowledge. Grapheme-based multilingual acoustic modeling already provides a globally consistent representation of acoustic unit set by definition [2]. Global phoneme sets such as SAMPA or IPA [17] may be used to express similarities between languages when using phoneme-based acoustic sub-word units. However, the use of context-dependent grapheme-based sub-word units eliminates the need to find common sets of acoustic sub-word units.

For the purpose of comparing the recognition performance of the phoneme-based sub-word units to grapheme-based ones in a similar framework, the same procedure was followed to create the acoustic models in both the grapheme-based and phoneme-based experiments. Cross-lingual acoustic models were trained in both ASR experiments. This was done to ensure that acoustic data can be shared across languages. The use of context-dependent grapheme-based sub-word units, called trigraphs, for cross-lingual acoustic modeling was heavily motivated by the distribution of graphemes across the languages of interest.

## IV. Experimental Design

For the monolingual framework, experimental ASR systems were partitioned as follows: SystemSep for Sepedi (SEP), SystemIsi for IsiNdebele (ISI) and SystemVen for Tshivenda (VEN) to associate them with their respective languages. A complete and functional system was developed for each language, i.e. the study resulted in three different systems which are uniquely identified by their language of focus. Two ASR experiments (ExpPho and ExpGra) were performed for each of the three different systems. Each experiment per system is differentiated by a pronunciation dictionary, i.e., it is either grapheme-based or phoneme-based dictionary.

### *A. Monolingual and cross-lingual speech corpora*

The speech training and testing data in this study were obtained from the LWAZI project. Therefore, it was not necessary for us to record the speech data from scratch, but instead used the data set. The monolingual experimental speech data setup is outlined in Table 2 and the cross-lingual one in Table 3. The speech databases were divided into the training and testing sets. The 80:20 ratio was used to divide the databases into 80% of the total data dedicated for training and 20% for testing.

TABLE 2. AMOUNT OF SPEECH DATA PER MONOLINGUAL SYSTEM

| System Type | Number of Speaker | | Number of Utterances | |
|---|---|---|---|---|
| | Train | Test | Train | Test |
| SystemSep<br>Duration (hours) | 120 | 20 | 4512<br>6.96 | 1128<br>1.74 |
| SystemIsi<br>Duration (hours) | 119 | 20 | 4810<br>7.52 | 1203<br>1.88 |
| SystemVen<br>Duration (hours) | 120 | 39 | 4751<br>5.1 | 1188<br>1.2 |

Ultimately, a total of 5640 utterances from 140 speakers, 6013 from 139 speakers and 5939 from 159 speakers were composed for SystemSep, SystemIsi and SystemVen, respectively. Over 19 hours of speech data was dedicated for training the cross-lingual system. The system was tested on more than 4 hours of speech data. The cross-lingual speech corpus is merely a combination of the monolingual corpora.

TABLE 3. AMOUNT OF SPEECH DATA PER CROSS-LINGUAL SYSTEM

| | Train | Test |
|---|---|---|
| # Speakers | 359 | 79 |
| # Utterances<br>(Hours) | 14073<br>19.58 | 3519<br>4.85 |

### *B. Monolingual and cross-lingual phoneme-based pronunciation dictionaries*

All the phoneme-based dictionaries were also obtained from the LWAZI project, except for the cross-lingual dictionary which is merely a combination of all the monolingual dictionaries. All the dictionaries used are the original versions of the Lwazi dictionaries [21] that contain no pronunciation variations. The details of the dictionaries are outlined in Table 4.

As observed from Table 5, only 24 of the total 55 monophones are shared across the languages. A staggering 56% of the monophones are uniquely distributed across the languages. This is one of the most disturbing characteristic of phoneme-based pronunciation dictionaries and one of

their central weaknesses especially in the context of cross-lingual speech recognition.

TABLE 4. PHONEME-BASED DICTIONARY SETUP PER LANGUAGE

| | Language | | | |
|---|---|---|---|---|
| | SEP | ISI | VEN | MULTI |
| Unique Words | 3317 | 4754 | 2490 | 10184 |
| Monophones | 43 | 48 | 41 | 55 |
| Triphones | 6262 | 5353 | 4856 | 11803 |

Grapheme-based dictionaries on the other hand, become useful due to the good coverage of letter similarities across languages, as reflected in Table 7. This characteristic grants graphemes the superiority over phonemes when dealing with under-resourced languages. Albeit the shared graphemes may not represent the same acoustic realization across languages, they become important for directly using orthography as acoustic units, as is often the case when linguistic expertise is unavailable.

TABLE 5. USAGE OF PHONEMES ACROSS LANGUAGES

| Phoneme | Language | | | |
|---|---|---|---|---|
| | SEP | ISI | VEN | ∑ |
| E J N O a d_0Z f i j k_> k_h l m n p_> p_h r s tS_h t_> t_h ts_> u w | x | x | x | 24 |
| !_b !_bg !_bh K_b b_< h k kx tK_> tK_h |_b |_bg |_bh | | x | | 13 |
| BZ G kx_h l_a m_j o pS_h p_bS p_bs ps_h tl_> tl_h | x | | | 12 |
| J_b Z d_Z p_b sw zw | | | x | 6 |

### C. Monolingual and cross-lingual grapheme-based pronunciation dictionaries

All existing phoneme-based pronunciation dictionaries were converted to grapheme-based dictionaries. To ensure the minimal time, linguistic knowledge and cost required for generating the dictionaries, the conversion did not follow any predetermined rules. We strictly used the most straightforward method of generating pronunciation dictionaries as sequence of graphemes and thus directly using orthographic sub-word units as acoustic models [3, 18]. The wordlists were obtained from the existing phoneme-based pronunciation dictionaries. An alternative method would be to derive lists of words directly from transcriptions, but we did not follow that route as we wanted to guarantee the same size of vocabulary in both (phoneme and grapheme) dictionaries. Since the phoneme-based pronunciation dictionaries did not cater for any pronunciation variants, so did the graphemic dictionaries and hence the pre-processing phase involved sorting all grapheme-based dictionaries with unique words to remove redundancies. The details of the resulting grapheme-based dictionaries are outlined in Table 6.

TABLE 6: GRAPHEME-BASED DICTIONARY SETUP PER LANGUAGE

| | Language | | | |
|---|---|---|---|---|
| | SEP | ISI | VEN | MULTI |
| Unigraphs | 27 | 27 | 30 | 32 |
| Tigraphs | 4115 | 3963 | 3899 | 6261 |

As it is supported by Table 6, there are generally fewer graphemes than there are phonemes in a language. Also, there are more letters shared across the three languages than there are phonemes in Table 5. About 22 of the total 32 graphemes are shared across the languages. As reflected in Table 7, only 31% of the total graphemes are uniquely distributed across the languages.

TABLE 7. USAGE OF GRAPHEMES ACROSS LANGUAGES

| Graphemes | Language | | | |
|---|---|---|---|---|
| | SEP | ISI | VEN | ∑ |
| a b d e f g h i j k l m n o p r s t u w x y | x | x | x | 22 |
| q | | x | | 1 |
| š | x | | | 1 |
| c v z | | x | x | 3 |
| ḍ ḷ ṅ ṇ ṭ | | | x | 5 |

This is an encouraging coverage and it is what makes graphemes much easier and less costly to use as sub-word modeling units than phonemes in the selected languages. This uniform distribution of graphemes across languages is one of the motivating reasons to use context-dependent grapheme-based sub-word units for cross-lingual acoustic modeling. However, cross-language data sharing using graphemes will only be successful for closely-related languages. For loosely-related or unrelated languages, data sharing with graphemes will not work.

### D. Description of speech recognition systems

As features, we extracted Mel-frequency cepstral coefficients (MFCCs) and compute delta features. MFCC window size was set at 25ms with a frame rate of 10ms. The same training and testing corpuses were used in both phoneme-based and grapheme-based experiments. Both experiment types are trained with the same TARGETKINDS: MFCC with zeroth cepstral coefficients, delta and acceleration coefficients and the zero mean static coefficients.

All ASR systems follow the same architecture; the Hidden Markov Model Toolkit (HTK) [19] was used to develop context-dependent tied-state acoustic models based on HMMs. Both grapheme and phoneme sub-word units were modeled by a 3 state left-to-right HMM. Eight Gaussian mixture models are employed as state-conditioned output probability distributions.

## V. Results and Analysis

### A. ASR accuracy and language models

The developed ASR systems' performances are commonly analyzed and compared in terms of phone error rate (PER) or WER, as is in our study. The same flat language model containing all the words in the entire dataset is used to recognize all our test sets. Although statistical language models can be used to attain better recognition accuracies [4, 5, 6, 7], we are specifically evaluating the effect of the acoustic models without recognition being guided by a language model. Therefore our systems are evaluated and compared in terms of WER with the only difference between systems being their pronunciation dictionaries.

### B. *Phoneme vs. Grapheme systems*

Tables 8, 9, 10 and 11 reflect the recognition results obtained from SystemSep, SystemIsi, SystemVen and the cross-lingual system, respectively. The results obtained from both experiment types for all ASR systems are compared. All results are given at word-level accuracy as opposed to phone-level and grapheme-level accuracy. We use the WER to evaluate the performance of our systems, as reflected on the right column of the results tables. The last row (Difference) is used to measure the superiority of the one experimental approach on another.

After testing both experimental approaches per system, we obtained the following WERs: 54.32% WER on graphemes and 53.59% on phonemes for SystemSep, 59.44% on graphemes and 61.79% on phonemes for SystemVen, 65.06% on graphemes and 65.22% on phonemes for SystemIsi and 64.59% on graphemes and 62.22% on phonemes for the cross-lingual system. The phoneme-based WERs obtained in this study are comparable to those in [22]. The slight discrepancies can most likely be attributed to the kind of language models used.

TABLE 8. EXPERIMENTAL RESULTS FROM SYSTEMSEP

| System Type | WER |
|---|---|
| ExpGra | 54.32% |
| ExpPho | 53.59% |
| Difference | 0.73% |

TABLE 9. EXPERIMENTAL RESULTS FROM SYSTEMISI

| System Type | WER |
|---|---|
| ExpGra | 65.06% |
| ExpPho | 65.22% |
| Difference | -0.16% |

TABLE 10. EXPERIMENTAL RESULTS FROM SYSTEMVEN

| System Type | WER |
|---|---|
| ExpGra | 59.44% |
| ExpPho | 61.79% |
| Difference | -2.35% |

TABLE 11. EXPERIMENTAL RESULTS FROM THE CROSS-LINGUAL SYSTEM

| System Type | WER |
|---|---|
| ExpGra | 64.59% |
| ExpPho | 62.22% |
| Difference | 2.37% |

As shown in Tables 9 and 10, graphemes outperform phonemes. Tables 8 and 11 however, demonstrate the superiority of phonemes over graphemes. For Sepedi – SystemSep in Table 8, the phoneme-based system (ExpPho) is 0.73% more accurate than the grapheme-based system (ExpGra). The cross-lingual ASR system in Table 11 – also sees phoneme-based sub-word units performing better with a WER reduction of 2.37%. The ExpGra is 0.16% more accurate for SystemIsi and 1.23% in SystemVen. In [8], better results were achieved with graphemes, obtaining a 1.68% better word-level recognition accuracy as compared to our average 0.14% improvement. It is also a very interesting observation that the context-dependent grapheme-based sub-word units perform better than the phonemic ones in our study as opposed to the study in [7].

As it can be seen, the grapheme-based approach has the potential of attaining better results: the phoneme-based system is about 0.59% worse on average for the monolingual systems. However, the phoneme-based approach still performed better for the cross-lingual system despite the design and development benefits offered by the grapheme-based approach for the framework. As observed from the experimental results, a very strong correlation exists between the amount of training data and the WERs. As noted from the dictionaries, the number of phonemes per language is almost twice as large as the number of phonemes, this means that a single phoneme has smaller amount of acoustic training data than a single grapheme. Therefore, it is arguable that given enough training data and/or small phoneme set, the phoneme-based approach will work better. Another important factor on the recognition accuracy observed from the results is the grapheme-to-phoneme relation of these languages.

A very similar study in [20] reported degradation in word recognition accuracy using graphemes for the Afrikaans language. Although the grapheme-based system performed worse than the phoneme-based system, the results are still comparable and the authors successfully identified a set of "problematic categories" as the causes of the low performance of the ASR system. It is should also be noted that the indigenous South African languages are more phonetic than most of the European-based languages.

Although the WER reduction is minimal, these results demonstrate that graphemes are indeed capable of attaining recognition results comparable to, and/or even superior to phonemes and thus relatively acceptable in ASR experiments.

Given the minimal effort required to build pronunciation dictionaries for the grapheme-based system, as compared to the effort required for the phoneme-based system, we are confident that our suggestion to consider using graphemes will massively contribute towards successful development of large vocabulary continuous speech recognition systems for more under-resourced languages. These languages currently have few or non-existing speech processing tools. And also, the efficiency (in terms of cost and time) offered by graphemes demonstrates the possible preference of the suggested approach. The study significantly demonstrates that context-dependent grapheme-based sub-word units can be reliable for both monolingual and cross-lingual speech recognition tasks for the selected languages.

## VI. CONCLUDING REMARKS

In this study, we have shown that grapheme-based speech recognition, which copes with the problem of low-quality or unavailable pronunciation dictionaries, is comparable to phoneme-based recognition for Sepedi, IsiNdebele and

Tshivenda continuous speech in both the monolingual and cross-lingual speech recognition tasks. This straightforward approach is advantageous especially in situation of under-resourced languages and could be successfully used for building more robust speech recognizers for rare and marginalized languages. We demonstrated that graphemes can attain superior recognition accuracies for some under-resourced languages, preferably phonetic languages. This finding implies that for these under-resourced languages, graphemes can be considered for substituting phonemes as sub-word recognition units to lessen the total effort and cost required in developing perfectly hand-crafted pronunciation dictionaries.

To improve the quality of our recognition results, we intend to train our systems with more quality speech data, potentially from the NCHLT speech corpora. Having observed the ease of creating a pronunciation dictionary for a grapheme-based system, we hope to develop speech recognizers for other under-resourced languages of South Africa and also train cross-lingual recognition systems more rigorously.

Since this study essentially investigated the potential of grapheme-based sub-word units for monolingual and cross-lingual speech recognition. More work remains to be done to ensure satisfactory and reliable recognition results with significantly reduced recognition error rates so that the local speech processing research community can consider adopting this method to build LVCSR systems for more languages with little or no linguistic resources. This will benefit various speaker communities that use most of these heavily under-resourced languages on daily basis by ensuring the delivery of automatic linguistic tools which may significantly help with language preservation, uplifting and general language-specific e-service provisioning tasks.

## VII. Acknowledgements

The data sets were obtained from TCoE4ST and also Human Language Technologies (HLT) division of Meraka Institute at the Council of Scientific and Industrial Research (CSIR). Their support with the scarce spoken language processing resources is greatly acknowledged. The research project funding received from our sponsors, namely, Telkom and the National Research Foundation (NRF), is highly appreciated.

## VIII. References

[1] S. Stuker, and T. Schultz, "A Grapheme Based Speech Recognition System for Russian", SPECOM'2004: 9th Conference Speech and Computer, St. Petersburg, Russia, 20-22 September, 2004.

[2] S. Kanthak and H. Ney, "Multilingual Acoustic Modeling Using Graphems", in Proceedings of European Conference on Speech Communication and Technology, Geneva, Switzerland, vol. 2, pp. 1145-1148, September 2003.

[3] M. Killer, S. Stüker and T. Schultz, "Grapheme Based Speech Recognition", in Proceedings of the EUROSPEECH, pp. 3141-3144, December 2003.

[4] A. Black, K. Lenzo, and V. Pagel, "Issues in building general letter to sound rules", In Proceedings of the ESCA Workshop on Speech Synthesis, Australia, pp. 77-80, 1998.

[5] S. Besling, "Heuristical and statistical methods for Grapheme-to-Phoneme conversion", in Proceedings of Konvens, Wien, Austria, pp. 23-31, 1994.

[6] R. Singh, B. Raj, and R. M. Stern, "Automatic Generation of Subword Units for Speech Recognition Systems", IEEE Transactions on Speech and Audio Processing, vol. 10, pp. 98-99, 2002.

[7] S. Kanthak and H. Ney, "Context-dependent Acoustic Modeling using Graphemes for Large Vocabulary Speech Recognition", in Proceedings ICASSP, Orlando, Florida, pp. 845-848, 2002.

[8] E.G. Schukat-Talamazzini, H. Niemann, W. Eckert, T. Kuhn, and S. Rieck, "Automatic Speech Recognition without Phonemes", In Proc. European Conf. on Speech Communication and Technology, Berlin, pp. 129-132, 1993.

[9] A. Black and A. Font Llitjos, "Unit SelectionWithout a Phoneme Set", in Proceedings of the IEEE TTS Workshop, Santa Monica, CA, pp. 77-80, 2002.

[10] S. Krauwer, "The basic language resource kit (BLARK) as the first milestone for the language resources roadmap", In Proceedings of the International Workshop Speech and Computer SPECOM-2003, Moscow, Russia, pp. 8-15, 2003.

[11] E. Barnard, M. Davel, C. van Heerden, "ASR corpus design for resource-scarce languages", In Proceedings of Interspeech, pp. 2847–2850, 2009.

[12] N.J. De Vries, M.H. Davel, J. Badenhorst, W.D. Basson, F. de Wet, E. Barnard, and A. De Waal, "A smartphone-based ASR data collection tool for under-resourced languages", Speech Communication Journal – Special Issue on Processing Under-Resourced Languages, vol. 56, pp. 119-131, 2014.

[13] E. Barnard, M. Davel and G.B. van Huyssteen, "Speech technology for information access: a South African case study", In Proceedings of the AAAI Spring Symposium on Artificial Intelligence for Development (AI-D), Palo Alto, California, pp. 8–13, March 2010.

[14] STATISTICS South Africa, "Census 2011 – Census in brief", Online: http://www.statssa.gov.za/Census2011/Products.asp, 2012.

[15] S. Stüker, "Integrating Thai Grapheme Based Acoustic Models into the ML-MIX Framework - For Language Independent and Cross-Language ASR," in Proc. of the Spoken Languages Technologies for Under-resourced Languages (SLTU), Hanoi, Vietnam, pp. 27-32, 2008.

[16] S. Stüker, "Modified Polyphone Decision Tree Specialization for Porting Multilingual Grapheme Based ASR Systems to New Languages," in Proc. of ICASSP, pp. 4249–4252, 2008.

[17] T. Schultz and A. Waibel, "Language independent and language adaptive large vocabulary speech recognition", Speech Communication, vol. 35, no. 1, pp. 31-51, 2001.

[18] C.H. Schillo, G.A. Fink and F. Kummert, "Grapheme Based Speech Recognition for Large Vocabularies", In Proceedings of ICSLP '00, pp. 129-132, 2000.

[19] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Veltchev, and P. Woodland, "The HTK Book", http://htk.eng.cam.ac.uk/, Cambridge University Engineering Department, 2005.

[20] W.D. Basson and M.H. Davel, "Comparing grapheme-based and phoneme-based speech recognition for Afrikaans", In 23rd Annual Symposium of the Pattern Recognition Association of South Africa, PRASA 2012, CSIR International Convention Centre, Pretoria, pp. 144-148, November 2012.

[21] M. Davel and O. Martirosian, "Pronunciation dictionary development in resource-scarce environments", In Proceedings of Interspeech, pp. 2850-2854, 2009.

[22] D. Henselmans, T. Niesler and D. van Leeuwen, "Baseline Speech Recognition of South African Languages using Lwazi and AST", In 24th Annual Symposium of the Pattern Recognition Association of South Africa, PRASA 2013, Pretoria, pp. 30-35, December 2013.

**Mabu Manaileng** received his honors degree in Computer Science from the University of Limpopo in 2012. He is presently studying towards a Master of Science degree at the same institution. His research interests includes, among others, Automatic Speech Recognition and Synthesis, Statistical Pattern Recognition, Programming Languages, Network Security, Encryption Algorithms and High Performance Computing.

# Towards Developing a Stemmer for the IsiXhosa Language

Mnoneleli Nogwina, Zelalem Shibeshi and Zoliswa Mali
Telkom Centre of Excellence in ICT for Development
Department of Computer Science
University of Fort Hare, P/Bag X1314, Alice 5700, RSA
Tel: +27 40 6022464, Fax: +27 40 6022464
Email: {mnogwina, zshibeshi,zmali}@ufh.ac.za

**Abstract: IsiXhosa language is one of the eleven official languages in South Africa. It's the second most widely spoken language, however it is known to be one of those languages that have not much of automated systems. An efficient retrieval of documents can be archived if we use a technique called stemming. Basically, information retrieval (IR) systems make use of a stemmer to index documents using the stems of document keywords and also to change terms in users' queries to their stems so as to retrieve matching documents. In this paper we present our efforts in developing a stemmer for the isiXhosa language. The stemmer can be used in IR systems. By using 14 rules and a set of exceptions, we are able to predict the stems of an independent test set of nouns with 92 % accuracy. The paper also presents recommendations and future work.**

**Index Terms— Stemmer, Information Retrieval Systems, IsiXhosa Language, Xhosa, Stemming**

## I. INTRODUCTION

Natural languages are languages spoken by humans and currently we are not yet at that point where these can be understood by computers in their unprocessed forms. Natural language processing can be defined as the collection of techniques employed to try and accomplish this goal of computers understanding the spoken languages in their unprocessed forms [2]. Developing these programs is very difficult because natural languages have much ambiguity; words have several different meanings in different contexts [13]. One such problem is where we try to search for documents using key words while the documents we are looking for are represented by different words than the ones we used in our queries. As a result simple string matching does not help us to bring together documents and queries and we have to use other techniques to retrieve relevant documents. Stemming is one technique of natural language processing that we use to solve this type of a problem. In this project the problem of processing queries submitted in natural language to retrieve relevant documents from a collection of documents and the solution to this problem is discussed.

### *A. DATA ORGANISATION AND RETRIEVAL*

One of the most popular ways of organizing data is keeping it in a database system. In a database system data is stored in an organized way that makes retrieval easy because one can only search for data or information that they know it's in the database. The database management system uses a particular syntax for data retrieval and if one uses a wrong syntax then it will be difficult to get an output. Data is stored in tables and arranged using indexes. Fields that are going to be used for searching are indexed so that searching for a particular term or value will be easy. The tables in a database are related making it easy to go from one table to another because of the relationship between them. This is different from an online information retrieval system where documents are written in natural language with all the different variations of terms to explain a particular concept. The queries are also submitted using natural language and therefore finding a matching term in this kind of situation may be difficult. For this purpose terms representing documents and also queries needs to come to common form so that searching will be easier. The following section discusses how information storage and retrieval works in general.

### *B. INFORMATION STORAGE AND RETRIEVAL*

When one searches for a document online they submit a query by assuming a matching term can be found in the documents they are looking for. As a result, when the query is submitted, the corpus of the documents is checked for a matching term then those documents that have the term will be returned as output. There are different problems with this approach because as mentioned above words could exist in different forms and a word used in a query may not exist in a document that the user may be looking for. Stemming is employed to avoid such situations because with stemming different forms of words will be converted into one word, called the root (stem) and we use stems to index documents [11]. Similarly users' queries will also be converted so that they use stems of the terms that the user uses for his query. As a result, documents that have same stems are indexed together and also returned to the user together.

Consider for example, a document that contains information related to computer. We may have documents that contain words such as computers, computation, computer, and computational. So if we use one of these words in our query but a document does not contain that particular form of the word, the document may not be retrieved but if we use stemming to index documents then all the above words could be stemmed to their root which is "comput" and if a user asks for computers all documents that contain these words will be retrieved as they fall under one stem (comput). This is done by using a tool known as a

stemmer. As stemmers are language dependent, we also need a stemmer to handle the indexing of isiXhosa documents that will make information retrieval systems for isiXhosa effective.

## II. BACKGROUND AND RELATED WORKS

The algorithms for stemming have been studied in Computer Science since 1968 [15]. In many search engines there is a process called conflation that treat words with the same stem as synonyms and this helps in query broadening [15]. In this section we present some of the algorithms that have been studied on stemming in the recent past years. These stemming algorithms vary from language to language and the reason for variations is because stemmers are language dependent.

### *A. ENGLISH STEMMERS*

The Porter stemming algorithm was developed by Martin Porter at the University of Cambridge in 1980. It is also called conflation stemmer [12]. This stemmer has five steps and applies rules within each step of word transformation [9]. Porter used a minimal length that is based on the consonant-vowel-consonant strings, also known as the measure remaining after a removal of a suffix [16]. A typical rule may be as follows:
(m > 0) *FULNESS - > *FUL
This means the suffix FULNESS should be replaced by the suffix FUL if and only if the resulting stem has a non-zero measure (m) [16]. Non-zero measure is the measure that is greater than zero.
Another rule in Porter Stemming Algorithm;
(m > 0) *FUL - > *NULL, which means the suffix FUL should be replaced by a null string if, and only if the resulting stem has the non-zero measure [16].

Rules are applied one after the other, for example, this second rule is invoked after the first rule, and thus HOPEFULNESS will be stemmed first to HOPEFUL and then to HOPE in the second iteration [16]. As mentioned above there are five steps in this algorithm, the first handles inflectional suffixes and the second three handle derivational suffixes and then the last being the recoding step [9]. An inflectional suffix is a grammatical change of a word maybe from plural to singular or changing the tense without changing the meaning. A derivational suffix makes a grammatical change of a word in such a way that the new word has a new meaning "derived from the old word".

The other English stemming algorithm that we reviewed here is Lovins stemming algorithm. The authors in [12] mentioned that this was the first stemming algorithm that was ever published in 1968. Lovins stemmer has 294 endings, 29 conditions and 35 transformation rules. Each ending is associated with one of the conditions [6]. This algorithm searches for the longest ending which satisfies the associated condition first and it is removed. It is then in the second step where the 35 rules are applied to transform the endings. This step is performed whether or not the ending is removed from step one [6]. The rules from this algorithm handle conditions such as the letter undoubling, i.e.; (sitting - > sitt - > sit) but only for the English language [6].

### *B. STEMMERS FOR NON-ENGLISH LANGUGAES*

In developing a stemmer for the Greek language, Ntais [7] used a stemming algorithm similar to Porter's stemming algorithm. Their system takes a word as an input and removes its inflectional suffixes according to a rule-based English algorithm but developed according to the grammatical rules of the Greek language. When this system was evaluated it showed accuracy of about 92.1 percent of correct results and this proved its performance to be better than other past stemming algorithms that were developed for the Greek language [7].

The authors in [5] developed a stemming algorithm for the Dutch language and in this algorithm they also followed Porter's stemming algorithm as it is the most widely used algorithm. They chose the version that was published by [4], because of its advantage that clearly distinguishes between separation rules and procedures which test the attached conditions. Plural endings and verbal inflections do not affect the basic meaning of the stem and therefore can be removed without the risk of losing information [5]. They outlined that only those derivational affixes that do not affect the information conveyed by the stem should be removed. Basically they created a six rule cluster of the Dutch Porter stemming algorithm. Each cluster representing a particular class of affixes and the rules within a particular class are ordered, meaning the first rule that matches is applied and no other rules are tried in the same cluster.

On the other hand the authors in [1] presented their work for the stemming algorithm for the Silt'e language in Ethiopia. In their work the authors emphasize more on the use of stems for information retrieval systems. They highlighted the fact that Silt'e uses affixation and reduplication to derive different word forms from stems. The algorithm developed is an iterative stemmer that uses context sensitive and recoding rules to remove prefix, suffix and reduplication of letters [1]. The stripping procedure was applied in this order: prefix, suffix then letter reduplication.

## III. THE ISIXHOSA LANGUAGE

AmaXhosa, also known as the Southern or Cape Nguni are composed of numerous groups of people concentrated mainly in the former Transkei, Ciskei and Eastern Cape regions, who speak isiXhosa. As other African languages isiXhosa is a tonal language that is governed by the noun which dominates the sentence [8].

Approximately 8 million people in South Africa speak isiXhosa as their mother tongue. As a result, it is used as the main language of instruction in many primary schools and some secondary schools. It is also studied as a subject at these schools and some universities. IsiXhosa has five simple vowels a, e, i, o and u. The isiXhosa language is also known for its clicks. It has three basic click consonants; a dental click with the letter "c [kl]", a lateral click with the letter "x [kll]" and a palatal click with the letter "q [k!]" manifested in both Orthographical and Phonetic representation [8]. IsiXhosa is one of the branches of the

Nguni languages which include; SiSwati, IsiNdebele and IsiZulu. All these languages share many linguistic features. These Nguni languages also form part of a bigger group of Bantu languages. The isiXhosa language has been grouped into several dialects namely; isiBhaca, isiMpondo, isiMpondomise, isiGcaleka, isiNgqika, isiHlubi, isiXesibe, isiThembu, isiMfengu and isiBomvana. Xhosa nouns are classified into fifteen morphological classes, with different prefixes for plural and singular. It is an agglutinative language consisting of a list of prefixes and suffixes attached to root words or stems [8].

### *A.ISIXHOSA NOUN CLASSES*

The authors in [3] mentioned that, in linguistics the term noun class refers to a system of categorizing nouns. A noun may belong to a given class because of characteristic features of its referent [3]. In this regard, isiXhosa, as well as most Bantu languages, has a class system. In the isiXhosa language all nouns belong to one of the thirteen different classes. These classes are usually numbered from 1 to 15 with classes 12 and 13 and the ones higher than 15 missing. These classes are missing because in Proto-Bantu the 12 and 13 noun class contained diminutives, which in isiXhosa are signified by a suffix not a prefix and in some other Bantu languages that still do have noun classes 12 and 13.

#### *a) ISIXHOSA NOUN CLASS PREFIXES*

A prefix is a formative or morpheme that is placed at the beginning of a root or stem of the word [10]. The reason for looking at noun class prefixes is because the stemmer developed in this research is a noun stemmer. It only stems nouns and therefore the main focus will be on the prefixes and suffixes that are attached to noun roots to form meaningful nouns. Unlike European languages the IsiXhosa nouns are classified according to prefixes. IsiXhosa like many other African languages, especially those categorized as Bantu languages by philologists such Greenberg and Meinhoff, classifies nouns according to prefixes. IsiXhosa has about fifteen noun class prefixes also known as *Amahlelo* in isiXhosa [10]. IsiXhosa unlike other languages such as English where the removal of only suffixes yield correct stems, it has a rich set of prefixes that has to be removed in order to get proper stems. Tables 1 and 2 show the list of prefixes and suffixes that were used in the development of this stemmer.

**Table 1: List of prefixes used in this research.**

| | | | |
|---|---|---|---|
| 1. | Um | 7. | Isi, is |
| 2. | Aba, Abe, Ab | 8. | Izi, iz |
| 1(a). | U | 9. | In, Im, I |
| 2(a) | Oo | 10. | Izin, Izim, Ii, Iin, Iim |
| 4. | Imi, Im | 11. | Ulu, Ulw, Ul |
| 5. | Ili | 14. | Ubu, Ub, Utyw, Uty |
| 6. | Ama, Ame | 15 | Uku, Uk, Ukw |

### *B.ISIXHOSA SUFFIXES*

In order to get correct stems in isiXhosa both the prefixes and suffixes must be removed. Therefore it is important to also compile a list of suffixes that are commonly found in the nouns of this language.

**Table 2: List of suffixes used in this research.**

| |
|---|
| An-a |
| Kaz-i |
| El-a |
| Ek-a |
| Is-a |
| W-a |
| En-i |

When one is looking at the suffix table it is easy to notice that the final vowels are separated from the suffixes. In the isiXhosa language both nouns and verbs end in a vowel and therefore the vowel is given the name ”*isigqibelo*” (final vowel) as it is found at the end of a word. In this regard there are two commonly occurring suffixes; **kazi** and **ana.** These are nominal suffixes that are used to differentiate between feminine, augmentative and diminutive. **-Kazi** can be used to show feminine or augmentative and **-ana** to show augmentative. The following examples illustrate the use of these suffixes. *Nkosi* “chief” + *-Kazi* “feminine” = *Nkosikazi* “female chief”. *Isizwe* “nation” + *-Kazi* “augmentative” = *Isizwekazi* “big nation”. For Diminutive suffixes we show the following examples. *Inja* “dog” + *-Ana* “diminutive” = *Injana* “puppy”. *Idolophu* “town” + *-Ana* “diminutive” = *Idolophana* “small town”.

### *C. ISIXHOSA STEMS/ ROOTS/ NOMINAL/ VERBAL*

In this isiXhosa, whether it’s a noun or a verb, the meaning of the word is carried by a root or stem. To form a word one attaches a prefixes and a suffix to a root word. Therefore in developing this system we also took this argument into consideration and came up with a simple affix removal technique.

## IV. DESIGN AND IMPLEMENTATION

In order to create a stemmer for the isiXhosa language a list of prefixes and suffixes was created and therefore the algorithm removed the longest possible match in either of the lists. There were also exceptions to be considered such as when a word does not need to be stemmed. In the removal process the stemmer considered exceptions first then go to the prefix list and finally the suffix list. The proposed algorithm is presented in the following figure.

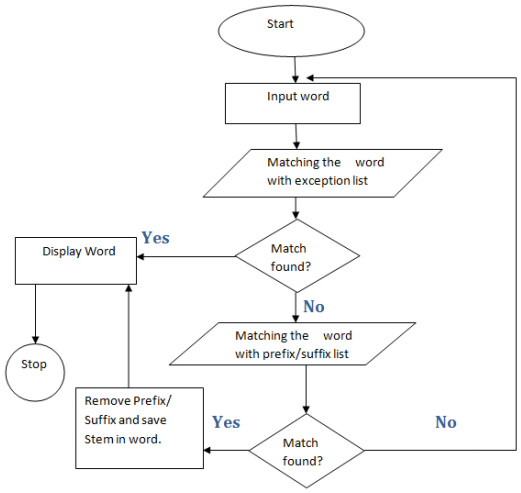



**Figure 1: IsiXhosa Stemming Algorithm.**

The designed algorithm clearly shows that a word that is found to be an exception is left untouched and it is returned as it is. If the input word is not found in the exception list, the algorithm then proceeds to the prefix list. It checks whether the word has a prefix or not, if prefix found it looks for the longest match then pass the remaining word to the suffix list. With the suffix list again a word is checked whether it has a suffix or not and if it contains a suffix the algorithm checks for the longest match then remove it. After both the prefixes and suffixes have been removed the remaining part of the word is then saved as a stem and presented to the user. If the word doesn't match either prefix or suffix list, it is also returned as it is. In order for this algorithm to accomplish all of these tasks, there are rules embedded in it. These rules specify how the word is read through the algorithm and how the removal techniques are performed. These are explained in the following section.

### *A. THE RULES*

The algorithm has about fourteen rules embedded in it. The rules seem to be many but because of the nature of the isiXhosa language, the algorithm cannot follow an iterative technique in its removal methods and therefore this is where this algorithm differs from the well-known Porter's stemming algorithm. Porter's algorithm has about five rules and it is iterative in its process. These rules were compiled from the literature; the researchers studied the language and also consulted with the linguist to come up with these sets of rules. The isiXhosa stemming algorithm that was developed in this research project is a light weight stemmer that simply removes affixes according to the list that is provided. One can say it is also a simple match algorithm because it takes the word and match it with the list of prefixes and if the match is found it is then removed and the same applies with the suffixes. In isiXhosa if one removes a suffix in most of the words they become roots and not stems. Therefore if one remains with a root it may be difficult to use this stemmer in the IR systems. This difficulty may be caused by the fact that in isiXhosa one root can represent many distinctive words and therefore for retrieval purposes it will return a lot of unrelated documents. But if we leave the suffixes untouched the stem remains and with the stem only related words will be represented and therefore making the retrieval technique precise. Also in this language stems end in a vowel making it easy to link related words and roots end with no vowel at all. There are cases where a root can work well in bringing together various words to one simple common representation.

The other important point one has to take into consideration is that for this algorithm we are only considering nouns. As mentioned before, there are also different forms that a noun can take from. These forms include Vocative, Locative, Predicative and Negative Predicative [14]. The Vocative is formed by dropping the augment "*iceba*". Locatives can be formed in many ways, but the most common one is to replace the augment by **e-** and replace the final vowel by **-ini**. Copulatives can be formed from both weak and strong classes. By weak classes we mean classes that their prefix contains a nasal, e.g. class 1,3,4,6 and 9 and strong classes are those that contain plosives in their prefixes [14]. It is formed in classes that starts with **u-** and **a-** by prefixing **ng-**, with exceptional case in class 9 where the prefix becomes **y-**. In these classes a copulative can be formed by repeating the plosive of the prefix. The copulatives have their own prefixes and are shown in the examples below.

(Class 6) Ngamakhwenkwe "they are boys" => Asingomakhwenkwe"they are no boys"

(Class 2) ngaBantwana/bangaBantwana "they are children" => Asingobantwana/Abangobantwana "they are no children"

## V. CHALLENGES FACING THE DEVELOPMENT

This kind of research requires the training of the algorithm using a huge collection of documents, also called the corpus. As per our investigation there is no standard corpus for isiXhosa language. In fact, isiXhosa is one of the languages which is resource scarce, meaning there are a limited number of resource of this language especially online. It is difficult to get valuable information on the language online. We believe this relate to the lack of tools like this.

## VI. RESULTS

For the purpose of this paper 200 nouns were used to test the stemmer. These nouns were read from a text file that was created using a list of random nouns chosen from an isiXhosa text book. From the noun formation section in the isiXhosa text book, we selected every second noun and the fourth noun for every form of the noun formation. We found that the system performed under stemming (prefixes or suffixes that were supposed to be removed but were not removed). The system also performed over stemming (prefixes or suffixes that were not supposed to be removed but were removed). Seven nouns were under stemmed and ten were over stemmed. The results show that from the 200 nouns that were on the test set only 17 were stemmed incorrectly, therefore this means the systems accuracy is 92 %. The test was repeated for a number of times to check consistency and accuracy of the results and same result were obtained.

## VII. DISCUSION

Through the difficulties that were encountered during the system development, we managed to finally design a system that produces proper stems for most of the nouns. The

results were appended in a log file where each noun is printed next to its corresponding stem, and then given to a linguist to determine which nouns were stemmed correctly and which ones were not. The results of the system are very encouraging and therefore proving the algorithm to be effective. 100% effectiveness could not be archived as there were exceptions to be considered and also the removal of the suffix **-ana** in most nouns resulted in over stemming, which requires further investigation.

## VIII. CONCLUSION AND FUTURE WORK

In this paper a stemmer for the isiXhosa language was developed and this stemmer will be used in IR systems to try and improve their performance when retrieving isiXhosa documents. The proposed algorithm is presented in a form of flow chat and the rules that are embedded in it are also explained. This is only a noun stemmer and therefore in the future it will be extended to stem verbs also. For now the stemmer focuses more on the prefixes because in this language plurals and singulars are determined by prefixation. Augments, feminine and diminutives are also taken care of as they are the most common suffixes found in the nouns. The result is very promising.

## IX. ACKNOWLEDGEMENTS

This work is based on the research undertaken within the Telkom CoE in ICTD supported in part by Telkom SA, Tellabs, Saab Grintek Technologies, Easttel, Khula Holdings, THRIP and National Research Foundation of South Africa (UID : 84006). The opinions, findings and conclusions or recommendations expressed here are those of the authors and none of the above sponsors accepts any liability whatsoever in this regard.

## X. REFERENCES

[1] Abedo, M. K. (2012). Designing a stemming algorithm for the Slit'e language. School of graduates studies department of information science (Master of Science dissertation, Addis Ababa University).

[2] Ahmad, S. (2007). "*Tutorial on Natural Language Processing"*. Artificial Intelligence, 810(161).

[3] Craig, C. G. (Ed.). (1986). *Noun classes and categorization: proceedings of a symposium on categorization and noun classification, Eugene, Oregon, October 1983* (Vol. 7). John Benjamins Publishing Company.

[4] Frakes, W. B., & Baeza-Yates, R. (1992). Information retrieval: data structures & algorithms. Prentice-Hall.

[5]Kraaij, W. & Pohlmann, R. (1995). "*Evaluation of a Dutch stemming algorithm*". The New Review of Document and Text Management. 1, 25-43.

[6]Lovins, J. B. (1968). "*Development of a stemming algorithm"*. MIT Information Processing Group, Electronic Systems Laboratory.

[7]Ntais, G. (2006). Development of a Stemmer for the Greek Language (Master of Science dissertation, Stockholm University).

[8] Pascoe, M. & Smouse, M. (2012). "*Masithethe: Speech and language development and difficulties in isiXhosa*". SAMJ: South African Medical Journal, 102(6), 469-471

[9] Porter, M.F. (1980). "*An algorithm for suffix stripping*", Program: electronic library and information systems, Vol. 14 Iss: 3, pp.130 – 137.

[10] Satyo, S. (1988). Igrama noncwadi lwesixhosa ibanga 10. 1st Edition. Via Afrika Limited.

[11] Sharma, D. (2012)."*Stemming Algorithms: A Comparative Study and their Analysis*." International Journal of Applied Information Systems, Foundation of Computer Science FCS 4.3 7-12.

[12] Smirnov, I. (2008)." *Overview of stemming algorithms*". Mechanical Translation.

[13] Tomita, M. (1985). "Efficient parsing for natural language: a fast algorithm for practical systems", Vol. 8. Kluwer Academic Pub.

[14] Tshabe, S.L., Putu, B.D., Mini, B.M., Mtuze, P. T., & Mkonto, N.V. (1988). IsiXhosa Sezikhuthali ibanga 10. 1st Edition. De Jager-HAUM.

[15] Ulmschneider, J. E. & Doszkocs, J. (1983). "*A practical stemming algorithm for online search assistance*". Online Information Review, 7(4), 301-318.

[16] Willett, P. (2006) "*The Porter stemming algorithm: then and now*", Program: electronic library and information systems, Vol. 40 Iss: 3, pp.219 – 223.

**Mnoneleli Nogwina** obtained his B.Sc. Honours (Computer Science) in 2012 from the University of Fort Hare and is presently studying towards his Master of Science degree at the same institution. His research interests include Natural Language Processing, Information Retrieval and End User Applications.

# Digital video shot boundary detector investigation

M.G. de Klerk, W.C. Venter and A.J. Hoffman
School of Electrical, Electronic and Computer Engineering
North-West University, Potchefstroom Campus, South Africa
Tel: +27 18 299 1978, Fax: +27 86 521 2569
Email: {20555466, 10063218, Alwyn.Hoffman}@nwu.ac.za

***Abstract*—Many algorithms have been proposed and evaluated for the detection of shot boundaries in digital video. Although these techniques have been verified, there remains a lack of standardised data to classify which techniques are best suited for certain applications. The Jensen-Shannon divergence (JSD) is one such technique used for shot boundary detection. In this article the JSD technique was adapted to handle monochromatic and RGB videos. This adaptation made it possible for the JSD technique to be evaluated in the RGB and monochromatic (grayscale) color spaces as well as the effect of video resizing in terms of recall, precision and execution time.**

***Index Terms*—shot boundary detector (SBD); Jensen-Shannon Divergence**

## I. Introduction

Digital video has become part of our everyday lives. It is not just found in television broadcasts any more but it is commonplace, everywhere from our homes to our workplace, and even on the go with the modern mobile devices. This advent of digital communication by means of video has brought forth a new set of challenges.

Managing video content is a laborious task that has grown beyond what is manually practical. Various techniques have been developed that can perform this task, but tend do be computationally intensive and slow to execute [1]. These techniques are discussed in literature, but a lack of standardised testing makes a quantitative comparison from literature impossible.

One of the core techniques utilized to manage video content is a shot boundary detector (SBD). The aim of this investigation is to quantitatively compare one such a SBDs by subjecting it to a multitude of controlled tests.

This article provides an overview of a SBD and previous work done in this regard in section II-B. This is followed by a breakdown of the methodology followed while evaluating the JSD SBD in section III. Lastly the experimental results of the investigation and a conclusion thereof is provided in sections IV and V.

## II. Background

Video segmentation refers to the process of breaking down a video sequence into the shots from which it is comprised [2]. This can be achieved by manually identifying shot boundaries or by implementation of a SBD. A SBD is an algorithm used to identify the shot boundary between consecutive shots in a video stream. This boundary is generally a considerable discontinuity in the visual-content flow of a video sequence [3].

In order to grasp the application of a SBD, it is important to understand the basic structure of digital video. The basic structure of digital video is inherently similar to that of traditional film video. The smallest component of any video is called a frame, which is a static image. A collection of multiple consecutive frames, captured by a camera in a single continuous take, is called a shot [4]. When multiple shots are combined into a single video, it is referred to as a sequence, or commonly referred to as a video as seen in figure 1.

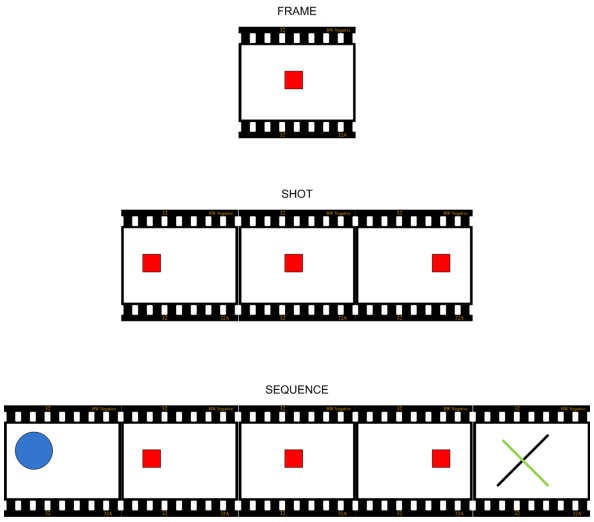



Fig. 1. Hierarchical structure of a video sequence

### A. SBD overview

Shot boundaries, as illustrated in figure 2, can be classified into two distinct classes: hard and soft cuts. A hard-cut refers to an abrupt change from one shot to another [5]. Soft transitions refers to a collection of transition methods used to soften the transition boundary between frames. This is commonly done by dissolving one frame into (fade-in) or from (fade-out) a black image [6]. Another technique involves fading one video section into another, called cross-fading. Soft-cuts are generally more difficult to detect than hard-cuts and can be easily falsely detected due to a drastic illumination difference e.g. a lens flare.

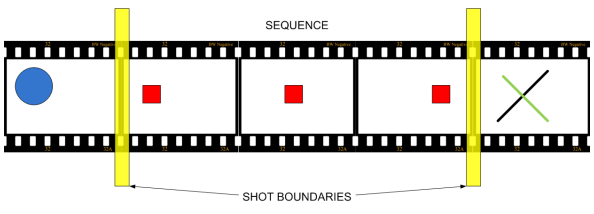



Fig. 2. Video structure hierarchy

A SBD algorithm identifies the start and end of a new shot. Once the start and end positions of a shot are known, it can be further analysed in manners similar to the SBD to determine the most representative frame of that shot, called the key-frame, for video indexing purposes.

### B. Related work

A lot of work has been done in the field of SBD algorithms. The majority of SBD algorithms are designed to function on color videos. Each pixel in these color frames are represented by a red, green and blue (RGB) value. This provides techniques like the Jensen-Shannon Divergence (JSD) algorithm with 3 separate *datasets* i.e. histograms of the red (R), green (G) and blue (B) components of each pixel. This abundance of information makes it easier to detect soft transitions between shots, but it comes at a price. An RGB video has 3 times more data to process than a grayscale video, thus requiring more time to process all the data.

The aim of this investigation is to determine how these various techniques can be adapted to function on grayscale data to increase processing speed but retaining the accuracy of the algorithm.

### C. SBD techniques

There are various techniques available to detect the shot boundaries in videos. The majority of these techniques can be grouped under pixel based or histogram based techniques.

#### *Pixel-based methods*

The simplest way of comparing two consecutive frames is by comparing the corresponding pixels in both frames and measuring the difference. A shot boundary may have been detected if the number of changed pixels exceeds a prescribed threshold. This technique is sensitive to camera motion [5] and noise since each pixel is evaluated as a single entity.

Statistical methods like the mean and standard deviation of the pixels' gray levels can be used to provide a measure of variance between frames. This method proved to be more tolerant of noise, but takes longer to perform calculations due to the complexity of the statistical formulas.

#### *Histogram-based methods*

Histograms SBD techniques applies the same basic principles as pixel-based techniques, but now the pixels are grouped into bins.

The most commonly used metric for shot boundary detection is the histogram difference between frames. Histograms are drawn of each frame by counting pixels with certain values based upon the number of bins chosen - typically 256 bins to quantize the image. Grayscale images deliver a single histogram per frame where color images require 3 histograms.

By drawing a histogram of each frame and subtracting them, one is left with a difference histogram indicating the frame-to-frame content change. Frame pairs with a high content change are usually indicative of a shot boundary. A linear combination or straight forward summations of these differences are then used as a measure of the total content change across the frames [7]. A predefined value for these metrics, called a threshold, is generally used to discern if this difference constitutes a shot boundary. These thresholds can be set manually or calculated automatically.

These histograms can then be compared using various techniques like the JSD, Chi-squared ($\chi^2$) comparison [8] and various other techniques.

This article investigates the JSD technique as it was adapted for both RGB (3 histograms) and monochromatic (1 histogram) videos. Since the JSD technique is a histogram based method, it is less susceptible to noise and camera movement, with the advantage of faster processing speeds which makes it ideal for applications such as advertisement tracking.

#### *JSD algorithm*

The JSD algorithm provides a means to determine the frame difference measure (FDM) between consecutive frames, which in turn can be used to detect if these frames constitute a shot boundary. As the name suggests, the JSD algorithm is a combination of the Shannon's entropy and the Jensen inequality.

The Shannon entropy [9] is a method used in information theory as a measure of information choice and uncertainty. The Shannon entropy function $H$ of the probability distribution $P = (p_1, p_2, p_3, \ldots, p_n)$ consisting of $n$ possibilities in the distribution is calculated by:

$$H(P) = -K \sum_{i=1}^{N} p_i \log_b p_i \tag{1}$$

where $K$ is a positive constant [10].

This entropy calculation is combined with the Jensen inequality measure between two consecutive frames' histograms as derived by Qing Xu [4] producing the JSD equation:

$$JSD(f_{i-1}, f_i) = H\left(\frac{P_{f_{i-1}} + P_{f_i}}{2}\right) - \frac{H\left(P_{f_{i-1}}\right) + H\left(P_{f_i}\right)}{2} \tag{2}$$

where $f_{i-1}$ is the previous frame and $f_i$ the current frame. The probabilities of the frames are respectively $P_{f_{i-1}}$ and $P_{f_i}$. The aforementioned probabilities are calculated from the applicable frames' histograms as expressed in equation 3:

$$P(f_i) = \frac{Histogram\left(f_i\right)}{Height\left(f_i\right) \times Width\left(f_i\right)} \tag{3}$$

where the histogram of each color component is divided by the number of pixels in that frame $f_i$. The number of pixels is calculated by multiplying the number of horizontal pixels by the number of vertical pixels in the frame.

The results from equation 2 are then used to create a moving average of the previous frames that serves as a threshold level to test the current frame-to-frame difference against. Trial and error has indicated that a *moving average window* of 5 frames produced satisfactory results for an *auto-adjusting threshold.*

## III. Methodology

A high level abstraction of the basic methodology followed during the SBD analysis can be seen in figure 3.

MATLAB was used as the primary programming environment due to the availability of mathematical and statistical tools as well as video functionality.

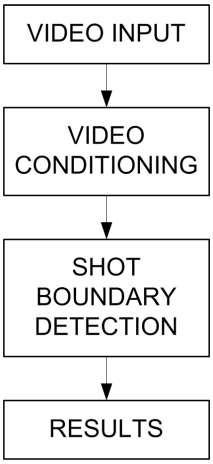



Fig. 3. SBD high level program flow

- *Video input*
  Each video is read into MATLAB's workspace as a video object with the `VideoReader` function, containing all the frames and relevant video information. Due to constraints of the aforementioned function, only the following video formats are supported:
  - All Platforms:
    - Motion Joint Photographic Experts Group (JPEG) 2000 (.mj2);
    - Audio Video Interleave (.avi);
  - All Windows® platforms:
    - Moving Picture Experts Group Phase 1 (MPEG1) (.mpg);
    - Windows Media® Video formats:
      - Advanced Systems Format (.asf);
      - Windows Media Video (.wmv);
      - Advanced Stream Redirector (.asx);
  - Windows 7 or later:
    - Apple Quicktime Movie (.mov);
    - Moving Picture Experts Group Phase 4 (MPEG4 including H.264 encoded video) (.mp4);
  - Moving Picture Experts Group Phase 1 (.mpg);
- *Video conditioning*
  Due to the input format constraints imposed by using the `VideoReader` MATLAB function, all the videos used in the comparison had to be converted to a suitable format as listed above. Many digital videos on the internet are packaged in a Adobe Flash video format (.flv) [11] which is not natively supported by MATLAB. These videos were converted to the AVI container format, using their original codecs in *Any Video Converter Ultimate 4.4.2* which utilizes the FFmpeg. FFmepeg provides a complete, cross-platform solution to record, convert and stream audio and video [12].
  The digital videos were subjected to some conversions during the analysis to determine the effect thereof on the accuracy and execution speed of the SBD. The RGB-to-Grayscale conversions were done in MATLAB on the video loaded into memory using `rgb2gray`, leaving the original video in its original state for further calculations. Resizing of the video to a smaller resolution of 320x240 was done beforehand by using *Any Video Converter Ultimate 4.4.2*.
  An arbitrary resolution of 320x240 was chosen for the resize conversion, since it fit all the source video's aspect ratio of 4:3.
- *Shot boundary detection*
  The aforementioned JSD algorithm was implemented on the video data that has been loaded into MATLAB's workspace.
- *Results*
  The frame index of each boundary detected by the JSD algorithm was stored as well as the processing time of the SBD algorithm. The frame indexes were used to visually confirm the frame identified as a shot boundary by displaying it alongside the previous and following frame as seen in figure 4. All the videos were also analysed by hand to identify all the shot boundaries. This manual process is called ground-truthing [1]. Based upon the results obtained from the comparison, it was possible to calculate the precision and recall factors as defined in section IV.


Fig. 4. SBD visual comparison example

The aforementioned execution time of the SBD algorithm does not include the *overhead* of the simulation that includes the loading and conditioning (wrapper conversion) of the video. This is attributed to the fact the current video analysis approach implemented in MATLAB was focused on comparing the speed, recall and precision of the SBD technique on the various videos.

## IV. Results

The Performance of the JSD algorithm, based upon recall and precision rates as expressed in equations 4 and 5 [13], was determined when the algorithm was applied to different test videos.

Recall rates provides a ratio of the number of relevant shot boundaries (correct detection) to the total number of relevant shot boundaries in the video, while precision rates provides the ratio of relevant shot boundaries to the total number of irrelevant shot boundaries (false detections) retrieved.

$$Recall = \frac{Correct}{Correct + Missed} \tag{4}$$

$$Precision = \frac{Correct}{Correct + FalsePositive} \tag{5}$$

In order to simplify the ground-truthing process and ensure accuracy, multiple test sequences were created to test the functionality of the technique. A linear depiction of these test sequences is illustrated in figure 5.

The sequences were constructed to have the following properties:

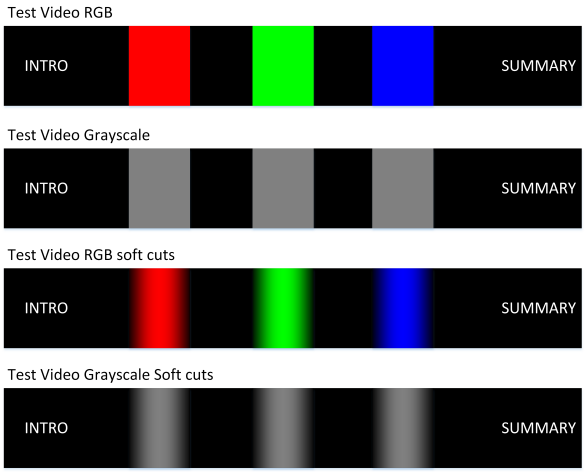



Fig. 5. Test video sequences

- *Test Video RGB:*
  Black video sequence to R, G and B sequence transitions featuring hard cuts;
- *Test Video Grayscale:*
  Black video sequence to gray sequence hard-cut transitions;
- *Test Video RGB soft cuts:*
  Soft transitions from black to R, G and B and back to black again. Cross-dissolve transition duration is 25 frames;
- *Test Video Grayscale soft cuts:*
  Black video sequence to gray sequence and back to black cross-dissolves.

The hypothesis behind the full color versus monochromatic representation was to test how the accuracy of the technique holds up with the loss of data as well as the effect thereof on the execution speed. It is important to note that the gray sections in the aforementioned test videos were created by a RGB value of R=128; G=128; B=128 on the RGB 8bit color map. This RGB-grayscale representation allowed the JSD technique be compared on RGB-gray (3 separate histograms) and monochromatic-gray (1 histogram) videos. All video sequences were converted to monochromatic representations where applicable in the various algorithms by the principle explained in equation 6:

$$monochromatic = \frac{R+G+B}{3} \tag{6}$$

where R,G and B are the individual component values of each pixel.

### A. Test video results

The execution times listed was obtained by averaging the execution times of the JSD SBD technique for 10 iterations. The results of the RGB test video shot boundary detection analysis are listed in table I, confirming the previous statement that an RGB comparison of the 3 histograms takes longer to execute compared to a monochromatic comparison with 1 histogram. The *JSD Grayscale* technique listed includes the time required for the RGB-to-Grayscale conversion with the `rgb2gray` function. The weak recall rate produced by the RGB techniques can be attributed to the thresholding criteria.

The thresholding for RGB videos is essentially the same as for the monochromatic videos. The only difference is that it is implemented on all three components. The threshold concept for monochromatic videos to detect a shot boundary (SB) is given by equation 7:

$$SB = \begin{cases} True & \text{if } JSD(f_{i-1}, f_i) > T_{aveGray} \\ False & \text{otherwise} \end{cases} \tag{7}$$

where $T_{aveGray}$ is the average moving threshold computed from the *moving average window*. Similarly for RGB videos, each color channel represented by equations 8, 9 and 10 must be evaluated:

$$SB_R = \begin{cases} True & \text{if } JSD_R(f_{i-1}, f_i) > T_{aveR} \\ False & \text{otherwise} \end{cases} \tag{8}$$

$$SB_G = \begin{cases} True & \text{if } JSD_G(f_{i-1}, f_i) > T_{aveG} \\ False & \text{otherwise} \end{cases} \tag{9}$$

$$SB_B = \begin{cases} True & \text{if } JSD_B(f_{i-1}, f_i) > T_{aveB} \\ False & \text{otherwise} \end{cases} \tag{10}$$

in order to calculate if a shot boundary has been found. A shot boundary has been identified if equation 11 is satisfied:

$$SB_{RGB} = \begin{cases} True & \text{if } SB_R \,\&\, SB_G \,\&\, SB_B = True \\ False & \text{otherwise.} \end{cases} \tag{11}$$

TABLE I
TEST VIDEO - RGB

| SBD Technique | Execution Time (s) | Precision (%) | Recall (%) |
|---|---|---|---|
| JSD Grayscale | 3.83 | 100 | 87.5 |
| JSD RGB | 5.08 | 100 | 14.29 |
| JSD Grayscale Resized | 1.11 | 100 | 87.5 |
| JSD RGB Resized | 1.85 | 100 | 14.29 |

The native grayscale test video produced better recall rates as can be seen in table II. The only difference between the RGB and grayscale test videos are the colors. The RGB video has definitive red, green and blue scenes, while the grayscale video has only gray scenes amidst the black timeline. These gray scenes contain equal values for all the RGB channels (R=G=B=128). Hence the boundary detection conditions as required by equation 11 is easily satisfied for if one channel is satisfied, the remaining two will be satisfied as well.

TABLE II
TEST VIDEO - GRAYSCALE

| SBD Technique | Execution Time (s) | Precision (%) | Recall (%) |
|---|---|---|---|
| JSD Grayscale | 3.85 | 100 | 87.5 |
| JSD RGB | 5.17 | 100 | 87.5 |
| JSD Grayscale Resized | 1.09 | 100 | 100 |
| JSD RGB Resized | 1.78 | 100 | 14.29 |

*Soft cut test video results*

The nature of a gradual transition makes it difficult to distinctively classify a certain frame as the shot boundary. The recall and precision rates listed in tables III and IV were calculated by determining if the detected boundary was within the frame-range constituting the gradual transition.

TABLE III
TEST VIDEO - RGB SOFT CUTS

| SBD Technique | Execution Time (s) | Precision (%) | Recall (%) |
|---|---|---|---|
| JSD Grayscale | 3.80 | 70 | 87.5 |
| JSD RGB | 4.93 | 100 | 87.5 |
| JSD Grayscale Resized | 1.10 | 77.78 | 87.5 |
| JSD RGB Resized | 1.81 | 50 | 91.3 |

The lower precision values are attributed to the false detections during the gradual transition of both the monochromatic and RGB videos. Only one shot boundary was allowed within the span of the gradual transition as the definition of a shot boundary implies that each shot has a start and stop boundary. A gradual transition from black to white is illustrated in figure 6 over a period of 10 frames. Adhering to the aforementioned guidelines regarding the gradual transition, only one frame should be identified as the shot boundary e.g. frame nr. 5, even though frame 6 might also provide a difference value higher than the current threshold.


Fig. 6. Gradual Transition

TABLE IV
TEST VIDEO - GRAYSCALE SOFT CUTS

| SBD Technique | Execution Time (s) | Precision (%) | Recall (%) |
|---|---|---|---|
| JSD Grayscale | 3.77 | 53.85 | 87.5 |
| JSD RGB | 4.83 | 53.85 | 87.5 |
| JSD Grayscale Resized | 1.10 | 80 | 100 |
| JSD RGB Resized | 1.81 | 80 | 12.5 |

Hence the low precision is caused by the identification of a second shot boundary alongside the current shot boundary within the same transition period.

All the test videos have an introduction title and information summary at the end of the sequence. It is debatable whether or not these text sequences can be classified as a shot or not. Although the majority of the techniques failed to identify the addition or removal of these text sequences, with the exception of both the resized implementations of the JSD algorithm on the grayscale test sequences.

### B. Other video results

The same analysis was done on a video generally available on the internet in order to determine how they fared compared to the best case scenarios of the test videos. The action-sports themed RedBull commercial contains multiple lens-flares and flash photography that in turn gave rise to a few false detections. The results listed in table V indicate worse recall and precision rates for the original video than compared to the smaller resized versions thereof.

Upon further investigation into this peculiarity, it came to light that the frame rate of the video has an effect on the accuracy of the techniques being evaluated. The original RedBull Commercial was a 1280x720 video file with a frame rate of 30 frames per second (fps). During the resize process, the video was down-sampled to 25 fps, the same as all the test videos.

A higher video frame rate constitutes more frames during a gradual transition, in effect causing less change from frame-to-frame. Thus in order for the JSD technique to function correctly at a higher frame rate, the size of the moving average window used for thresholding should be increased while the threshold might be lowered to detect the *slower changes* in a frame-to-frame context.

TABLE V
THE WORLD OF REDBULL TV COMMERCIAL 2013

| SBD Technique | Execution Time (s) | Precision (%) | Recall (%) |
|---|---|---|---|
| JSD Grayscale | 7.31 | 72.73 | 92.31 |
| JSD RGB | 10.49 | 74.19 | 88.46 |
| JSD Grayscale Resized | 1.87 | 76.67 | 100 |
| JSD RGB Resized | 2.94 | 87.5 | 100 |

The final video used in the comparison was a wildlife video, depicting various scenes with slow and fast moving animals, hard cuts between shots and no lens-flares. The results, as listed in table VI, shows a perfect score for both recall and precision across all the SBD techniques.

TABLE VI
WILDLIFE

| SBD Technique | Execution Time (s) | Precision (%) | Recall (%) |
|---|---|---|---|
| JSD Grayscale | 3.11 | 100 | 100 |
| JSD RGB | 4.29 | 100 | 100 |
| JSD Grayscale Resized | 0.88 | 100 | 100 |
| JSD RGB Resized | 1.40 | 100 | 100 |

## V. CONCLUSION

The SBD technique comparison confirmed the logical assumption that RGB comparisons take longer to execute than a monochromatic comparison. Overall the RGB techniques had higher precision values than the grayscale techniques, but their recall rates plummeted in certain instances.

Amongst other factors, it is clear that lens-flares and flash photography can adversely affect the outcome of these detection algorithms by providing false positives. Heng et al. [14] proposed various techniques to remove false alarms caused by abrupt illumination changes which can be implemented.

It can be concluded that for most instances of this comparison, the JSD grayscale technique on resized videos at 25fps produced the best results in the shortest amount of time.

There are a few impact factors on SBD that require further investigation:

- Effect of the frame rate;
- Effect of various file wrappers and codecs;
- Text over video.

The general convention for video composition focusses the subject in the middle of the frame, hence the most detail and movement tends to be focussed here. This convention calls for further investigation into a special attention areas as described by Gu et al. in [15].

## VI. Acknowledgement

I would like to thank my supervisor, Prof. W.C. Venter for his help and guidance throughout the duration of the project.

## References

[1] U. Gargi, R. Kasturi, and S. H. Strayer, "Performance characterization of video-shot-change detection methods," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 10, no. 1, pp. 1–13, 2000.

[2] I. Koprinska and S. Carrato, "Temporal video segmentation: A survey," *Signal processing: Image communication*, vol. 16, no. 5, pp. 477–500, 2001.

[3] A. Hanjalic, "Shot-boundary detection: unraveled and resolved?" *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 12, no. 2, pp. 90–105, 2002.

[4] M. Vila, A. Bardera, Q. Xu, M. Feixas, and M. Sbert, "Tsallis entropy-based information measures for shot boundary detection and keyframe selection," *Signal, Image and Video Processing*, vol. 7, no. 3, pp. 507–520, 2013.

[5] J. S. Boreczky and L. A. Rowe, "Comparison of video shot boundary detection techniques," *Journal of Electronic Imaging*, vol. 5, no. 2, pp. 122–128, 1996.

[6] W. J. H. N. Ngan, "Shot boundary refinement for long transition in digital video sequence," *IEEE TRANSACTIONS ON MULTIMEDIA*, vol. 4, no. 4, pp. 434–445, 2002.

[7] M. R. Naphade, R. Mehrotra, A. M. Ferman, J. Warnick, T. S. Huang, and A. M. Tekalp, "A high-performance shot boundary detection algorithm using multiple cues," in *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on*, vol. 1. IEEE, 1998, pp. 884–887.

[8] I. Williams, N. Bowring, and D. Svoboda, "A performance evaluation of statistical tests for edge detection in textured images," *Computer Vision and Image Understanding*, vol. 122, pp. 115–130, 2014.

[9] C. E. Shannon, "A mathematical theory of communication," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 5, no. 1, pp. 3–55, jan 2001. [Online]. Available: http://doi.acm.org/10.1145/584091.584093

[10] M. Menéndez, J. Pardo, L. Pardo, and M. Pardo, "The jensen-shannon divergence," *Journal of the Franklin Institute*, vol. 334, no. 2, pp. 307–318, 1997.

[11] J. Emigh, "New flash player rises in the web-video market," *Computer*, vol. 39, no. 2, pp. 14–16, 2006.

[12] Ffmpeg. [Online]. Available: http://www.ffmpeg.org/

[13] V. Raghavan, P. Bollmann, and G. S. Jung, "A critical investigation of recall and precision as measures of retrieval system performance," *ACM Transactions on Information Systems (TOIS)*, vol. 7, no. 3, pp. 205–229, 1989.

[14] W. J. Heng and K. N. Ngan, "High accuracy flashlight scene determination for shot boundary detection," *Signal Processing: Image Communication*, vol. 18, no. 3, pp. 203–219, 2003.

[15] X. Gu, Z. Chen, and Q. Chen, "Refinement of extracted visual attention areas in video sequences," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 966–969.

**M.G. de Klerk** is currently studying towards an PhD degree in Electrical and Electronic Engineering at the North-West University after receiving his M.Eng. degree on renewable energy simulations in 2013.

# A Model for Context Awareness for Mobile Applications using Multiple-Input Sources

Direshin Pather, Janet Wesson and Lester Cowley
Department of Computing Sciences
Nelson Mandela Metropolitan University, P. O. Box 77000, Port Elizabeth, 6031
Tel: +27 41 504 2323, Fax: +27 41 504 2831
Email: {Direshin.Pather, Janet.Wesson, Lester.Cowley}@nmmu.ac.za

**Abstract - Context-aware computing enables mobile applications to discover and benefit from valuable context information, such as user location, time of day and current activity. However, determining the users' context throughout their daily activities is one of the main challenges of context-aware computing. With the increasing number of built-in mobile sensors and other input sources, existing context models do not effectively handle context information related to personal user context. This paper aims to investigate how personal user context can be inferred from multiple input sources. A model is proposed using multiple inputs obtained from mobile sensors and other sources to support context-awareness for mobile applications. A prototype will be built using the proposed model and evaluated to determine if the proposed model can effectively support context-awareness in mobile applications.**

**Index Terms — Context-awareness, personal user context, ubiquitous computing, multiple input sources, sensors.**

## I. INTRODUCTION

Context awareness is essential in mobile computing. When users are on the move, their context of use changes faster than when they are still and tied to a desktop device. For example, as people move, their location changes, the devices and people they interact with change more often and their goals and information needs change. Mobility provides opportunities for leveraging context awareness but requires robust gathering and processing of context to try and understand how the user's goals are changing. This places extra burdens on the mobile computing platform, as it needs to sense potentially rapidly changing context, synthesize it and act upon it [1].

Dey & Abowd (2001) define *context* as any information that can be used to characterize the situation of an entity. Real-time access to context information can support time-critical applications, such as emergency healthcare, and location-based services that exploit knowledge about where a user is located [2]. This is due to the extensive adoption of mobile phones with context-awareness features enabled through sensors and personal information, such as calendars. Another factor contributing towards utilizing context-awareness is the tendency of individuals to carry their phones with them everywhere [3]. However, the system infrastructure to support context awareness is not widely available to developers, resulting in less context-aware applications from which users can benefit.

Context-awareness is a key requirement for achieving ubiquitous computing. Ubiquitous computing enables information technology to be invisible while still being integrated within our daily lives [4]. Improved human-computer interaction can be achieved by facilitating context-awareness in mobile devices such as smart phones [5].

Location awareness is an important aspect of context-awareness that is needed to characterize the situation of an entity. Location awareness enables services to access information relevant to the current situation, such as a patient's location. Patient location is important contextual information that is required in healthcare systems and especially in remote health monitoring [2]. Location becomes a crucial attribute for patients who suffer from memory loss diseases such as Alzheimer's disease. Having access to the patient's location can help to provide timely medical assistance in emergency and life-threatening situations [6].

Mobile phones can be used to capture contextual information to help individuals to better understand the situations that affect their daily lives. Location, which is part of the environmental context of an entity, can be acquired by accessing the mobile phone's sensors, such as the accelerometer and the gyroscope [5]. Other sources of contextual information such as the schedule from a user's calendar, can also be obtained. A suitable model and structure to support context information could be used to help determine the user's context. Determining the users' context throughout their daily activities is one of the main challenges of context-aware computing [7]. Using multiple input sources with a robust context-aware model to identify and predict context, could help to solve this problem [8].

The aim of this paper is to investigate how personal user context can be inferred from multiple input sources. The personal user context is determined in terms of four dimensions: physical context, user activity, health and preferences. These dimensions form the basis of the proposed model for context awareness. Only those sensors commonly found in mobile phones are considered. These sensors include the accelerometer, gyroscope, light sensor and temperature sensor. Other input sources include the GPS receiver, web services and user profile information.

This paper is structured as follows: Section II provides an overview of related work on context-awareness and mobile solutions. Section III describes the analysis and design of the proposed model. Section IV describes the preliminary implementation of a prototype. Section V discusses future work and concludes this paper.

## II. RELATED WORK

### A. Context-Awareness

Context-awareness in computing is the existence of computer systems and applications, which can collect and understand "*information about the immediate situation such as the people, roles, activities, times, places, devices, and software that define the situation*" [9]. Based on this perceived context, the systems and applications should perform appropriate and related actions. These actions can involve the presentation of customized or specially formatted information, the performance of some action to avoid a potentially dangerous situation, or assistance in the case of an emergency [9]. As a result, context needs to be managed and interpreted correctly, as it may be acquired from multiple and heterogeneous sources.

Context awareness typically involves a number of steps [10]:

- Acquisition of contextual information
- Monitoring contextual information
- Filtering contextual information
- Storing of contextual information
- Representation of contextual information
- Interpretation of contextual information.

In order for context-awareness to be effectively used to support context-aware applications, an underlying architecture needs to be in place.

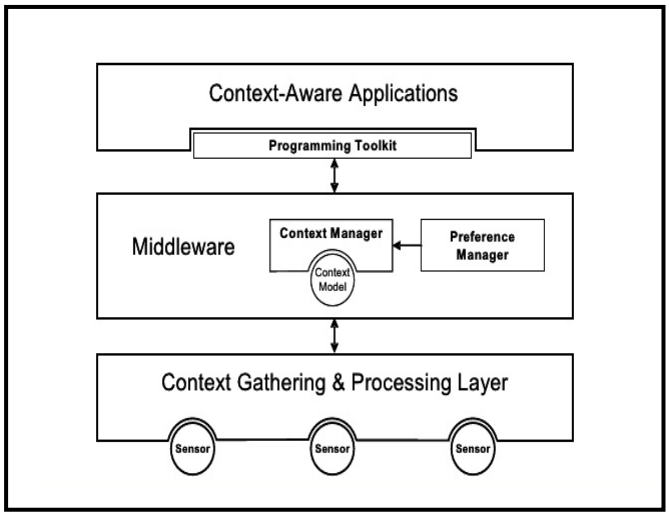



Figure 1: Typical Architecture to Support Context-Aware Applications [11]

The typical architecture is illustrated in Figure 1 and includes the context gathering and processing layer, middleware layer and context-aware application layer. The context gathering and processing layer, also known as the context provider, involves the acquiring, monitoring and filtering of contextual information. The middleware layer involves the storing, representation and interpreting of contextual information. The context-aware application layer, also known as the consumer layer, is the layer in which context becomes an input for applications to use, in order to appropriately change these applications' behaviour.

A number of existing issues in context-awareness were identified from literature. Incorporating device, user and environmental information using advanced sensors and sensor networks is still a major challenge [12]. Some sensor-based systems merely display the sensor information without actually expressing or defining what it means in terms of high-level context [13]. For example, displaying context as separate variables, such as the current time and the current GPS coordinates. Furthermore, these context variables are not discussed in terms of the integration of the different sensor information or what they mean at a higher level.

Another challenge is determining high-level context information such as the user's current activity [14]. Due to sensor uncertainty and unreliability, the sensor data can become invalid or missing. The goal of a context-aware system is to retain as much location-awareness capability as possible in such circumstances.

Some possible solutions are evident in existing literature. One approach to dealing with high-level context information, such as the user's current activity, is machine vision, which is focused on image processing and camera technology [14]. Another potential approach is to check the user's calendar in order to identify what the user is meant to be doing at a specific time. Another promising approach involves using artificial intelligence techniques to recognize context by integrating several low-level sensors [15]. These techniques can include rules or machine learning.

Context-aware applications are often designed around a set of *if-then* rules. For example, if the application senses a particular situation, then it should perform a particular action. Rules can be created as all the knowledge for each rule is represented in a similar format. It is possible to develop a rule-based system as there are existing rule engines that can determine when the criteria of a rule have been met.

An alternative approach to a rule-based system is to use machine learning. Instead of creating a sequence of rules about how an application should adapt its behaviour, an application developer can gather data on the types of situations or contexts that a user will experience and the types of desired adaptations. Machine learning can then be applied to learn the probabilistic relationships between the situations and the associated adaptations, instead of having these relationships hardcoded and deterministic. This still requires that the application provides the ability to perform context inferencing to map the sensor data to user situations.

One approach to dealing with context ambiguity is to combine multiple, disparate sources of the same type of context to improve the accuracy or dependability of the provided context. A method commonly known as *sensor fusion* is used to combine sensor-related data. For example, in activity recognition, a Hidden Markov Model can be used with different sensors and the fused results can be represented as a confusion matrix over the set of possible activities [16]. Matching fused sensor data with other input sources, such as calendar and web-services, can help deal with context ambiguity.

### B. *Mobile Solutions*

Context information is one of the most significant information sources for mobile applications and the development of mobile devices with embedded sensors has made context data widely available [17]. A need exists for access to real-time information and services anytime and anywhere for people on the move. Context information such as location and calendar data, obtained from mobile phones, can greatly improve the user experience. Mobile smart phones have become the central device of users' day-to-day activities and since the context of mobile users changes dynamically, so do the current needs of the users [18].

Mobile smart phones incorporate several sensors that make it possible to capture environmental contextual information to help individuals to better understand the surroundings that affect their daily lives [15]. The built-in sensors in modern smart phones include GPS (for location context monitoring), image and video sensors (camera), audio sensors (microphones), light sensors, temperature sensors, direction sensors (magnetometer and gyroscope sensors), and movement sensors (accelerometers and rotation sensors) [19].

Collecting and analysing data from sensors requires increased processing, storage and energy resources on mobile devices [15]. Lee *et al.* (2012) suggested that future developers should not only extract high-level context from raw sensing data but also make an effort to support continuous sensing and processing. This optimization process is quite challenging since resource availability of sensor devices and demands from other concurrent applications change dynamically [20].

Extracting useful and meaningful high-level contextual information from low-level smart phone sensor data and other data such as activity data has not been fully explored [19]. This gap has provided an opportunity for mobile applications to leverage user contexts more actively, such as their location, activity, social relationship, and health status [20]. Context is essential in cases such as anti-theft or near-emergency services [7]. To provide these types of services, mobile devices need to be able to clearly identify the specific context of the user. Current mobile devices include sensors from which data such as position, lighting or sound can be obtained, which can help to determine the user's context. However, this raw context data is meaningless for computers so a suitable data model is required to represent and manage the user's context [21]. Mobile context modelling is a process of recognizing and reasoning about contexts and situations in a mobile environment. This process is a fundamental research problem in order to successfully leverage the rich context information of mobile users whilst on the move [22].

Several models were reviewed and the Situation-Aware User Model of Fausto and Alberto [23] was selected as the most complete model. This model, depicted in Figure 2, highlights the separation of user and physical contexts each with their own unique context attributes. However, it does not include a health context, which is seen to be an important part of personal user context [24].

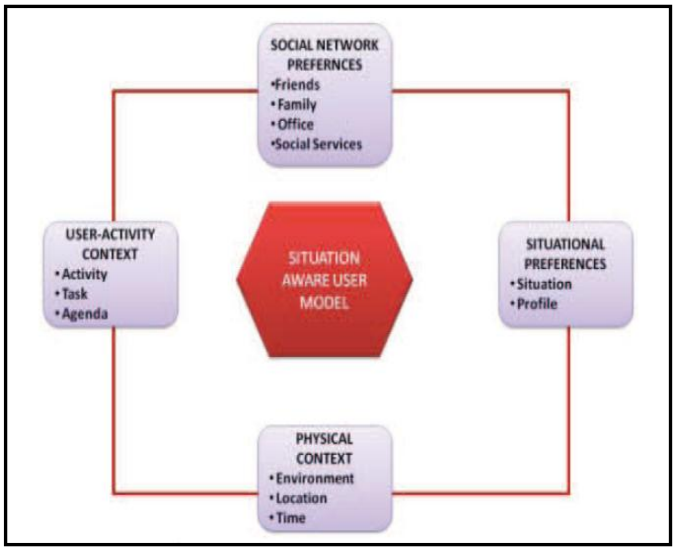



Figure 2: Situation-Aware User Model [23]

The typical architecture, depicted in Figure 1, and the Situation-Aware User Model illustrated in Figure 2, were considered to be the most complete models. However, none of the reviewed models and architectures met all of the requirements to successfully model the personal user context of a mobile user. These two models, together with additional elements, such as an added health context and additional input sources, were used as the basis of the proposed mobile context-aware model.

## III. ANALYSIS AND DESIGN

### A. *Context Attributes*

Context attributes are the building blocks of context dimensions. Context dimensions include a set of context attributes that define the specific context dimension. For example, the model depicted in Figure 2 has a Physical Context dimension with context attributes, which are environment, location and time. Several context attributes for personal user context exist in related literature. The most relevant context attributes were combined and summarized, as depicted in Table 1.

These summarized context attributes are listed with the multiple input sources required to gather data for that attribute. For example, in order to obtain an accurate user location, inputs from the GPS, calendar and web-services can be used. Each attribute was also given example values. The example values can be classified as absolute or fuzzy logic, depending on the actual values. If the raw data is sufficient, the value type will be absolute, or if a semantic meaning can be used to better represent the raw values, then the value type will be classified as fuzzy.

| Attribute | Input Source | Example Value |
|---|---|---|
| Environment | Ambient temperature sensor, sound sensor, light sensor, pressure sensor, web-services. | 32 degrees, cloudy |
| Location | GPS, Time, Calendar, IPS, Wi-Fi, web-services. | lat-35 lng 24/ Summerstrand |
| Time | Mobile device | 04:47 PM/ Afternoon/ 14 March 2014 |
| Spatial | Accelerometer, linear acceleration, orientation sensor, gyroscope, rotation vector and GPS. | 34km/h, North-East |
| Network | Wi-Fi, 3G | WiFi,3G/ NMMU student Wi-Fi, MTN |
| Device Capabilities | Mobile device | Accelerometer available |
| Activity | Accelerometer, linear acceleration. | Driving, still, on foot. |
| Schedule | Calendar | Masters Meeting, Cake day |
| Identity | Facebook, User input(Profile) | Male, Tom, 7 Jan |
| Social | Facebook, Mobile Device | Richard, Friend |
| Physiological | Body Sensors | 64 degrees |
| Medical | User input (Profile), PHRs | Type 2 Diabetes, A+, Hay fever |
| Situation (Device state) | Mobile device | Silent, Loud |
| Interest & needs | User input(Preferences), Facebook | The Beatles, Pizza |
| Points of interest | User input(Preferences) | KFC, Boardwalk |
| Availability | Calendar, Accelerometer, linear acceleration. | Available, Busy |

Table 1: Summarized Context Attributes for Personal User Context

### B. *Extended model*

The proposed model depicted in Figure 3 was designed based on the summarized context attributes illustrated in Table 1, and the Situation-Aware User Model depicted in Figure 2. The proposed model was extended to include a health context with physiological and medical context attributes.

The context attributes of spatial, network and device capabilities were added to the Physical Context. The Preferences of the original model were combined to include a social attribute and additional context attributes were added including interests and needs, points of interest and availability.

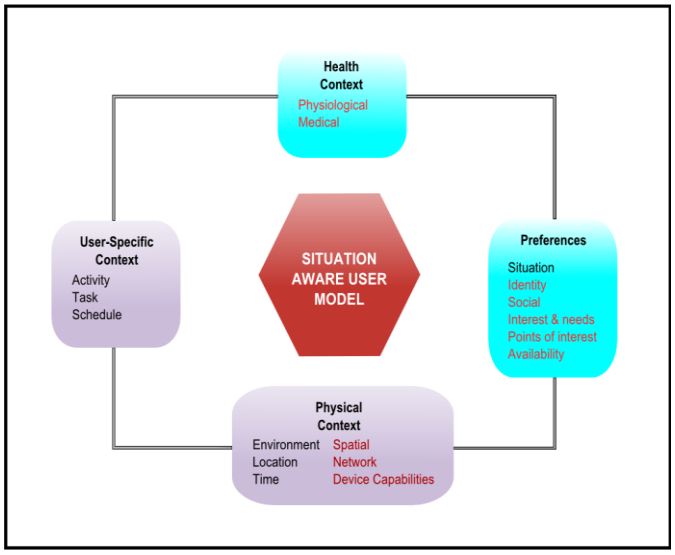



Figure 3: Extended Model with Four Core Dimensions

The proposed model provides a more complete approach to modelling personal user context than the Situation-Aware User Model illustrated in Figure 2.

### C. *Architecture*

The typical architecture depicted in Figure 1 has several aspects such as the middleware layer with a context and a preference manager that are necessary to support the proposed context model. This architecture, however, does not cater for multiple input sources and does not clearly illustrate the different context dimensions supported by the proposed model.

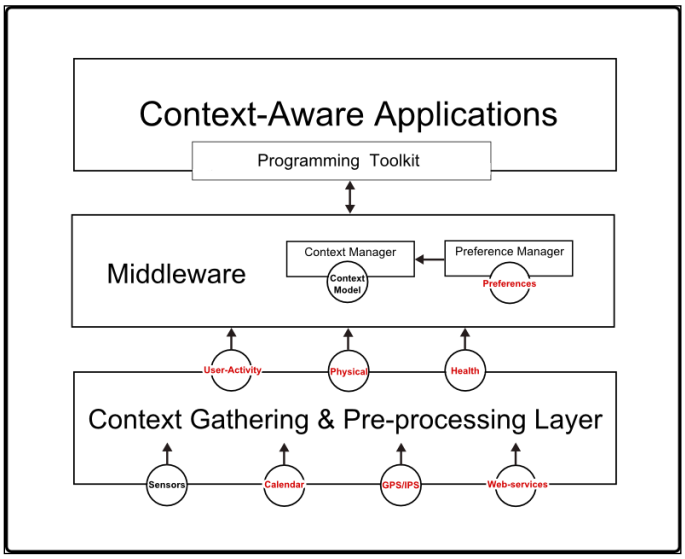



Figure 4: Adapted Architecture with Multiple-Input Sources and Context Dimensions

Thus this architecture was adapted by adding additional input sources such as calendar, GPS and web-services, to the gathering and pre-processing layer. These inputs together with sensor inputs will be processed and combined to form context dimensions, which will represent outputs from the gathering and pre-processing layer.

### D. *Data*

Using the proposed model, the high-level data design was formulated as a UML class diagram, as illustrated in Figure 5.

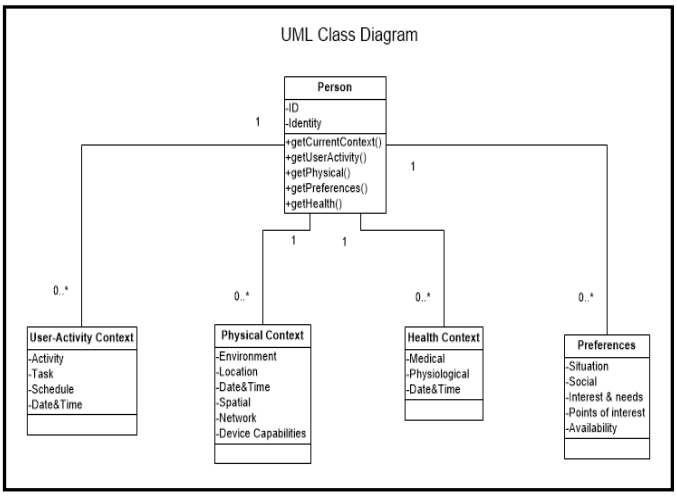



Figure 5: UML Class Diagram High-Level Data Design

As personal user context is focused on the context of a person, the data design was structured around a Person class. This Person class is associated with each dimension class with a multiplicity of 0..*. Each dimension class will be accessible from the Person class. For example, to access the data from the Physical Context class, a method person.getPhysicalContext() will be called. The User-Activity and Health Context Classes include a date and time attribute in the data model as these two classes are dependent on date and time.

## IV. IMPLEMENTATION

The implementation of a prototype based on the proposed model involved the following steps:

- Gathering data
- Filtering data
- Defining Thresholds
- Training data

The implementation of the prototype followed an iterative development process and the Physical Context dimension was implemented first. The implementation of the prototype is in its initial stages and is being developed on Android on a Samsung Galaxy S4. The Samsung Galaxy S4 was selected as it contains a large number of sensors such as the ambient temperature and humidity sensors.

### *A. Data gathering*

The data gathered for the physical context will be obtained from multiple input sources including ambient temperature and humidity sensor, sound sensor, light sensor, pressure sensor, web-services and GPS. The most suitable web-services were selected based on the input required and the data returned, which include:

- OpenWeatherMap API for obtaining weather data;
- Google Geocoding API for determining addresses.

The sensor data that is acquired from the mobile device is regarded as raw sensor data and needs to be pre-processed and filtered to cater for errors.

### *B. Filters*

The most suitable filters were selected based on the errors they deal with and the amount of processing required for the data to be useful. The Bandpass filter was selected as it consists of a high-pass filter as well as a low-pass filter. The high-pass filter will filter out slow drift and offset while giving higher frequency changes. The low-pass filter will have a smoothing effect on the data and filter out high-frequency signals or noise. The Bandpass filter will therefore filter out both low-frequency and high-frequency data and keep the data at a frequency range with fewer errors.

### *C. Thresholds*

Setting thresholds for the sensor data will allow for a higher level meaning to be obtained than simply reporting the absolute value. For example, the light sensor reports its readings in lux and has a dynamic range between 0 and 30,000 lux. The smallest difference in light that the sensor can detect is 1 lux. A value of 10,000 lux is regarded as an overcast day. The following numbers represent typical values that can be expected:

- No moon - 0.001 lux
- Full moon - 0.25 lux
- Cloudy - 100 lux
- Sunrise - 400 lux
- Overcast - 10000 lux
- Shade - 20000 lux
- Sunlight - 110000 lux
- Sunlight max - 120000 lux

These values cover a broad range and cannot always be accurately represented by a qualitative measure such as an "Overcast Day". It is therefore required to train this data with machine learning to overcome this problem. The data collected for the proposed model will use thresholds where feasible and only train data if necessary.

### *D. Training*

As most training algorithms are complex to implement from scratch, the WEKA software package was selected to assist with training the data. WEKA is an open source machine learning framework, which was developed at the University of Waikato in New Zealand [25]. This framework is compatible with several mobile platforms as it was developed and tested in Java and Android. The framework implements various well-known machine learning algorithms including: Naive Bayes, nearest neighbour, neural networks and decision trees.

The most suitable training algorithms for the proposed model still need to be selected. Once the algorithms are selected and the data gathered are trained, a model will be generated for each data set. These trained models, together with set thresholds, will form the input gathering and pre-processing layer.

## V. CONCLUSIONS

Existing context models do not effectively deal with the contextual information related to overall personal user context such as physical, user-activity, health and changing preferences. A suitable model for context-aware applications is required, which should provide support for several tasks dealing with context. These tasks include acquiring context from multiple sources such as device sensors, databases and web-services and performing context interpretation.

This paper has proposed a model to effectively facilitate context-awareness in mobile applications. This model was based on the Situation-Aware User Model, which was adapted to include four core dimensions, physical, user-activity, health and preferences. The model was also adapted to include elements that it lacked, which included a Health Context and additional context attributes in other dimensions.

This research has contributed to existing work by extending an existing model and architecture to support context-awareness in mobile applications. Future work will involve completing the implementation and evaluation of the prototype to determine if the proposed model can effectively support context awareness in mobile applications. Thereafter, the completed prototype will be turned into middleware in the form of an API.

## REFERENCES


[1] A. K. Dey and J. Hakkila, "Context-Awareness and Mobile Devices," p. 13, 2008.

[2] K. Elgazzar, M. Aboelfotoh, P. Martin, and H. S. Hassanein, "Ubiquitous Health Monitoring Using Mobile Web Services," *Procedia Computer Science*, vol. 10, pp. 332–339, Jan. 2012.

[3] P. Klasnja and W. Pratt, "Healthcare in the pocket: Mapping the space of mobile-phone health interventions.," *Journal of Biomedical Informatics*, vol. 45, no. 1, pp. 1–15, Feb. 2011.

[4] J. Zhu, P. Chen, H. Pung, M. Oliya, S. Sen, and W. Wong, "Ubiquitous Computing and Communication Journal Coalition : A Platform for Context-Aware Mobile Application Development Ubiquitous Computing and Communication Journal," vol. 6, no. 1, pp. 722–735, 1992.

[5] K. Mihalic, M. Tscheligi, and U. Unit, "Interactional Context for Mobile Applications," 2006.

[6] N. Bricon-Souf and C. R. Newman, "Context awareness in health care: a review.," *International journal of medical informatics*, vol. 76, no. 1, pp. 2–12, Jan. 2007.

[7] D. R. Ferreira, P. C. Diniz, and P. Chainho, "Context Inference for Mobile Applications in the UPCASE Project," pp. 1–14, 2009.

[8] M. J. Mitchell, "Context and Bio-Aware Mobile Applications," 2011.

[9] D. Traynor, E. Xie, and K. Curran, "Context-Awareness in Ambient Intelligence," *International Journal of Ambient Computing and Intelligence*, vol. 2, no. 1, pp. 13–23, 2010.

[10] S. De and K. Moessner, "A framework for mobile, context-aware applications," *2009 International Conference on Telecommunications*, pp. 232–237, 2009.

[11] B. Hardian, "Context Awareness in Mobile Computing," *Seminar Riset Tenologi Informasi*, 2011. [Online]. Available: http://www.slideshare.net/bhardian/context-awareness-in-mobile-computing. [Accessed: 05-Jun-2013].

[12] P. Demeester, "Context Aware Services," 2010. [Online]. Available: http://www.slideshare.net/guest3cf4991/5-context-aware-services. [Accessed: 05-Jun-2013].

[13] F. Ntawanga, A. P. Calitz, and L. Barnard, "An integrated logical context sensor for mobile web applications," *Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference on - SAICSIT '13*, p. 320, 2013.

[14] D. Huang, W. Liu, and X. Li, "A survey on context awareness," *2011 International Conference on Computer Science and Service System (CSSS)*, pp. 144–147, Jun. 2011.

[15] M. Marcu, N. Ghiata, and V. Cretu, "Extracting high-level user context from low-level mobile sensors data," *2013 IEEE 8th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, pp. 449–454, May 2013.

[16] J. Lester, T. Choudhury, and G. Borriello, "A Practical Approach to Recognizing Physical Activities," pp. 1–16, 2006.

[17] H. Wolf, K. Herrmann, and K. Rothermel, "Robustness in context-aware mobile computing," *2010 IEEE 6th International Conference on Wireless and Mobile Computing, Networking and Communications*, pp. 46–53, Oct. 2010.

[18] G. Eichler, K. Lüke, and B. Reufenheuser, "Context information as enhancement for mobile solutions and services," pp. 1–5, 2009.

[19] A. M. Otebolaku and M. T. Andrade, "Recognizing High-Level Contexts from Smartphone Built-In Sensors for Mobile Media Content Recommendation," *2013 IEEE 14th International Conference on Mobile Data Management*, pp. 142–147, Jun. 2013.

[20] Y. Lee, Y. Ju, C. Min, J. Yu, and J. Song, "MobiCon : Mobile Context Monitoring Platform," pp. 109–111, 2012.

[21] D. Martin, C. Lamsfus, and A. Alzua, "Mobile Context Data Management Framework," *2011 Fifth FTRA International Conference on Multimedia and Ubiquitous Engineering*, pp. 73–78, Jun. 2011.

[22] T. Bao, H. Cao, E. Chen, J. Tian, and H. Xiong, "An Unsupervised Approach to Modeling Personalized Contexts of Mobile Users," *2010 IEEE International Conference on Data Mining*, pp. 38–47, Dec. 2010.

[23] M. Fausto and P. Alberto, "Context Planning and User Profiling in Mobile Services," no. Mc, pp. 301–306, 2010.

[24] D. Zhang, Z. Yu, and C.-Y. Chin, "Context-aware infrastructure for personalized healthcare.," *Studies in health technology and informatics*, vol. 117, pp. 154–63, Jan. 2005.

[25] M. Hall, H. National, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software : An Update," vol. 11, no. 1, pp. 10–18.



**Direshin Pather** received his BCom Honours degree in 2012 from the Nelson Mandela Metropolitan University (NMMU). He is presently studying towards his MCom degree at the same institution. His research interests include mobile computing, context-awareness, location awareness and ubiquitous computing.

# Optical Character Recognition using Minutiae based feature detection

Pieter Erasmus, Trevor Ho, Yuko Roodt
Hypervision Research Laboratory, Department of Electrical and Electronic Engineering
University of Johannesburg
South Africa
Email: p.erasmusjnr@gmail.com, trvrho@yahoo.com, yukor@uj.ac.za

***Abstract*—Optical Character Recognition (OCR) forms part of image processing where automatic identification techniques are used for pattern recognition. In OCR, there are two types of techniques used for pattern classification, namely: matrix matching and feature extraction. Generally characters are handwritten text or printed text which consists of various types of fonts. This makes pattern recognition a cumbersome process therefore feature extraction techniques are explored. This paper focuses on using minutiae extraction to identify features of skeletonized characters in order to perform OCR. A unified approach is explored to deal with both printed and hand-written text with the exclusion of cursive handwriting.**

***Keywords*—OCR, Minutiae Extraction, Feature Extraction, Skeletonization.***

## I. Introduction

Optical Character Recognition is the process of performing automatic identification of characters by recognition of patterns within an image. There are numerous applications for OCR such as automatic number plate recognition systems[1] or document analyses systems[2].

There is, however, a problem in standardization of an image before actual pattern recognition can commence. This will be regarded as the pre-processing of an image which includes noise removal. This convolution process determines which regions need to be discarded and which contains substantial amounts of information. Noise can be introduced when a photograph is taken or a scan is made. If lighting is considered, the noise added due to the scanning process of a document on a flat piece of paper is considerably less than that of a scan of a thick manuscript or textbook. The latter causes the ridges to be exposed to less light resulting in text and shadows regarded as a single object. On the contrary, too much light will subdue the edges on the object making it deficient. This phenomenon is also present when using a camera to capture the document. A single light source incident on the document can change the contrast of the image drastically and make segmentation problematic as discussed by Wojciech et.al[3]. To account for this the image is binarized with a threshold. These techniques are explored by Seeger et. al[4] and by using dynamic threshold calculations [5] the best approximation is obtained. This, however, is not as effective as some thresholds can leave behind unwanted pixels and remove some necessary segments.

Another drawback arises if affine transformations are applied to an image which changes an object's thickness. Consequently, when filtering is applied, some objects may be discarded. When isolating objects for character segmentation, unidentified characters will be detected when using a matrix matching technique such as OSTD[6] for OCR.

Matrix Matching is the most common way in which OCR is achieved. This technique entails the test character matrix to be compared to a predefined character matrix template of a certain character. Small changes are made to the test character image such as scaling and rotating. The similarity of these matrices are considered and a match is made based on the highest similarity. This technique is particularly useful when recognizing consistent fonts and handwritten characters of the same authors. Nevertheless, this method yields inconsistent classification results when the deviation is large between similar characters.

This leads to another method developed for OCR which encapsulates features to identify characters. A feature can be defined as a collection of neighbouring pixels that present specialized properties[8]. In feature extraction, the properties are extracted and compared to trained descriptions of the character that is acquired during the training phase. For this purpose, the features defined in this context are called minutiae points. An attempt to use the same method of minutiae extraction for biometrics is applied to optical character recognition.

Similar to feature extraction using Least Square Support Vector machine for OCR by Jianhong[9], minutiae points are distinctly defined by their curves and certain ridges. Minutiae points are primarily used in fingerprint matching that forms part of the science of biometrics. These points are the critical points in a skeletonized image of the fingerprint. There are three main categories of minutiae points known as ridges, bifurcations and terminations, which are used as features when fingerprint matching is done.

In this paper we shall extract minutiae points from the skeletonized image and use these points for features to do character recognition. The types of features that will be exploited for identification of a character are as follow: minutiae point type, number of minutiae points, connection type and distance between points.

The first section of this paper consists of the methods used to pre-process the image. Subsequently the feature extraction process will be explained. The experimental setup follows where test datasets are explained and then the results will be discussed and analyzed.

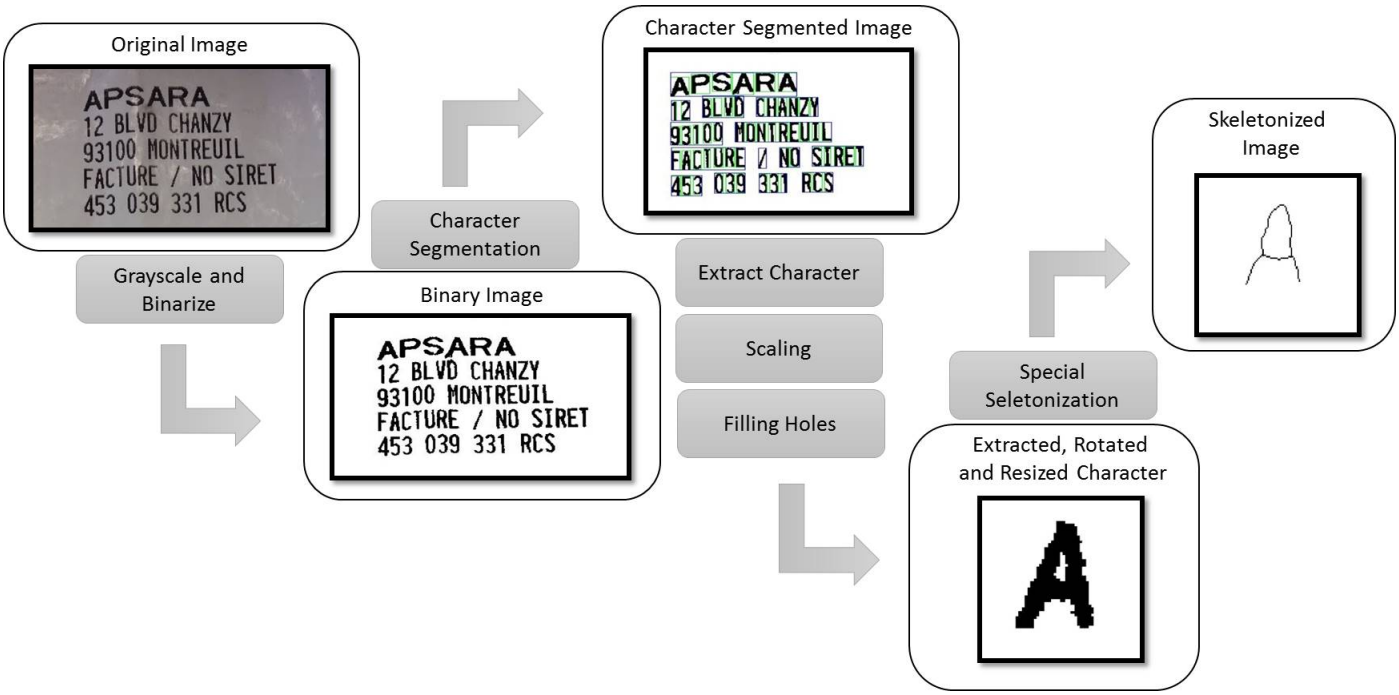



Fig. 1. Overview of pre-processing steps

## II. Pre-Processing

Pre-processing is focused on getting an image containing only a single character in the center of the image pane and furthermore skeletonizing the image appropriately. Figure 1 indicates the steps that are taken to enhance the image for feature extraction. Consider the original image in Figure 1 which will be used for analysis. This image is grayscale and binarized in order to make character segmentation easier. Each character is then extracted, rotated and resized in order to produce a normalize image. This image is then skeletonized in order for minutiae points extraction to be performed.

### A. Grayscale and Binarization

The image first undergoes a transformation to a grayscale image, and then the image is binarized. Real-time adaptive image thresholding from Bradley[10] is used. This result of binarization of an input image is presented as Binary Image in Figure 1.

### B. Character Segmentation

Character segmentation is now done on the binary image by using the method of connected component labelling[11]. This method entails labelling all connected pixels as one object using 8 connected neighbours. The 15% trimmed mean height and width is calculated. All possible objects within three times the mean height are considered possible characters. All objects that are square-like and have a mass of more than 82% are considered possible dots. Possible dots and characters are combined by observing their relative positions to one another and also taking into account the distances between their center points. Character Segmented Image in Figure 1 shows the different characters extracted and group them to form words.

### C. Image Moments

After character segmentation the character image is standardized such that the same characters have similar orientations. Firstly the image orientation is determined using image moments[12]. For a m x n image having moment M of degree (p + q) and centroid $\{\bar{x}, \bar{y}\}$ the central moment is computed using equation 1. The image is then rotated by the angle obtained by using equation 2.

$$u_{pq} = \sum_{m}^{p} \sum_{n}^{q} \binom{p}{m} \binom{q}{n} (-\bar{x})^{(p-m)} (-\bar{y})^{(q-n)} M_{mn} \quad (1)$$

$$\theta = \frac{1}{2} arctan(\frac{2u'_{11}}{u'_{20} - u'_{02}}) \quad (2)$$

### D. Scaling

The processed images can be of different sizes, hence the image has to be rescaled for feature detection to be generic. Scaling must be executed appropriately such that the result is small enough to be processed quickly, yet large enough not to lose important features. A proportionality resize to a height of 100 pixels achieves this objective. Figure 1 shows the extracted character ”A” after rotation and scale was standardized.

### E. Filling Holes

The image is analysed to fill small holes in order to obtain a correctly skeletonized image. This is achieved by negating the binarized segmented image. Using connected component labelling of 4 connected neighbours, possible objects are identified. The objects that have a mass of less than 40 pixels is considered a hole and are filled.

### F. Special Skeletonization

To extract the best skeleton image, skeletonization is performed using 20 different thresholds and the number of end-points and branch-points of each is computed. The skeletonization proposed by Howe[13] was used. The first image that produced the least amount of both these feature points is selected. Furthermore a morphological thinning method is performed on the image to ensure a width of one pixel skeletons. Skeletonized Image in Figure 1 shows the result.

## III. Feature Extraction

This section defines the features that are used for character classification. These features include: Identified Minutiae points, which types of minutiae points are connected, how these points connect and beginning direction of a point. Possible corners are identified based on direction changes, distance between points and relative positions of the points are also considered. All off these features combine into a vector for which every character has a set of these feature vectors. Recognition is done based on the similarities between the character's feature vectors. Consider the following skeletonized image containing an "R" in Figure 2 as an example.

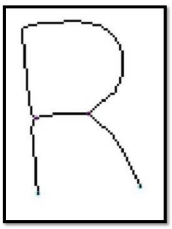


Fig. 2. Skeletonized character - "R"

### A. Minutiae Extraction

The whole image is examined to identify where branch-points and end-points occur. This is done by extracting a 3x3 matrix of pixels from every point in the skeleton image. The position of a point is found in the centre of the submatrix. The submatrix is then analysed to determine if it has the properties of a selected feature. An end-point pattern has the centre pixel and any one of the surrounding 8 pixels as illustrated in Figure 3(a). A connected pattern has the centre pixel and any two of the surrounding 8 pixels as illustrated in Figure 3(b). A branch-point pattern has the centre pixel and any three of the 8 surrounding pixels as illustrated in Figure 3(c).

Now the minutiae points are extracted by using these feature points. A minutiae end-point will be a feature end-point such as displayed in Figure 4 points E1 and E2. Minutiae branch-point will be a feature branch-point such as the points B1 and B2 illustrated in Figure 4.

### B. Connecting Minutiae Points

Adding to the feature vector is the minutiae points the current minutiae point is connected to. By analyzing each of the possible routes of the current point, it has to be connected to another minutiae point. To determine the end minutiae point, a marker is moved along every possible path, ensuring that it does not move back on the same path. By keeping track of all the minutiae points for a given image a connected feature will always end up at exactly the position of another feature, or itself. Figure 4 indicates that points B1 is connected to E1 and twice to B2.

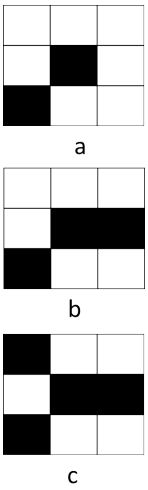



Fig. 3. (a) Endpoint Pattern (b) Connected-point Pattern (c) Branchpoint Pattern

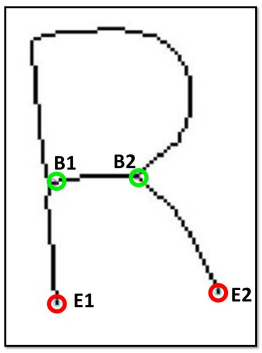



Fig. 4. Minutiae Points Identified

### C. Distance between points

By incrementing the distance as the marker moves along each path, the distance between the different minutiae points can be determined. In Figure 4 the shortest distance between B1 and B2 is 24 pixels where the longest distance is 118 pixels.

### D. Connection Type

The connection type describes the way in which two minutiae points are connected to one another. The number of new direction changes determine the curvature. Figure 5(a) shows a 3x3 sub matrix where there is no direction change. Figure 5(b) shows the sub matrix when there is a single direction change.

In Figure 4 B1 and B2 are connected through 2 line segments with curvature 2 and 8 respectively.

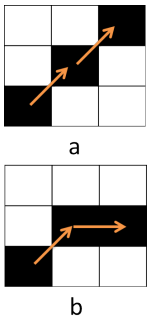



Fig. 5. (a) No Direction Change (b) Direction Change

### E. Possible Corners

Detecting the position of possible corners makes use of the direction changes between two minutiae points. Based on a subset, consisting of 40 direction changes, if there are more than 3 new direction changes, the subset is considered to have a corner in it. In Figure 6 points B1 and B2 are connected through a straight line segment and a curved line segment. From the example shown in Figure 6, the curved line segment was found to have 2 corners.

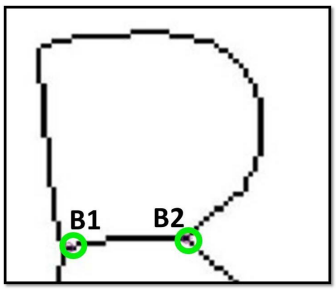



Fig. 6. Possible Corners Identified

### F. Relative Position

Consider two minutiae points connected to one another. Based on their positions on the image, the relative position of the one point can be obtained with respect to the other and the angle can be calculated. If B2 is the reference point, B1 will be West where E2 will be South-East as shown by Figure 7.

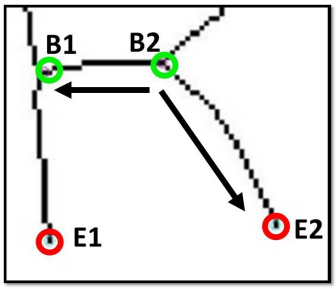



Fig. 7. Relative Position Identified

### G. Begin Direction

Every minutiae point has starting directions to the minutiae points it is connected to. Extracting this feature enables better classification of the current position of the minutiae point if two points are connected through a curvature. As the first and second direction change may be part of curvature it does not resembles the actual begin direction. Therefore the third direction change is considered. Figure 6 indicates that B2 has a begin direction West to B1 for the shortest distance and begin direction North-East for the longest distance.

## IV. Experimental Setup And Results

The system was trained by a set of characters to identify a set of reference features. Training began by using ideal enhanced electronic letters. Testing was done by using photographed tickets and alphabets of various fonts. Every match had a percentage associated with it which was calculated in the following manner:

$$\bar{a}_{ij} = |\bar{b}_i - \bar{c}_j| \tag{3}$$

$$A_{ij} = \frac{2}{7}(\sum_{x=1}^{6} 5 - \bar{a}_{ij}(x) + (\sum_{x=7} 5 - \frac{1}{5}\bar{a}_{ij}(x)) \tag{4}$$

$$S = 10(n) + 7\frac{\sum_{y=1}^{c} max(A_y)}{c} \tag{5}$$

Where $\bar{b}$ is the feature vector of the reference character and $\bar{c}$ is the feature vector of the computed character. A score matrix $A$ is constructed with dimensions [$r$,$c$]. $r$ is the number of feature vectors in the reference character and $c$ the number of feature vectors in the computed character. $S$ is the score out of a hundred and n is the number of the various correctly counted features.

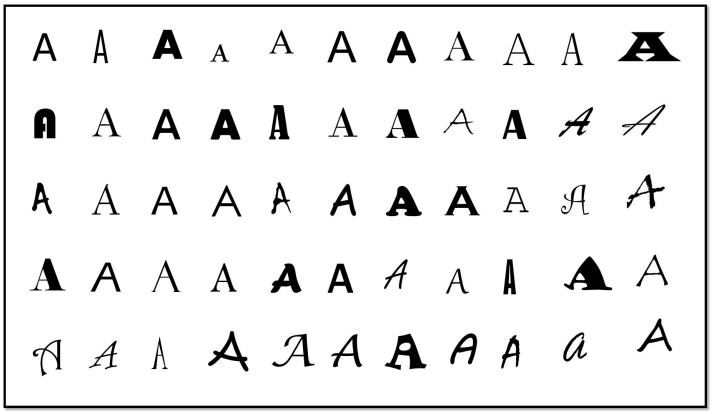


Fig. 8. Example of the letter "A" variations in training data set

The system is trained with a dataset. This is done by taking variations of the same character to detect the distinct features. Figure 8 shows the data set containing 55 different fonts that were used for training A to Z, a to z and 0 to 9. This results in a training data set of 3410 characters. This dataset is chosen for generic character recognition, but can be adapted for specific font types.


Fig. 9. Example of captured input image

### A. Ideal Enhanced Character Analysis

A screen shot of the training set with various different fonts were analysed. Figure 10 shows the percentage correctly detected characters for the different fonts. These results exclude characters that looks very much alike such as ”Z” and ”z” or ”0”,”o” and ”O” as these results are significantly lower as seen in Figure 11.

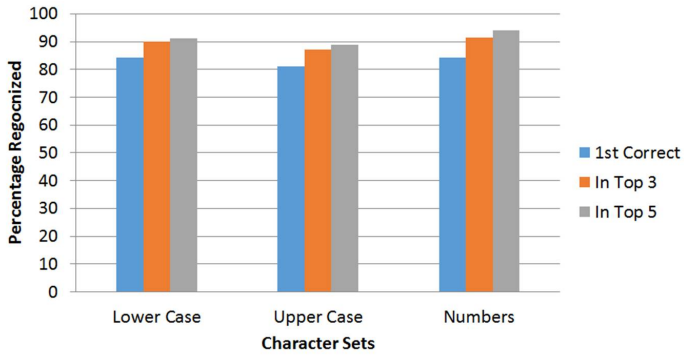



Fig. 10. Results for Ideal Text Recognized.


Fig. 11. Results for similar looking characters.

### B. Photographed Tickets Analysis

The original image in Figure 9 was used for analysis. Figure 12 shows the precentage correctly recognized characters. Due to the limited training set and the font difference of the ticket, less than 80% recognition was obtained. The result can be improved through additional training as well as training for specific fonts.

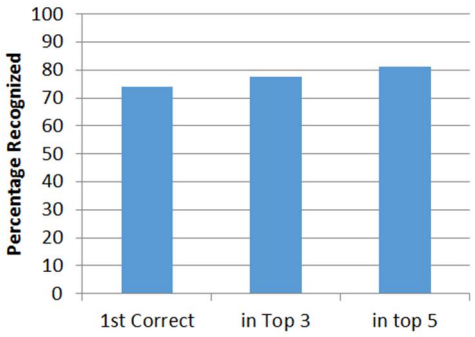



Fig. 12. Results for ticket analysis

### C. Handwritten Character Analysis

Classification of hand-written text is a difficult problem to solve due to the large variability of characters. The letters itself change constantly from person to person and even when written by the same person. Therefore the results for the hand-written text have relatively low percentages of success. Strokes and extra lines from a person’s handwriting are also of great concern. An over stroke or under stroke can affect the matching of a letter to the correct character. Figure 13 shows an example of a hand written ”A” for two candidates. Figure 14 shows the results for hand written characters for 2 different candidates.

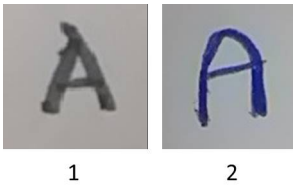



Fig. 13. Example of a handwritten ”A” from 2 candidates

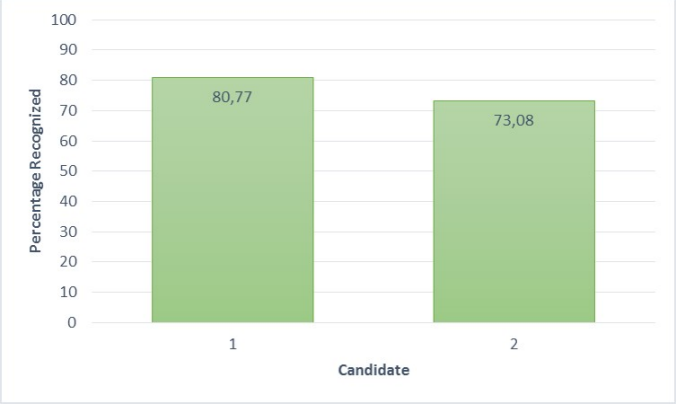



Fig. 14. Precentage handwritten characters recognized

The printed text performed relatively well compared to that of the hand-written text, due to the fact that the printed text was compared to that of a database contained characters trained text that are printed characters. It is also more accurate because of conformity and stability in fonts where the shape of the character stays constant while in hand-written characters, they

change constantly even with the same test subject writing the same character.

### *D. Affine change in character image*

Capturing a character image from different angles can increase classification difficulty when OCR is done as seen with different variations of captured images in Figure 15. The pre-processing stage normally tries to account for affine changes. However, Figure 16 shows results when affine correction is discarded and OCR is done directly on the input image. These results show minor attenuation in recognition percentages when an affine transformation is introduced.

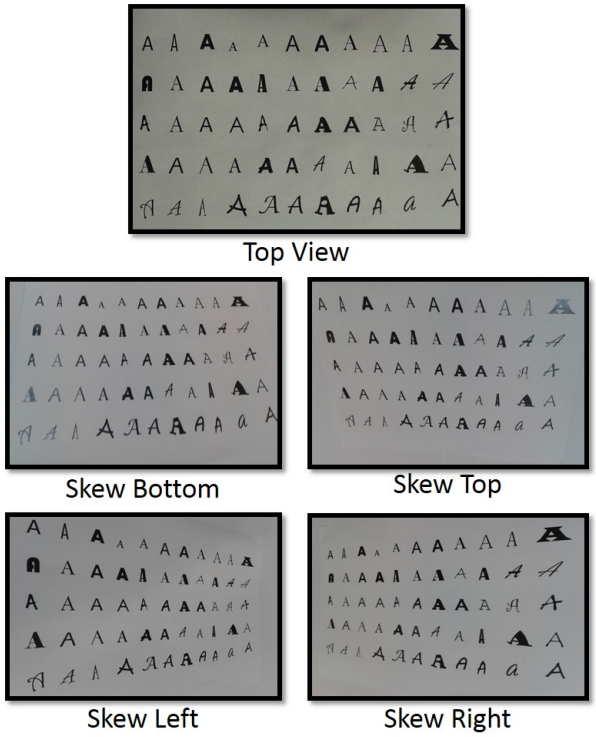



Fig. 15. Various Captured Images with applied affine transformation of training set "A"


Fig. 16. Precentage of characters recognized for various skewness types

## V. Conclusion

Using minutiae extraction to identify features from a skeletonized image requires a dynamically structured adaptive pre-processing system. This would enable feature extraction by using specified minutiae to be done more accurately. An OCR system is developed that extracts minutiae features used to recognize a combination of hand-written and printed text. Printed text is easier to normalize and extract features resulting in better character recognition results than that of hand-written text. Affine transformation introduction has a minor influence on the recognition of characters in this way. Additional feature extraction and larger training sets along with post processing can improve minutiae bases optical character recognition.

## References


[1] N Sulaiman, S.N.H.M Jalani, M Mustafa and K Hawari, "Development of automatic vehicle plate detection system",*System Engineering and Technology (ICSET), 2013 IEEE 3rd International Conference*,pp.130-135, Aug. 2013.

[2] V Kluzner, A Tzadok, D Chevion and E Walach , "Hybrid Approach to Adaptive OCR for Historical Books" *Document Analysis and Recognition (ICDAR), 2011 International Conference*, pp. 900-904, Sept. 2011.

[3] W Bieniecki, S Grabowski, W Rozenberg, "Image Preprocessing for Improving OCR Accuracy",*Memstech, Lviv-Polyana, Ukraine*, 2007.

[4] M seeger, C Dance, "Binarising camera images for OCR",*Document Analysis and Recogniton. Proceedings. Sixth International Conference on, pp.54-58*, 2001.

[5] C Ma, Q Bai, X Chen "The new improvement of multi-threshold dynamic binarization for bill images,"*Software Engineering and Service Science (ICSESS), IEEE 3rd International Conference on, pp.67,70, 22-24 June 2012*

[6] M Imran, J Hossain, T Dey, B Debroy, A Abir, "OSDT: Outer Shape Detection Technique for Recognition of Bangia Optical Character",*"12th International Conference on Computer and Information Technology, (ICCIT 2009)*, Dept Computer Science & engineering, Khulna University, December 2009

[7] A Nikolaidis, "Affine transformation invariant image watermarking using moment normalization and radial symmetry transform," *Image Processing (ICIP), 18th IEEE International Conference on, pp.2729-2732, 11-14 Sept. 2011.*

[8] R Ramanathan, A Nair, L Thaneshwaran, S Ponmathavan, N Valliappan, "Robust Feature Extraction Technique for Optical Character Recognition",*International Conference on Advances in Computing, Control, and Telecommunication Technologies*Department of Electronics & Communication Engineering,Amrita Vishwa Vidyapeetham,Coimbatore, India, 2009.

[9] J Xie , "Optical Character Recognition Based on Least Square Support Vector Machine",*Third International Symposium on Intelligent Information Technology Application*School of Electronics, Jiangxi University of Finance and Economics, Nanchang, China, 2009.

[10] D Bradley, G Roth,"Adaptive Thresholding Using the Integral Image", Carleton University, Canada, 2005.

[11] L Di Stefano, A Bulgarelli, "A Simple and Efficient Connected Components Labeling Algorithm", University of Modena, Modena, Italy, 1999.

[12] L Kotoulas and I Andreadis, "Image analysis using moments",Democritus University of Thrace, pp.74-85, 2005.

[13] N Howe, "Better Skeletonization", Internet: http://www.mathworks.com /matlabcentral/fileexchange/11123-better-skeletonization , Jul. 2007.

# HADEDA: A Concurrent Music Synthesis Project for the XMOS startKIT

James Dibley and Karen Bradshaw
Department of Computer Science
Rhodes University
Grahamstown 6140
Email: g11d3593@campus.ru.ac.za, K.Bradshaw@ru.ac.za

***Abstract*—The HADEDA (Highly Adaptive Device Emitting Digitally-synthesised Audio) project implements a software-defined audio synthesiser with stored musical sequences and real-time synthesis controls, using the XMOS startKIT low-cost development platform. This paper describes the design and implementation of the project as a practical demonstration of XMOS technology. We also demonstrate the use of XC programming techniques and the xTIMEcomposer development tools to build and test a concurrent real-time embedded application easily.**

***Index Terms*—audio synthesis, concurrency, embedded systems, real-time control, multimedia, software development**

## I. Introduction

The HADEDA (Highly Adaptive Device Emitting Digitally-synthesised Audio) project implements a software-defined audio synthesiser based on the XMOS startKIT low-cost development platform [1]. The project provides a clear, self-contained 'working example' of message-passing concurrency and a number of XMOS microcontroller (MCU) programming techniques. It can also be used as a demonstration of aspects of the XMOS 'xTIMEcomposer' development tools.

This project engages with a novel architecture and a rich set of development tools and techniques. The XMOS MCU is proving to be a popular research platform [2] [3] [4]. This project aims to provide a demonstration of some of its capabilities. The primary purpose is to demonstrate:

- message-passing concurrency on the XMOS platform
- some commonplace XC programming techniques
- some typical uses of the xTIMEcomposer development tools

The secondary objective of the project is to develop a programmable musical instrument: to produce a minimal working device, the user only needs to attach earphones or a speaker to one of the startKIT's expansion pins.

### A. Project resources

The XMOS XS1 series of MCUs implement a rich set of concurrency, synchronisation, timer and general-purpose input-output (GPIO) features in hardware, with the goal of providing programmable microcontrollers capable of deterministic concurrent execution defined in software [5] [6]. Software for the XS1 MCU can be programmed in C or C++. The XC programming language provides 'multicore extensions' to C for access to the MCU's specialised features [7].

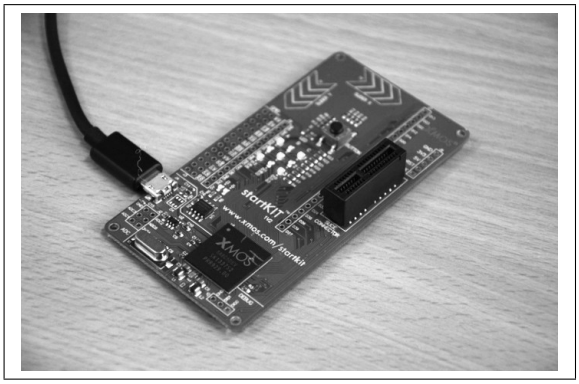



Fig. 1. The XMOS startKIT

The XMOS startKIT (Fig. 1) is a low-cost educational platform based around a single XS1-A8-64-DEV MCU. The startKIT provides a wide variety of hardware interfaces, including a GPIO header for use with the Raspberry Pi, an integrated analog-digital converter (ADC), a PCI Express slot, and a range of other user-assignable pins, as well as onboard hardware (eleven LEDs, one button, and two four-zone capacitance touchplates) and flash RAM.

Applications for the startKIT can be built within the xTIMEcomposer software tools [8]. The startKIT may run these applications interactively, or they may be flashed to the device to run independently of the software tools. There is no limitation to the number of times a startKIT can be flashed.

The HADEDA application requires the user to connect a monophonic audio connector to two pins on the board. No other external hardware is required, although adding a Class D audio amplifier will improve the quality of audio playback. The application depends on two freely available XMOS modules, `module_capacitive_sensing` and `module_startkit_gpio` [9].

The project is finally intended to provide a basis for experimentation. The central components of the HADEDA application are a sequencer process (that generates musical events) and a

This work was undertaken in the Distributed Multimedia CoE at Rhodes University, with financial support from Telkom SA, Tellabs, Genband, Easttel, Bright Ideas 39, THRIP and NRF SA (TP13070820716). The authors acknowledge that opinions, findings and conclusions or recommendations expressed here are those of the author(s) and that none of the above mentioned sponsors accept liability whatsoever in this regard.

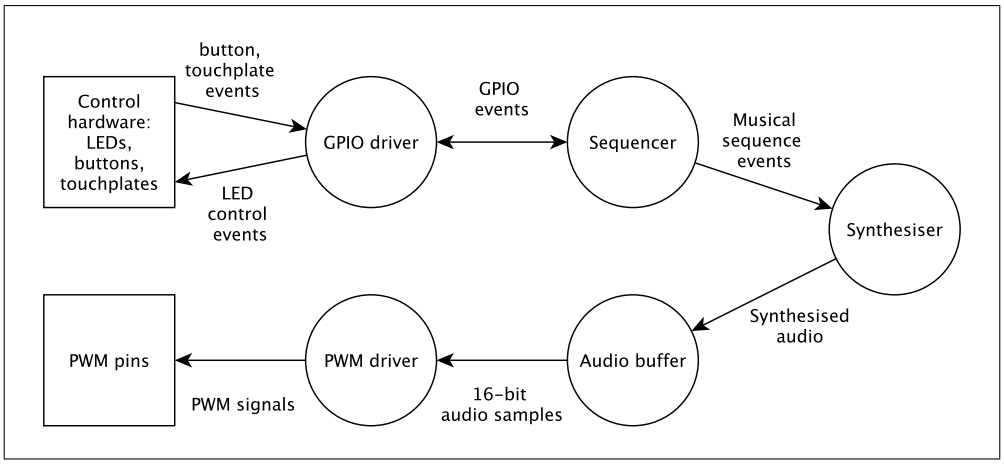



Fig. 2. Top-level design of the HADEDA project, showing channel and port communications between processes

synthesiser process (that generates audio on the basis of these events). Each process is user-programmable. Since the XMOS approach to concurrency facilitates clear separation between functional processes, either of these processes may be replaced with an alternative implementation with negligible alteration to the other, as long as the basic message-passing interface is maintained.

## II. Project requirements

The HADEDA produces a single digital audio signal at a 48kHz sampling rate. It requires preset sequencer and synthesiser programs, a means to switch between them, and real-time control over some of the synthesis parameters.

The startKIT is not equipped with a DAC or other audio hardware, but [10] contributes open-source code demonstrating the use of pulse-width modulation (PWM) to deliver audio signals over two output pins. Ideally this signal would drive a Class D audio amplifier. In practice, the pins can be connected directly to a small speaker or earphones with moderately acceptable results[1].

The startKIT has limited hardware controls - a button and two touch plates, with a 3×3 LED matrix. The button should be used to step through a series of 'preset' configurations for the sequencer and synthesiser, the LED matrix indicating the current configuration, while the touch plates should be used to implement limited real-time control over synthesis parameters.

### A. The synthesizer

The synthesiser must be capable of generating an audio sample fast enough to meet the 48kHz sampling rate (that is, in approximately 21 $\mu$s). Synthesiser configuration presets must be programmed in device firmware.

It must produce a single periodic audio signal, with frequency continuously variable from around 0Hz to 24kHz. The synthesiser must be able to produce a variety of waveforms (or 'timbres'). The synthesiser must implement 'envelopes' that control the amplitude and timbral variation of the output signal over time (that is, in relation to the beginning and end of musical notes).

[1]This relies on the physical filtering properties of the speaker or earphones to filter out the PWM carrier, and is not optimal. However, any better solution would require extra hardware.

### B. The sequencer

The startKIT's limited controls are not sufficient for the input of musical sequences, so these must be programmed in device firmware.

Sequences of musical events correspond to an overall speed or *tempo* measured in beats per minute (BPM). The musical sequences must provide control over placement, pitch, duration, and dynamic of each musical event. In addition to the musical sequence, the sequencer's presets must store and recall 'metadata' – e.g., the tempo of the sequence, the number of steps in the sequence, and the number of steps that make up a musical 'beat'.

The sequencer must be capable of playing back sequences at a wide range of tempi (e.g., 50–200BPM). Each programmed sequence should recall its own metadata. The sequencer should provide a visual indication of the current sequence's tempo.

## III. Design

The top-level design of the instrument is informed by a Real-Time Structured Analysis & Design (RTSAD) approach [11], and is depicted in Fig. 2. Communication between processes is performed via XC *channels*. Communication involving hardware resources is performed via XC *ports*.

Of the processes shown, the PWM driver and audio buffer are derived from [10], and the GPIO driver is provided by XMOS [9]. The sequencer and synthesiser processes are original to this project, and this section of the paper focuses on these.

### A. The synthesiser

The synthesiser accepts musical event sequences and real-time control messages, including a 'load preset' instruction. The synthesiser implements phase modulation (PM) synthesis, as used by many digital synthesisers, most notably the Yamaha DX7. For a detailed discussion of PM, see [12] [13].

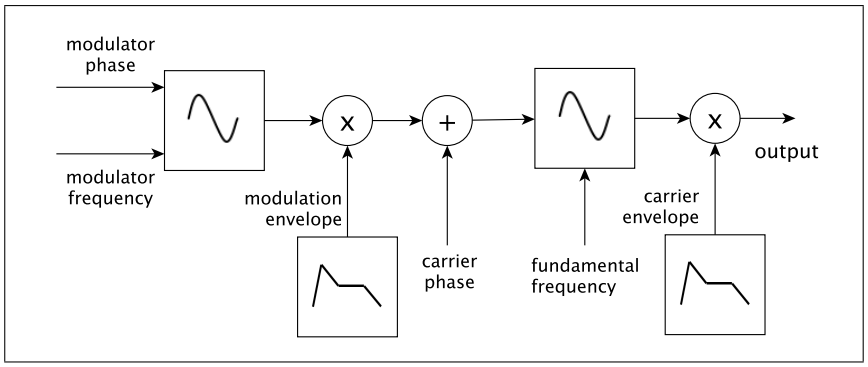



Fig. 3. Phase modulation oscillator topology

The synthesiser is based around a pair of sinusoidal oscillators (Fig. 3). One oscillator (the 'carrier') provides a signal tuned to a fundamental frequency (e.g., a musical pitch). The other oscillator (the 'modulator') is tuned to a ratio of the fundamental. Its output modulates the phase accumulator of the carrier oscillator, producing a wide variety of timbres.

The sinusoidal oscillators are implemented through Direct Digital Synthesis [14]; a lookup table holds samples of one quadrant of a sine waveform; table entries are read, sometimes in reverse order and/or then inverted, and interpolated to render a full sinusoidal cycle.

Each oscillator's output is shaped by an envelope generator (EG), providing a standard attack/decay/sustain/release (ADSR) envelope. These control the modulation imposed on the carrier signal and the amplitude of the carrier's modulated output. The attack phase is triggered by the start of a musical event, and the release phase is triggered by the end of a musical event.

The EGs were adapted from Redmon's tutorial [15]. This required the use of the C standard math library and some floating point calculations. Since XC does not support the C float type, these calculations are implemented in C functions, providing an opportunity to demonstrate integration of XC and C code. All other calculations within the synthesiser are implemented in fixed-point arithmetic.

All of these synthesis parameters may be 'preset' in device firmware. The startKIT's touchplates provide programmable real-time control of these parameters.

### B. The sequencer process

The sequencer process generates musical events from preset sequences. It also intercepts real-time control events from the GPIO process (button presses, touchplate swipes) and processes these as required, in some cases by relaying to the synthesiser.

Each preset sequence, in addition to musical events, contains metadata as defined in Subsection II-B. This interval controls the rate at which musical events are generated, and is recalculated whenever a new preset is loaded.

The sequencer is patterned as a typical 'step sequencer'. A musical sequence is encoded as a series of steps that are 'played' in a continuous loop. Each step is encoded as a 32-bit unsigned integer, meaning that any step can be sent to the synthesiser process in a single XC channel operation.

Any step may encode the beginning or end of a musical note, its pitch, its loudness, and synthesis parameters. The length of a step in musical terms can be set on a per-sequence basis; longer note durations can be programmed by 'slurring' consecutive steps. Sequences are programmed in device firmware and cannot be edited in real-time.

## IV. Implementation

### A. Concurrency and communication between processes

The top-level organisation of the software (Fig. 2) is expressed straightforwardly in XC. In Listing 1, `chan` and `interface` declarations precede a `par` statement that launches each of the following functions as a parallel process.

XC *channels* provide a communication link between processes. The `c_pwm` and `c_synth_audio` channels convey 16-bit audio samples between the synthesiser, buffer and PWM driver processes. Where messages may be multi-purpose, as between the sequencer and synthesiser processes, an XC *interface* can abstract channel messages into a set of *functions*. The interface between the sequencer and synthesiser is shown in Listing 2.

An interface exists between a *server* process and one or more *client* processes. In this example, the synthesiser process is the server and the sequencer process is a client. The functions declared by the server are implemented in the context of the server's `select` statement, which is described in the following subsection.

### B. Event-based processing

The XC model of concurrency defines processes as 'event-driven' [7, 13]. Event sources can be diverse, including messages received over channels, signals received from hardware interfacing, and internal timer intervals. Events that drive a process are defined in an XC `select` block.

For example, the synthesiser process is driven by four events. Three of these events are the sequencer messages defined in the `pm_oscil_if` interface (Listing 2):

- a sequencer step, containing musical event information;
- an event that loads a synthesiser preset;
- a real-time control event that alters a synthesis parameter.

The fourth event is a tick from an internal timer with an effective rate of 48kHz. When this event fires, the synthesiser

must generate and communicate an audio sample to the audio buffer process.

The implementation of all four events is shown in Listing 3. Line 9 shows the fourth event and its handler. The `generate_sample` function returns an unsigned int containing a 16-bit audio sample. This is directly output (`<:`) over the `c_synth_audio` channel. Line 14 shoes the sequencer step processing; the remaining cases update synthesis parameters. On line 21, a real-time control event is parsed to determine which of the two touchplates was manipulated.

### C. Timed processing

The sequencer process offers another example of timed processing, of greater interest because it demonstrates dynamic programming of the timer interval. Listing 4 shows responses to two events: a user button-press, which loads the next preset sequence from `sq_p` and recalculates the timer interval based on the information in `sq_p_meta`; and the XS1 timer `t` reaching a timer interval, at which point it sends the stored musical event to the synthesiser process across the `osc_if` interface.

Line 21 implements a visual indicator ('metronome') of the sequencer tempo; on each 'beat' within the sequence, the `p_metro` port is driven low, then driven high on the following step, flashing an LED indicator.

### D. Hardware input-output

Hardware input-output within the HADEDA project is performed in three contexts:

- Real-time control of the synthesiser and sequencer through the XMOS `module_capacitive_sensing` and `module_startkit_gpio` modules.
- The visual metronome described in the previous section.
- Timed PWM output, as originally implemented in [10].

Listing 4 shows examples of the first (line 9) and second (line 21) varieties.

### E. Integration of C and XC

Computing the envelope generator coefficient involves floating-point calculations. While XC does not support the float type, an XC function can safely call a C function, as long as the function parameters and return value are XC-supported types (Listing 5). In the `eq_calc_coef` function, parameters are converted from fixed-point values, the calculation is performed, and the result returned as a fixed-point value.

A number of other types, including channel 'ends', arrays and pointers can also be passed between C and XC.

### F. Real-time monitoring

The xTIMEcomposer development tools include xSCOPE, a tool for monitoring states, variables and other conditions in real time [16]. Variables within application code can be registered with 'probes'; once registered, the real-time behaviour of these variables is graphed within xTIMEcomposer's real-time scope display. Listing 6 shows the registration of probes.

Fig. 4 shows a typical reading. Here, the synthesiser's response to two consecutive musical events is visible, in terms of three separate traces:

- The output of the carrier oscillator's EG
- The output of the modulation oscillator's EG
- The output audio samples from the synthesiser

xSCOPE probes can be registered on any number of concurrent processes within the software. Real-time visualisation is of great value. For example, xSCOPE was used during development to verify the output of the sinusoidal oscillator (accurate inversion, reversion and interpolation; variable frequency and amplitude producing the expected output).

### G. Timing analysis

The execution timing of any one process upon an XMOS processing tile is contingent on the activity of the other processes executing on the same tile. For $n$ concurrent processes on a tile, a minimum amount of processing time $m$ is guaranteed. This enables static analysis of XMOS binaries to provide worst-cast execution time (WCET) measurements [17].

As discussed in Section III, the synthesiser has a hard time constraint: whatever configuration is loaded or musical event has been generated, the synthesiser must be capable of generating at least one sample within a 21 $\mu$s interval in order to produce a digital audio signal at the required sample rate. As shown in Listing 3, the synthesiser executes `generate_sample` and outputs the 16-bit return value over the channel. Static analysis of the `generate_sample` function yields a WCET measurement of 4.68 $\mu$s (Fig. 5), satisfying the timing requirement. Static timing analysis of any

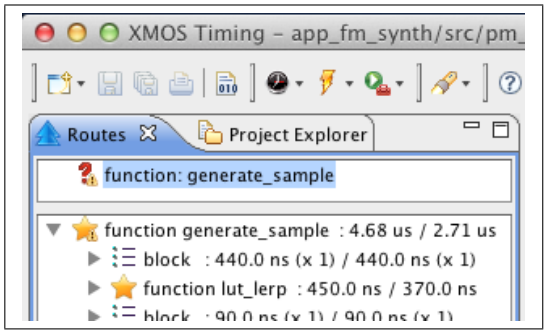



Fig. 5. Timing analysis of the generate_sample function

number of time-sensitive functions can be *scripted* into the build process, so that breaking a hard real-time requirement will also break the build.

## V. Conclusion

This paper has described a concurrent synthesis project that functions as a self-contained electronic musical instrument and as a tutorial for developers new to the XMOS MCU. The implementation of the software follows closely from a standard RTSAD process and the XC language enables clear description of interfaces and interactions between concurrent processes.

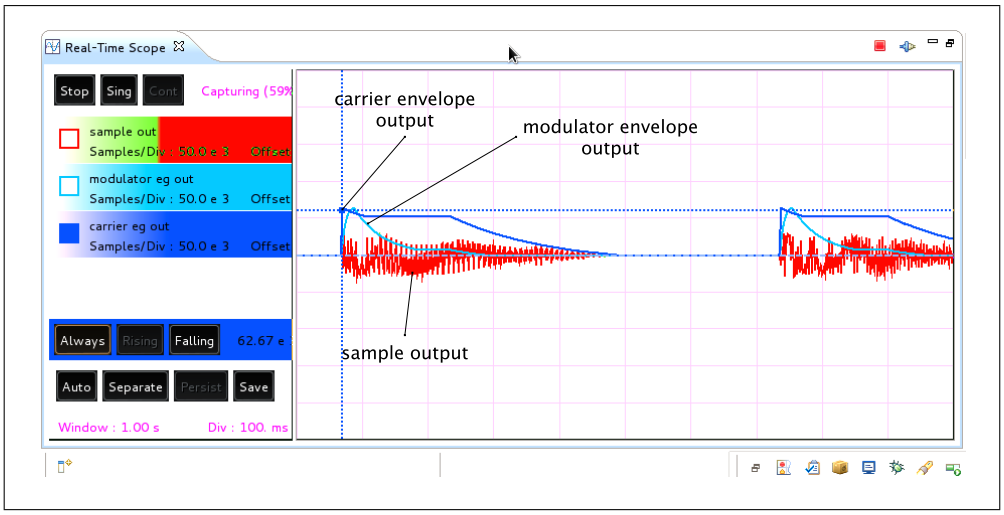



Fig. 4. Real-time monitoring with xSCOPE

The implementation of the project demonstrates some entry-level-to-intermediate XMOS programming and development techniques. All resources to build the application described in this paper have been made freely and publicly available. The project website is located at [18], featuring video documentation.

The HADEDA application demonstrates a number of XMOS programming techniques (message-passing between concurrent processes, integration of standard C code within concurrent processes, timed operations) and tools (real-time monitoring of software variables and worst-case execution time analysis).

This paper has demonstrated that the XMOS approach to message-passing concurrency enables the straightforward design and construction of a software application comprising several components with independent real-time requirements. In a non-multicore embedded system, these real-time requirements would have to be balanced by the developer, making maintenance or alterations to the application a precarious process. Within the XMOS architecture, however, deterministic process scheduling and the opportunity to measure WCETs within a given concurrent configuration provide a robust basis for further expansion and experimentation with the application.

This capability makes XMOS technology a compelling low-cost candidate for rapid prototyping and implementation of concurrent applications with strict real-time requirements.

## References


[1] XMOS. (2014) XMOS startKIT microsite. [Online]. Available: http://www.xmos.com/startkit

[2] T. Hayashi and K. Ohmori, "An autonomous vehicle using a multi-thread and event-driven processor," in *Integrated Circuits (ISIC), 2011 13th International Symposium on*, Dec 2011, pp. 305–308.

[3] G. Martins, A. Moses, M. J. Rutherford, and K. P. Valavanis, "Enabling intelligent unmanned vehicles through XMOS technology," *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, vol. 9, no. 1, pp. 71–82, 2012.

[4] M. Marshall, T. Carter, J. Alexander, and S. Subramanian, "Ultra-tangibles: creating movable tangible objects on interactive tables," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 2185–2188.

[5] D. May, "The XMOS architecture and XS1 chips," *IEEE Micro*, vol. 32, no. 6, pp. 0028–37, 2012.

[6] XMOS, *xCORE architecture overview*. XMOS, 2014. [Online]. Available: https://www.xmos.com/download/public/xCORE-Architecture(X9650D).pdf

[7] ——, *XMOS programming guide*. XMOS, 2014. [Online]. Available: https://www.xmos.com/download/public/XMOS-Programming-Guide-(documentation)(B).pdf

[8] ——, *xTIMEcomposer user guide*, 13th ed. XMOS, 2014. [Online]. Available: https://www.xmos.com/download/public/xTIMEcomposer-User-Guide(X3766B).pdf

[9] ——. (2014) GPIO driver for startKIT. [Online]. Available: http://goo.gl/TzC1Bg

[10] (2014) SID emulator on startKIT. [Online]. Available: http://www.xcore.com/projects/sid-emulator-startkit

[11] R. Williams, *Real-Time Systems Development*. Heinemann, 2005.

[12] D. Austin. (n.d.) No static at all: Frequency modulation and music synthesis. [Online]. Available: http://www.ams.org/samplings/feature-column/fcarc-synthesizer

[13] B. Frei, *Digital sound generation part 1: Oscillators*. Institute for computer music and sound technology., 2010. [Online]. Available: http://goo.gl/KxAzNx

[14] National Instruments, *Understanding Direct Digital Synthesis (DDS)*. National Instruments, 2013. [Online]. Available: http://www.ni.com/white-paper/5516/en/pdf

[15] N. Redmon. (2013, June) Envelope generators – ADSR part 1. [Online]. Available: http://www.earlevel.com/main/2013/06/01/envelope-generators/

[16] XMOS, *Use xTIMEcomposer and xSCOPE to trace data in real-time*. XMOS, 2013. [Online]. Available: https://www.xmos.com/download/public/Trace-data-with-XScope(X9923A).pdf

[17] ——, *XMOS timing analyzer whitepaper*, 1st ed. XMOS, 2013. [Online]. Available: https://www.xmos.com/download/public/XMOS-Timing-Analyzer-Whitepaper(1.1).pdf

[18] J. Dibley. (2014, May) HADEDA sequenced synthesiser. [Online]. Available: http://www.xcore.com/projects/hadeda-sequenced-synthesiser



**James Dibley** is a PhD candidate in the Department of Computer Science at Rhodes University. He received his Masters degree in 2014 from Rhodes University.

Listing 1. Declaring parallel processes and communication channels

```
chan c_pwm, c_synth_audio; /* XC channels */
startKIT_led_if i_led;
startKIT_button_if i_button;
slider_if i_slider_x, i_slider_y;
interface pm_oscil_if o1;

par {
on stdcore[0]: startKIT_gpio_driver(i_led, i_button,
     i_slider_x, i_slider_y, gpio_ports);
on stdcore[0]: sequencer(p_metro, i_led, i_button,
     i_slider_x, i_slider_y, o1);
on stdcore[0]: pm_oscil_task(o1, c_synth_audio);
on stdcore[0]: buffer(c_synth_audio, c_pwm);
on stdcore[0]: pwm_server(c_pwm, p_speaker,
     SAMPLE_RATE);
}
```

Listing 2. The synthesiser interface

```
interface pm_oscil_if {
    void sequence_step(unsigned int step);
    void load_oscil_preset(unsigned char preset_id);
    void rt_ctrl(unsigned char ctrl_type, unsigned
        int ctrl_change);
};
```

Listing 3. The synthesiser process is defined in terms of events

```
#define SAMPLE_PERIOD (XS1_TIMER_HZ / SAMPLE_RATE)
void pm_oscil_task(interface pm_oscil_if server
    sq_if, chanend c_synth_audio) {
unsigned int time;
timer t;
struct PM_Oscil o;
initialise_pm_oscil(o);
t :> time;
while(1) {
    select
    {
    case t when timerafter(time) :> int _:
       c_synth_audio <: (short) generate_sample(o);
       time += SAMPLE_PERIOD;
       break;
    case sq_if.sequence_step(unsigned int step):
        unpack_step(o, step);
        break;
    case sq_if.load_oscil_preset(unsigned char
        preset_id):
        configure_pm_oscil(o, preset_id);
        break;
    case sq_if.rt_ctrl(unsigned char ctrl_type,
        unsigned int ctrl_change):
        if (ctrl_type == SLIDER_X) {
            eg_set_decay_rate(o.m_amp_eg,
                ctrl_change);
        } else if (ctrl_type == SLIDER_Y) {
            eg_set_release_rate(o.c_amp_eg,
                ctrl_change);
            eg_set_decay_rate(o.c_amp_eg,
                ctrl_change);
        }
        break;
    }
} }
```

Listing 4. Dynamic timed processing in the sequencer process

```
static const unsigned int sq_p_meta[SQ_COUNT][3];
static const unsigned int sq_p[SQ_COUNT][16];
 . . .
unsigned char bpm = sq_p_meta[sq_preset][
    SQ_TEMPO_BPM];
unsigned char steps_per_b = sq_p_meta[sq_preset][
    SQ_BEAT_LENGTH];
unsigned int i = 0, tempo = XS1_TIMER_HZ / ((bpm   *
     steps_per_b ) / 60);
 . . .
case i_button.changed():
    if (i_button.get_value() == BUTTON_DOWN) {
        sq_preset++;
        if (sq_preset == SQ_COUNT) sq_preset = 0;
        bpm = sq_p_meta[sq_preset][SQ_TEMPO_BPM];
        steps_per_b = sq_p_meta[sq_preset][
            SQ_BEAT_DIV_ST];
        tempo = XS1_TIMER_HZ / ((bpm * steps_per_b )
            / 60);
        i = 0; // reset to 1st step of seq
    };
    break;
 . . .
case t when timerafter(time) :> int _:
    t :> time;
    if (i % sq_p_meta[sq_preset][SQ_BEAT_LENGTH] ==
        0) p_metro <: 1; else p_metro <: 0;
    osc_if.sequence_step(sq_p[sq_preset][i]);
    i++;
    if (i == sq_p_meta[sq_preset][SQ_TOTAL_STEPS]) i
         = 0;
    time += (tempo);
    break;
```

Listing 5. Computing envelope coefficients in a C function

```
#include <math.h>
 . . .
uint32_t eg_calc_coef(uint32_t rate, uint32_t
    target_ratio) {
float f_target_ratio =
    target_ratio * (1.0 / 0xFFFF);
float result = expf(-log((0x1 + f_target_ratio) /
    f_target_ratio) / rate);
uint32_t fixed_result =
    (uint32_t) (result * 0xFFFF);
return fixed_result;
}
```

Listing 6. Registration and use of xSCOPE probes

```
// Executed before main()
void xscope_user_init(void) {
    xscope_register( 3,
        , XSCOPE_CONTINUOUS, "sample␣out",
            XSCOPE_INT, "n"
        , XSCOPE_CONTINUOUS, "modulator␣eg␣out",
            XSCOPE_UINT, "n"
        , XSCOPE_CONTINUOUS, "carrier␣eg␣out",
            XSCOPE_UINT, "n"
}

...
    o.output_sample = (o.output_sample * o.
        c_amplitude) >> OSCIL_M_AMP_SHIFT;
    // Record current value of o.output_sample
    xscope_short(0, o.output_sample);
...
```

# Development of Soundex Algorithm for isiXhosa Language

Zukile Ndyalivana, Z.S.Shibeshi
Department of Computer Science
University of Fort Hare, Private Bag X1314, Alice 5700
Tel: +27 40 602 2746, Fax: +27 40 602 2464
Email: {ZNdyalivana, Z.Shibeshi} @ufh.ac.za

**Abstract- String matching is the common way of finding items from textual database. Because of the way people write some types of text, like names, string matching may not be helpful in text processing, and other mechanisms like phonetic matching needs to be included if we want to develop an efficient matching scheme. Phonetic matching is defined as a method of identifying a set of strings that are likely to be similar in sound to a given keyword. The problem of writing words (e.g. names) differently is a common problem in all languages. As a result there is need of a system which will match terms phonetically regardless of the type of errors introduced. There are many errors or variations that can be considered but we are referring to typographical errors, spelling errors as they differ in vowel and matching of consonants. In this paper, we will present a phonetic algorithm that is developed for isiXhosa language which matches terms written in Xhosa by approximating their meaning based on their sound. The algorithm is developed based on the principles of Soundex algorithm.**

**Index Terms—Xhosa phonetic algorithm, Soundex algorithm, Phonetic matching.**

## I. INTRODUCTION

In any languages there are cases where people write some terms differently. Names are the major sources of problems in this regard. The way people speak and even how they write some terms are different, but their meaning is the same. Phonetic matching deals with the equivalence of two or more terms by pronunciation irrespective of their actual spelling. This means, although terms are spelled differently, they can be matched phonetically. Instance in the Just-Dial service, where a telephone operator gets a name to find specific information is one example where matching terms with sound will be very helpful. The operator estimates the possible spelling of the names pronounced to her and looks from the database for exact match or approximate match.

Another typical example is when searching for a name in a large database, where we would be looking for a name such as Clair, and if there is only one word in the database that is written in different spelling like Clare, then there is a problem with the retrieval of that certain word in that particular application.

As can be seen from the above example, both names (Clair, Clare) sound the same. As a result, we need a method where we can search a term using its sound. The common used technique to solve this kind of problem is to use an algorithm similar to soundex algorithm. Soundex algorithm is an algorithm that indexes terms making use of the sound rather than their spelling. It focuses on homophones to encode them into one representation, so that they match despite their minor differences in spelling.

In trying to solve this problem mentioned above which is also occurring in isiXhosa, we create a phonetic scheme for isiXhosa language that can be used by different applications such as spelling checkers, transliterations and name word searching applications together with information retrieval systems. The complicated and frequently inconsistent rules of Xhosa words were a motivation for bringing together an original phonetic algorithm.

Our initial point of interest was making an extensive analysis of Soundex encoding scheme with some other techniques to check their effectiveness on Xhosa terms. We finally developed phonetic algorithm that works correctly for Xhosa terms. Based on the algorithm we developed a system that compares the similarity of terms and also retrieves related words through a given query. The system also makes use of the Xhosa phonetic algorithm to give related terms for terms that are misspelled.

## II. RELATED WORK

In this section we review the different methods that are used by different researchers for phonetic matching. Modern methods were highly effective in matching terms approximately, in our project we also use such techniques to match terms approximately.

A large number of phonetic matching algorithms are developed for Latin-based languages. Algorithms that use the phonetic characteristics of terms been investigated for English and string equivalence methods have gathered attention because of their communication of self-ruling methodology(Aqeel et al, 1998). In suggesting a proper procedure on how spelling errors can be automatically fixed, the authors in (Li et al, 2011) looked at four most important classifications, which are presented in **Table 1**.

**Table 1: Common spelling errors in English**

| | English | |
|---|---|---|
| Type of error | Baseline name | Deviation |
| Insertion | Fisher | Fischer |
| Omission | Johnston | Johnson |
| Substitution | Katherine | Catherine |
| Transposition | Hagler | Halger |

We also investigated similar spelling errors in the Xhosa language, which are presented in **Table 2**.

**Table 2**: **Common spelling errors in isiXhosa**

| | Xhosa | |
|---|---|---|
| Type of error | Baseline name | Deviation |
| Insertion | Zintle(beautiful) | Zinthle |
| Omission | Chuma(prosper) | Cuma |
| Substitution | Isipaji(wallet) | isipatshi |
| Transposition | Bhuti(brother) | buthi |

As can be seen from the above table, even though the words are spelled differently they sound the same and we can recommend the correct spelling using algorithms like Soundex. In Soundex systems algorithm numbers are assigned to characters and names constituting the different characters that are assigned the numbers will be concatenated to have one number for the word so that names with similar sounds will get the same number. The numbers are referred to as soundex encodings. A soundex based searching application will not directly search for an exact name directly but will search for the soundex encoding. By so doing, it will get all names that sound like the name being sought. The following sections present the different soundex algorithms developed for English.

### *A. Russell Soundex*

This was the first soundex algorithm created as explained in (Aqeel et al, 1998). Its encodings are produced by allocating numbers to single letters in the name, and using the same number for letters that are closely related (e.g. m and n). As soon as the sequence of numbers gets a definite length, no more letters in the name are looked at. This restricts the name-matching to only the first portion of the term. **Table 3** presents Soundex encoding scheme that is used to assign to characters by this algorithm.

**Table 3: English soundex codes**

| **Soundex Coding Scheme** | **Characters** | **categories** |
|---|---|---|
| 0 | b, f, p, v | Labial |
| 1 | c, g, j, k, q, s, x, z | Guttural and sibilants |
| 2 | d, t | Dental |
| 3 | l | Long liquid |
| 4 | m, n | Nasal |
| 5 | r | Short liquid |
| 6 | b, f, p, v | Labial |

### *B. Metaphone Algorithm*

In 1990 the metaphone algorithm was published in an article relating to a more progressive soundex system the authors named it **Metaphone** (Lawrence Phillips, 1990). This system tries to fundamentally yield its encoding based on how a term is pronounced rather than its spelling, and it is merely based on the English pronunciation, therefore removing spelling inconsistency. Corresponding to Daitch-Mokotoff algorithm it makes use of groups of letters rather than just single letters. In contrast all its predecessors, this encoding technique is based on the complete name rather than shortening after noting only some first portion of the term. It did not walk on the footsteps of the algorithm of Daitch-Mokotoff's lead of letting a name to have extra encoding, but as an alternative by creating one encoding for each name as was completed in the previous systems.

The code makes use of the following 16 consonant symbols "0 B F H J K L M N P R S T W X Y". The '0' stands for "th", 'X' represents "sh" or "ch" and the other characters stand for their usual English pronunciations. The vowels "A E I O U" are used only at the start of the code. The algorithm is known for its long list of rules

- The following is the summary of this encoding scheme. 1. Delete duplicate adjacent letters, except for C.
- If the word starts with 'KN', 'GN', 'PN', 'AE', 'WR', drop the first letter.
- Delete 'B' if it comes after 'M' and if it is at the end of the word.
- 'C' should be changed to 'X' if is followed by 'IA' or 'H' (unless in latter case, it is part of '-SCH-', where it should be changed to 'K'). 'C' also need to be changed to 'S' if is followed by 'I', 'E', or 'Y'. Otherwise, 'C' changes to 'K'.
- 'D' is transformed into 'J' if followed by 'GE', 'GY', or 'GI'. Otherwise, 'D' transforms to 'T'.
- Delete 'G' if accompanied by 'H' and 'H' and is not at the end or before a vowel. Drop 'G' if accompanied by 'N' or 'NED' and is at the end.

- 'G' changes to 'J' if it comes before 'I', 'E', or 'Y', and it is not in the form 'GG'. Otherwise, 'G' changes to 'K'. Reduce 'GG' to 'G'.
- Delete 'H' if after vowel and not before a vowel.
- 'CK' changes to 'K'.
- 'PH' changes to 'F'.
- 'Q' changes to 'K'.
- 'S' changes to 'X' if accompanied by 'H', 'IO', or 'IA'.
- 'T' changes to 'X' if accompanied by 'IA' or 'IO'. 'TH' changes to '0'. Drop 'T' if accompanied by 'CH'.
- 'V' changes to 'F'.
- 'WH' transforms to 'W' if at the beginning. Drop 'W' if not followed by a vowel.
- 'X' transforms to 'S' if at the beginning. Otherwise, 'X' transforms to 'KS'.
- Drop 'Y' if not followed by a vowel.
- 'Z' transforms to 'S'.
- Drop all vowels unless it is the beginning.
- This algorithm uses a more complex collection of rules to provide more accurate phonetic comparisons.

### *C. Double metaphone*

The authors in (Lawrence Phillips, 2009) developed yet another algorithm related to their original metaphone algorithm, which they named it as Double Metaphone. The name originates from the fact that the algorithm creates two encodings for a single name. It does not have the strength of Daitch-Mokotoff algorithm of having ability of holding more encodings. One new feature of the Double Metaphone algorithm is that it has included foreign pronunciations, but it is not exhaustive in including all the foreign rules and it does not differentiate which rule relates to which particular language. The author (Lawrence Phillips, 2009) dropped one of the improvements of their previous Metaphone – to be precise the encoding is again constrained to the initial part of the term.

### *D. N-gram Analysis Technique*

N-grams are n-letter sub sequences of terms or strings where n typically is one, two or three (Mishra et al, 2013). One letter n-grams are denoted to as unigrams or monograms; two letter n-grams are mentioned to as bi-grams and three letter n-grams as trigrams. Over-all, n-gram detection technique works by observing each n-gram as an input string and looking it up in a precompiled table of n-gram strings to determine its existence or its frequency of terms or strings that are found to contain nonexistence or highly infrequent n-grams are recognized as misspelling.

### *E. Algorithm Comparison*

In the methods detailed above, it can be seen that the soundex algorithm was the most suitable for implementation of different algorithms because most of the described algorithms provided a bit of differing approaches with some kind of complexity, The original soundex, which is the Russell Soundex name-matching procedure is an often used method and has been changed easily for suitable use with other languages different from the English language. We also found out that it is most suitable to implement the original Russell Soundex phonetic algorithm instead of the more extended specialized ones.

## III. ISIXHOSA LANGUAGE

Xhosa belongs to the category of the Bantu language family tree, and is most popular throughout the south of the equator of Africa (M. Pascoe et al, 2012). The term 'Bantu' was studied in language studies since the beginning of the "1857"(Niesler et al, 2005). The Nguni, which is one of five language sets which includes the South-Eastern Bantu languages, consists of a classification of sibling languages such as Zulu, Ndebele and Swazi.

### *A. The vowel inventory in the Xhosa language*

The language has about seven distinct vowels, but in writing system these vowels are represented as five with the following symbols a, e, i, o, u. Symbols such as e and o just represent two vowels each .

### *B. The consonant inventory of Xhosa language*

The Xhosa language is composed of a library of consonants, which are the pulmonic aggressive sounds (same as those found in English), the velaric sounds (i.e. clicks), and the glottic sounds which are the (implosive) (Sand et al, 2007). Xhosa has got also sounds such as the velar nasal, and also contains the huge number of nasal clicks, which can therefore be allophonic with the phonemes next to them. It has also a massive inventory of stop consonants, which are called voiced, voiceless, and aspirated, including the series of stops which are only three. **Table 4** shows various clicks that exist in isiXhosa language.

**Table 4: Xhosa clicks**

| | Dental | Lateral | Alveo-palatal |
|---|---|---|---|
| Oral | Voiceless | c | x |
| | Aspirated | ch | xh |
| | Voiced | gc | gx |
| Pre-nasalized | Voiceless | Nc/nkc | Nx/nkx |

A pronunciation clicks in **Table 4** displays that clicks every time stands out from the sounds that exist on either side of them. The authors in specified that pronunciation and word origin can have an influence on spelling. In fact, as spelling is very observable, it becomes an alternative for literacy, more especially for the general public. The fact that isiXhosa has so many clicks poses a greater challenge in spelling of terms as mentioned in (Gxilishe, 2004).

Owing to the fact that many clicks include considerable deal of acoustic energy. Co-articulation with the next vowel is hard to attain (Gxilishe, 2004). Clicks are said to be in all likelihood the greatest salient consonants established in a human language. They are practically never disoriented with non-click letters. The Xhosa language has about seven manners of articulation and seven places of articulation, these are:

- Plosives (implosive and explosive)
- Fricatives
- Affricates
- Nasals
- Laterals
- Semi-vowels
- Trill(roll)

The combination of plosive and fricative form an affricate like ts, dz, and tsh there are also seven places of articulation. In **table 5** we display at the codes allocated to isiXhosa language.

**Table 5: isiXhosa soundex scheme**

| CODE | NUMBER | LETTER | NAME |
|---|---|---|---|
| **A** | 0 | a,e,i,o,u,w,h | Short & semi vowel |
| **B** | 1 | b,p ,bh,ph, Mb ,mp, f, v, Mf,mv | Bilabials, voiced &voiceless & Labio-dentals |
| **C** | 2 | M,n,mh,nh | Nasals (plain & aspirated) |
| **D** | 3 | t,th,d, dh | Alveolar |
| **E** | 4 | s,z,nz,dz,nts | Fricative and affricates |
| **F** | 5 | sh,tsh,ntsh, y,ny,nty,ndy,j | Pre & mid Palatal ,semi vowel and nasal explosive |
| **G** | 6 | k,kh,kr,gr,g ,nk,ng | Velar |
| **H** | 7 | hl,dl,ntl,ndl | Laterals |
| **I** | 8 | l,r | glottal and a rough guttural |
| **J** | 9 | q,qh,gq,nq, nkq,ngq | Palatal, lateral |
| | | c,ch,gc,nc,nkc,ncx,xh,gx,nx,nkx,ngx | & alveolar clicks |
| **A** | 0 | a,e,i,o,u,w,h | Fricative and affricates |

*C. Proposed Xhosa-Phonetic Algorithm*

In designing the soundex algorithm method for IsiXhosa, XSOUNDEX, we started with the method carried out in the original Soundex of combining identical consonants with similar sounding property. This is the approach of the soundex algorithm for isiXhosa language; the main idea is to come up with an algorithm to encode isiXhosa terms based on their sound. For this phonetic structure, we are assigning the code as shown in the (**Table 5)** so that strings can match phonetically**.**

Grouping the combinational forms such as 'sh' and 'tsh', we can combine the characters by adding another category to the encoding scheme, since some of the terms are mostly found in isiXhosa language. The presence of clicks also requires a different encoding for them. On the other hand, the presence of h in the language is more than crucial as it gives more meaning to most words, unlike Soundex for English as well as Phonix both certainly disregard the letter 'h' for the fact that its phonetic properties is not clearly pronounced as well also tend to alter in some cases such as in forms and combinations as 'sh' and 'ch'. We also group all the vowels to their separate encoding which is zero. Each and every segmental phoneme is therefore classified as stated to their segmental characteristics. While analyzing the phonetics of Xhosa letters, it was found out that some of the letters had to be given attention. There are twelve click consonants, which are C, CH, NC, GC, Q, QH, and NQ, GQ, X, XH, NX, and GX which need to be considered separately.

Consequently we have come up with the following algorithm:

- Pre-processing includes capitalizing all the letters of the string.
- Keep the first character of the string
- Remove all of the vowels unless they are at the beginning
- Iterate through the length of the string from left to right.
- Remove all of the vowels unless they are at the beginning.
- If two or more characters with identical numbers were next to each other in the original name (before step 1), then omit all but the first.
- Keep the maximum length of the code
- Return the Xsoundex code.

## IV. RESULTS

The algorithm was tested using a variety of Xhosa words and names to check if it can accurately match for any style of writing of a word in Xhosa, which will be equivalent phonetically. We developed a database and entered words

together with their code. We then designed a system where users will enter a term, and then request the system to search for similar words. The systems take the term, code it using the algorithm (XSOUNDEX) and match with codes of other words that are already in the database, and then display similar words. Accordingly we assess the effectiveness of the system by reviewing the results together with linguists. We started our experiment with a list of words that people tend to confuse with one another. **Table 6** presents these words.

**Table 6: List of terms with incorrect and correct spellings**

| Baseline Names/words (in Xhosa) | Deviation(In Xhosa) | Meaning in English |
|---|---|---|
| Zintle | Zinhle | Person name(beautiful) |
| edoropini | Edolophini | A town |
| isandla | Ihlanza | A hand |
| ikhomputha | Ikomputa,ikompiyutha | computer |
| intombazana | Tombazana, | A girl |
| iyandiphazamisa | uyandipazamisa | disturbing |
| intloko | inhloko | Head |
| ingangane | Inkankane/ing'ang'ane | Bird |
| inkciyo | inciyo | Animal skin |
| ukugromba | Ukugrhumba | dig |
| isisu | isusu | stomach |
| isiporho | isipokro | ghost |

The Xsoundex codes for these correct/incorrect words are presented in **Table 7**.

**Table 7: Xhosa words with Xsoundex code**

| **Words/Names** | | **Codes** |
|---|---|---|
| Zintle | Zinhle | Z248 |
| edoropini | Edolophini | E38172 |
| isandla | Ihlanza | I4238 |
| ikhomputha | Ikomputa,ikompiyutha | I672137 |
| intombazana | Tombazana, | I23214 |
| iyandiphazamisa | uyandipazamisa | U2317 |
| intloko | inhloko | I2487, I287 |
| ingangane | Inkankane/ing'ang'ane | I272 |
| inkciyo | inciyo | I2796,I296 |
| ukugromba | Ukugrhumba | U7782, U77882* |
| isisu | isusu | I55 |
| isiporho | isipokro | I528, I5278 |

Recall and precision are two methods of evaluating systems of this kind. **Table 8** presents the recall and precision of the system we developed. The XAverage is the Average precision and recall for the Xhosa terms.

**Table 8: List of Xhosa words with Xsoundex scheme displaying precision and recall**

| Analyzed word in Xsoundex | Precision =(P) C/B | Recall =R C/A | Correct Matches (A) | Retrieved Matches (B) | Correct +Retrieved (C) |
|---|---|---|---|---|---|
| edoropini | 0.51 | 0.32 | 0.80 | 0.33 | 0.432 |
| intombazana | 0.63 | 0.22 | 0.50 | 0.61 | 0.11 |
| ingangane | 0.78 | 0.70 | 0.54 | 0.55 | 0.93 |
| ukugromba | 0.43 | 0.34 | 0.23 | 0.54 | 0.77 |
| intloko | 0.55 | 0.46 | 0.43 | 0.63 | 0.60 |
| | XAverage=0.58=58 % | XAverage=0.41=41 % | | | |

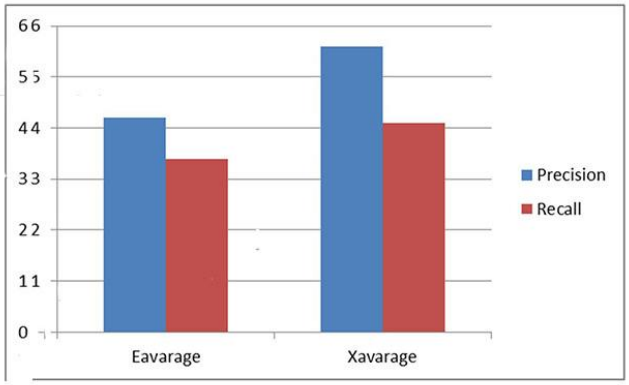
**Table 9: List of Xhosa words with English soundex displaying precision and recall**

| Analyzed words in English soundex | Precision =(P) C/B | Recall= R C/A | Correct Matches (A) | Retrieved Matches (B) | Correct +Retrieved (C) |
|---|---|---|---|---|---|
| edoropini | 0.51 | 0.32 | 0.50 | 0.30 | 0.32 |
| intombazana | 0.43 | 0.22 | 0.40 | 0.50 | 0.11 |
| ingangane | 0.28 | 0.01 | 0.34 | 0.55 | 0.36 |
| ukugromba | 0.43 | 0.34 | 0.73 | 0.45 | 0.20 |
| intloko | 0.45 | 0.56 | 0.43 | 0.37 | 0.60 |
| | Eavarage 0.42=42% | Eavarage 0.31=31% | | | |

**Table 9** shows the recall and precision of the Xhosa words

**Graph 1** shows the average (recall and precision) of Xhosa words using English soundex and average of Xhosa words using Xsoundex. The average for English soundex is Eaverage and for Xhosa is Xavarage.

**Graph 1: English and Xhosa soundex averages of the different Xhosa words.**




## V. DISSCUSION

Our results based on the words that were taken from list provided by several people who speak the language and also the staff of the linguistic department and the names were entered on the Xhosa soundex algorithm, the results show that our algorithm works well, but the authors noted that for some words the algorithm tends to duplicate some numbers which is due to the fact that in Xhosa we do not have a definite number of consonants but combinations of letters mostly, although our algorithm caters for most words for that but it has difficulty in differentiating some complex combinations like nkx.

Retrieval was performed by finding the Soundex encoding (or Xhosa soundexing) of each query and retrieving each and every name indexed by the Soundex encoding or Xhosa soundex. Our algorithm (Xhosa soundex) really shows that based on the information provided in (**Graph 1)**. From (**Graph 1**) the soundex for isiXhosa is doing well in terms of precision and a little more for the recall. From the previous discussions we can concluded the following. Firstly it is not possible to obtain a matching algorithm that is competent into locating all exact matching word that you use for query.

Xhosa there are very strange words which we believe would be very difficult to find a consensus on how they are written. We asked different students one particular word that has different spelling to write it for us and we found the following variations: “inkankane, ingangane, ingkangkane” given by different students but still the word is incorrectly spelled. This word can be written many times but will never have an exact spelling.

## VI. CONCLUSION AND FUTURE WORK

In this paper the authors have presented a phonetic algorithm that they proposed for isiXhosa which has worked well. Significant work has been done and portrayed broadly in the area for English and a few related other languages (e.g. for Oriya, Arabic, and other languages) but there is very little work done in Xhosa language. As a future work, the author plan to design and carry out a research to develop a plugin that can be used with spell checkers to help in matching words of Xhosa language that may be written differently. We also suggest for an extensive research to be done on the language so that different rules can be included to increase the accuracy of the algorithm presented in this paper.

## REFERENCES


Aqeel, S. U., Beitzel, S., Jensen, E., Frieder, O., & Grossman, D. (1998). On the Development of Name Search Techniques for Arabic وجاهت صابر احمد حميد ديميمد دما رباسد تاهجو, (0991).

Gxilishe, S. (2004). THE ACQUISITION OF CLICKS BY XHOSA-SPEAKING CHILDREN, *20*(2), 1–12.

Li, D., & Peng, D. (2011). Spelling correction for Chinese language based on pinyin-soundex algorithm. In *2011 I*nternational Conference on Internet Technology and Applications, iTAP 2011 - Proceedings.

Mishra, R., Kaur, N., & Technique, A. D. L. (2013). A Survey of Spelling Error Detection and Correction Techniques, *4*, 372–374.

Niesler, T., Louw, P., & Roux, J. (2005). Phonetic analysis of Afrikaans, English, Xhosa and Zulu using South African speech databases. Southern African Linguistics and Applied Language Studies, 23(4), 459–474. doi:10.2989/16073610509486401

Sands, B., Brugman, J., Exter, M., Namaseb, L., & Miller, A. (2007). ARTICULATORY CHARACTERISTICS OF ANTERIOR CLICK CLOSURES IN N ǀ UU, (August), 401–404.

Lawrence Phillips, Hanging on the Metaphone, computer Language, 7(12), 1990.

Lawrence Phillips, the Double Metaphone Search Algorithm, C/C++ Users Journal, 18(6), June, 2000.M.

Pascoe and M. Smouse. “*Masithethe: Speech and language development and difficulties in isiXhosa*”. SAMJ: South African Medical Journal, 2012, Vol. 102(6), pp 469-471.

Zobel, J. and Dart, P. [1996]. Fnetik: An integrated system for phonetic matching. Technical Report 96-6, Department of Computer Science, RMIT



**Zukile Ndyalivana has** received his Honors degree in 2014 from the University of Fort Hare and is presently studying towards his Master of Science degree at the same institution. His research interests include Natural Language Processing (NLP), Social Media, Web development and cloud computing.

# CORE NETWORK TECHNOLOGIES

# Implementation of EPC Mobile Networks using NFV and SDN

Joyce Mwangama, Neco Ventura
University of Cape Town,
Department of Electrical Engineering
Cape Town, South Africa
{joycebm, neco}@crg.ee.uct.ac.za

***Abstract*—For a mobile network operator, the mobile broadband revolution with its rapidly increasing traffic volumes has resulted in a number of challenges. Customers want ubiquitous network coverage, high bandwidths, and reliable services at reasonable prices. At the same time, investors and owners require constant efficiency improvements, reduced operational costs, and higher profits. Being able to increase the capacity of networks to be able to cope with exponential increases in network bandwidth demand by investing in new and more efficiently operating infrastructure becomes quite vital for mobile network operators. This work investigates the infrastructure sharing aspect for network operators looking to deploy the LTE/EPC network. Virtualization and Software-Defined Networking allows networks to operate at optimal and efficient levels, and serves as a good building block for networks of the future. In this paper, we describe the design, implementation and evaluation of a cloud computing and virtualization solution for next generation mobile network sharing among operators.**

***Keywords – Evolved Packet System, NFV, OpenFlow, Open vSwitch, SDN.***

## I. INTRODUCTION

The current trend within the mobile networking environment sees the evolution of the radio and core network architectures [1]. This new network design aims to achieve high-bandwidth availability for services and applications of future mobile networks. It is designed to enable truly mobile broadband services and applications and to ensure a smooth experience for both operators and end-users. The Evolved Packet System (EPS), which comprises the LTE radio technology and the EPC core network architecture, is the reference implementation for future mobile operators looking to deploy fourth generation and beyond mobile networks. Thus, the survival of mobile network operators lies in their ability to leverage the need to increase network capacity, while handling the ever increasing, but variable demand for network bandwidth. An important challenge for mobile network operators will be to meet the predicted increase in mobile traffic. To meet this demand, operators will need to invest heavily in the CAPEX and OPEX of their network infrastructure [2]. This creates situations such that the full ownership model, where a single network operator owns the entire infrastructure in their network, becomes too expensive to maintain. As a natural consequence, a new business model has arisen, seeing network operators sharing common network infrastructure to reduce expenditures. They can do this by investing in the adoption of virtualization and cloud computing techniques to deploy their network functions and resources.

Virtualization is a technique where abstract versions of physical resources are created. In the context of network infrastructure, virtualization is the creation of a virtual version of a physical resource (network, router, switch or server) based on an abstract model that is often achieved by partitioning (slicing) and/or aggregation [3]. The task of virtualization of commercial mobile networks, deployed by network operators either on a regional or national scale, requires a complex undertaking. This approach, however, has many advantages over the previous approaches as end-to-end connection management remains in the full control of the network operator, while security and separation is achieved over the shared infrastructure.

Along these very same lines comes another new idea that has recently received substantial attention in the research community: Software-Defined Networking (SDN). This goes hand in hand with the need to decouple hardware and software components in the networks, that is, the shift from hardware-centric network components and elements to software defined radio and SDN utilizing general purpose hardware [4]. Within the context of these ideas, although a huge amount of research activities has been spent on future network technologies, very few research results on next generation networks are available in the for future mobile networks. The key issues that need to be investigated can then be summed up as follows:

- How can network operators maximize profits using EPS infrastructure while minimizing costs to be incurred in deploying physical infrastructure. How can the provision of infrastructure sharing, optimal resource utilization, isolation and performance guarantees be achieved in LTE/EPC deployments?
- How can the scalability and elasticity of network deployments become easier to achieve and implement without increasing complexity in the mobile network architecture?
- What tools can facilitate on demand instantiation and tear down of network resources to ensure that over provisioning does not negatively affect the budget of the network?

The contribution of our work could be summarized as follows:

1. In this paper, we propose the use of a solution that allows for the deployment of an EPS mobile network in the cloud.
2. We propose to use network and resource virtualization to achieve a cost effective yet performance enhancing infrastructure deployment solution for mobile network operators.
3. We present the architecture for virtual mobile networks and furthermore present an initial implementation of EPC/LTE which is built over OpenFlow and Open vSwitch tools.
4. We evaluate the network function of the cloud network vs. a physical network prototype implementation

The remainder of this paper is structured as follows. Section II introduces the motivation and scenarios where this work is applicable in the context of the EPC architecture. Section III provides an overview of related works and literature. Section IV introduces the framework architecture. Section V discusses the implementation tools such as the network testbed as well as OpenFlow and Open vSwitch, the tools used to provide the SDN and network virtualization capabilities, and provides an evaluation of the framework. Section VI concludes the paper and presents the future work.

## II. SCENARIO AND REQUIREMENTS

### A. *Need for New Architectures*

Traditional network deployments are disadvantaged because of the static nature of the network infrastructure. The majority of currently deployed networking equipment is highly specialized and monolithic. It is envisioned that in the future, network operators will opt for dynamic control and management capabilities that allow for rapid creation and deployment of services and/or resources. This will be extremely difficult to achieve as the status of most currently deployed networking equipment is highly specialized and sometimes vendor locked networking [5]. This creates huge incentives to implement networking functions on generalized equipment and opting for software-defined solutions for greater flexibility and control.

Current mobile networks operate the data and control layers on the same plane but this is increasingly not efficient as data traffic might scale differently from control traffic. This has remained unchanged for many years. Additionally, network complexity continues to increase as new technologies, hardware and services are introduced. Scalability will become an important issue as bandwidth speeds and traffic rates increase. It will become more expensive to evolve mobile networks towards EPS architectures, and if network operators originating from the same markets don't consider to do this cooperatively they can miss out on CAPEX and OPEX reducing opportunities. Thus a need for network sharing that does not undermine the operation of each operator is needed.

In current network infrastructures elasticity is difficult to achieve, and thus another drawback resulting is the difficulty to archive network sharing in mobile networks. One of the biggest influences to the cost of network is the physical equipment. By sharing physical resources between several mobile operators it is possible to achieve greater efficiency of existing resources, fewer site builds and broader coverage [5]. While standardized and vendor solutions exist for network sharing mobile operators will seek greater flexibility and dynamicity from network sharing solutions than those that are currently available. For example, the ability to create and destroy (offer and revoke) on demand networking resources for mobile network operators will be useful for network sharing [2]. Dynamic and flexible sharing of the infrastructure would be required in this case on timescales smaller than the current times needed for contract agreements and implementation [6].

### B. *Benefits of Software Defined Networking?*

Software-defined Networking solutions have been realized for implementations of data centers and enterprise networks, as the most promising technology to facilitate network virtualization. SDN allows centralized, programmable control planes and data planes abstraction, where control and data are separated, so that network operators can control and manage directly their own virtualized resources and networks without recognizing detailed hardware technologies. This separation allows control to be directly programmable and manageable in a centralized manner and data planes to be simplified and abstracted rather than relying on specialized hardware.

There are many benefits for going in the direction of SDN; some of these are presented below:

- SDN facilitates and allows for the separation of the control plane from the data plane
- SDN dramatically reduces the complexity of operation and management in networks
- SDN creates network programmability capabilities where there is direct control or orchestration of the behavior of thousands of routers and switches
- SDN creates the opportunity for the development of emerging service-aware networking
- SDN can work in conjunction with network virtualization where the creation and provision of Infrastructure-as-a-Service (IaaS) is possible. This is the case where infrastructure elements, logical and physical, can be offered as on-demand resources
- Similarly, virtual switches and Network-as-a-Service (NaaS) can be exploited where network entities or functions are virtualized
- SDN is designed for high network performance and scalability
- SDN is enabled with the use of open interfaces between the devices in the control plane (controllers) and those in the data plane

### C. *Requirements in Future Network Deployments*

The main focus of this work is to facilitate network sharing. This inherently brings up some key considerations and functional requirements needed within the network deployments. The 3GPP technical specification on service requirements for network sharing highlights some crucial requirements to be considered [7]. User requirements entail that network sharing bear no noticeable difference to the user, in terms of service quality, from non-network sharing deployments. Requirements placed on the network specify that service capabilities should not be restricted by network-sharing scenarios. Network operators should be able to differentiate service offerings from other operators in a shared network. This also requires that service continuity,

handover and roaming not be diminished in a shared network.

Additionally, for the deployment of virtualized networks the following additional issues arise. How is high reliability/availability/network quality sustained within the shared network? Maintaining secure isolation of operators, in terms of each individual network's traffic and their related customers' traffic that is traversing through the shared infrastructure is of high importance. The security and privacy level among the sharing network operators must ensure that no opportunity for information leaks or control takeover by any operator can be allowed

## III. BACKGROUND AND RELATED WORK

Many researchers are active in this area of future network technologies. Network Virtualization is one of the most promising techniques to enable innovation in the network [8] and is seen as a solution for evolving the current limitation of networking functionalities mentioned above [9]. With network virtualization, it is possible to have multiple virtual networks running simultaneously on top of a shared physical infrastructure. Network management with virtualization support, however, introduces challenges that need to be addressed in order to fully achieve effective and reliable networking environments. This need has contributed to the numerous investigations of novel management solutions.

Similarly, mobile networks stand to gain many benefits by evolving towards the "software-defined" paradigm where we see control and data planes being completely separated. One of the first work investigating this is from [10]. In this paper, the authors describe an evolution of the mobile Evolved Packet Core Network (EPC) utilizing SDN that allows the EPC control plane to be moved into a data center. This work notes that while mobile networks already deploy a considerable amount of software control, the ability of OpenFlow to decouple the control and data planes for IP routing provides the opportunity to simplify the configuration and maintenance of mobile aggregation networks by eliminating the distributed IP routing control plane. The authors extend the OpenFlow 1.2 protocol with two vendor extensions, one defining virtual ports to allow encapsulation and decapsulation and another to allow flow routing using the GTP Tunnel Endpoint Identifier (TEID). The result enables an architecture where the GTP control plane can be lifted up out of network elements then run a simplified OpenFlow control plane, enhanced with GTP TEID routing extensions. The GTP protocol implementation, which has proven to be a popular protocol especially when providing roaming capabilities across different network operators, greatly benefits from this extension. This work represents the first steps of future mobile network deployment and highlights an early stage reference implementation.

In [2], the authors investigate the concept of network function virtualization. The motivation for this work, as in ours, is to identify the driving business models and corresponding network sharing functionalities needed for mobile network deployment. This work however focuses on the case of radio access network (RAN) sharing for 3GPP networks. This work also analyses the feasibility and business viability of this proposed solution for LTE practical deployments.

In [11], the authors discuss the necessary steps for the migration from today's residential network model to a converged access/aggregation platform based on SDN and OpenFlow. One of the steps highlighted is the integration of SDN into LTE/4G mobile networks, at the base stations, femto cells, eNodeBs, and Serving Gateway/PDN Gateways, as well as the integration of OpenFlow with the MME's session control and mobility management functionalities. Work in [4] discusses how to achieve a successful carrier grade network with SDN, based on the Open Networking Foundation (ONF) definition of SDN. In this article, SDN is presented as a viable solution for simplifying and unifying network management and provisioning across multiple domains. The authors attempt to analyze along four axes of issues they perceive to be the main challenges of SDN: performance vs. flexibility; scalability; security; and interoperability. They propose possible solutions on how a programmable switch may be achieved; how to enable the controller to provide a global network view; how an SDN may be protected from malicious attacks; and how SDN solutions may be integrated into existing solutions. Conclusions are drawn on the significant issues that must be addressed in order to meet the carrier's expectations, and how SDN is likely to pave the way for highly optimizes ubiquitous service architectures. Other than the identification of the required steps forward, no mention is made on how to achieve this nor is a prototype implementation presented for analysis.

Currently only one paper showcases a prototype implementation of a 3GPP mobile network, incorporating the concepts of SDN. In [12] the authors present their definition of a "software-defined mobile network" (SDMN) based on a software-driven forwarding substrate which enables on-demand creation of any type of mobile network and opens the network to innovation through new service enablers without mandating any change to mobile terminals. This work presents an on-demand mobile network (ODMN) prototype, which was developed to demonstrate the core benefits of the proposed SDMN approach.

Although a huge amount of private and public research projects have been spent on future network technologies, very few research results on next generation networks are available in the literature. Similar to [12], this paper attempts to add to the state of the art in the public peer-reviewed SDN literature for mobile network deployments with the added focus of network sharing and emphasis of virtualization of functions within the mobile network. We are also then able to showcase the proof of concept prototype implementation of our architecture. The research objectives are noted as follows:

- The Incorporation of SDN in an LTE/EPC mobile network
- Implementation of virtual network functional entities
- Cater for the requirements of the network sharing scenario
- Perform a performance evaluation, comparing for example to regular implementation non SDN/virtualized networks, to identify the feasibility, benefits and/or drawbacks in a practical deployment scenario

## IV. ARCHITECTURE

### A. Proposed Architecture

The main goal of our work is to facilitate in a core network-sharing solution for multiple network operators, ensuring service delivery to operators' respective users. We extend the normal architecture of the EPC network to facilitate for this with the following outcomes:

- It should extend the scope of network sharing such that end-to-end sharing is achieved in the EPC network.
- It should result in the isolation and separation of operators in both the data and control plane with the use of virtualization, to split the core network nodes and interconnecting interfaces.
- It should ensure privacy and security of operators and their users.
- It should be flexible in accommodating operators with different services tailor suited for their customers' needs.
- It should be able to provide end-to-end QoS connectivity to all users/services.

The proposed architecture can be seen in the figure below.

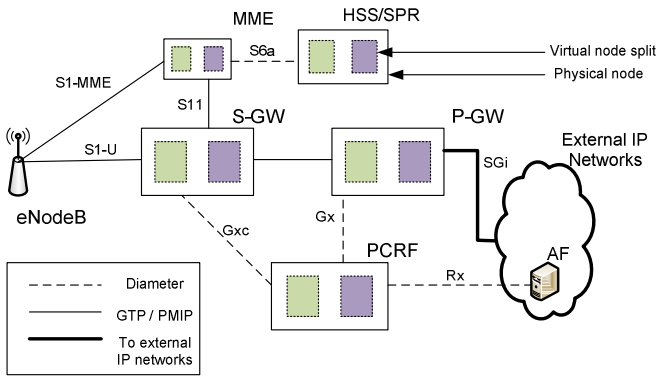



Figure 1. The Proposed Architecture

#### 1) Core Network Splitting by Virtualization

It is envisioned that the physical resources of the network are shared by all the network operators to allow for end-to-end sharing. This is achieved by providing a virtual split at network nodes and linking interfaces. This ensures that full network separation is achieved as well as privacy and security among the network operators.

#### 2) Subscription Management and Service Provisioning

Each individual operator entity should maintain a virtual Subscription Profile Repository (SPR) / Home Subscriber Server (HSS) that contains subscription information. This leads to each network operator being responsible for their own user subscription database implementations. Each operator will provide access to different Packet Data Networks (PDNs) with different services. For example, one PDN could be the public Internet. If a user establishes a PDN connection to this 'Internet PDN', the user can browse websites on the Internet or access other services available on the Internet. Another PDN could be a specific IP network set up by the telecom operator to provide operator specific services, for example based on the IP Multimedia Subsystem (IMS).

#### 3) Evolved Packet Core

The EPC aims to provide a smooth evolution of past and present network technologies towards a common core. This results in seamless mobility between the different generations of mobile access network technologies [1]. Other benefits include the support of different kinds of service types; advanced privacy and security options; and reduction of network complexity in the new flat IP architecture. In the network sharing case, the data and control planes are also separated. Figure 2 illustrates the updated network view for the Evolved Packet Core Architecture.

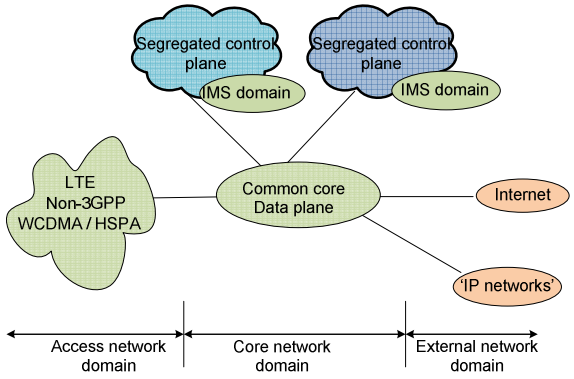



Figure 2. : High level Conceptualization on Mobile Networks

## V. IMPLEMENTATION AND VERIFICATION

### A. Implementation Tools

The following section details the tools used to develop the practical network testbed implemented based on the proposed design architecture detailed in section IV.

#### 1) OpenEPC playground

The FOKUS OpenEPC [10] platform is a non-open source EPC platform enables research, development and testing of protocols and applications on a live testbed. The OpenEPC is conformant to 3GPP specifications (Release 8). This platform was chosen because of its high performance capabilities, adaptability to different deployment scenarios and configurability.

#### 2) OpenFlow

The software defined networking enabled by OpenFlow protocol has recently been proposed as a control framework that supports programmability of network protocols and functionalities by decoupling the data plane and the control plane in networking equipment [1]. In addition to being a unified control plane candidate, OpenFlow provides an abstraction of the network forwarding path, which enables network slicing (partitioning).

OpenFlow architecture splits apart the control and data planes in network switching and routing equipment by moving control to a separate, centralized controller. The controller communicates with the switching and routing network equipment using the OpenFlow protocol to program the switches with flow specification that control the routes of packets through the network. The switches only need to run an OpenFlow control plane, this considerably simplifies their implementation.

*3) Open vSwitch*

In data centers, hypervisors need the ability to bridge traffic between virtual machines and with the outside world. Open vSwitch is targeted at these multi-server virtualization deployments. To cater for on demand resource creation and destruction in virtual environments, Open vSwitch supports a number of features that allow a network control system to respond and adapt as the environments changes. Open vSwitch supports a network state database (OVSDB) that supports remote triggers. Therefore, a piece of orchestration software can "watch" various aspect of the network and respond if/when they change. This is used heavily today, for example, to respond to and track VM migrations.

Open vSwitch also supports OpenFlow, as a method of exporting remote access to control traffic. Open vSwitch allows for control path to be able to control a pure software implementation or a hardware switch. The goal with Open vSwitch is to keep the in-kernel code as small as possible (as is necessary for performance) and to re-use existing subsystems when applicable.

## B. Experimental Evaluation

In this section, we experimentally evaluate the performance of our virtual mobile network deployment first by determining if proof of concept has been achieved in the implementation, secondly we perform comparative tests to determine the performance of our solution versus basic EPC/LTE implementation on physical hardware without integration of SDN.

*1) Experimental Setup*

The performance of individual procedures can be better analyzed if the following proposed changes are incorporated into the testbed in stages and evaluated at each increment; hence three scenarios are introduced for the test procedures. The first scenario is the reference case that implements the standard OpenEPC reference implementation. The second scenario details the reference implementation of the OpenEPC in a bottle. These first two scenarios are out of the box implementations with no added functionality other than the fact that in the first case, each core network entity is house in a physical personal computer, and in the second case, each core network entity is housed in a VMware virtual machine. The final scenario details the implementation of the OpenEPC with distributed elements (Client, eNodeB) residing on physical machine whereas the core network elements (SGW/MME, PGW, IMS, PCC) residing on KVM virtual machines. The switches interconnecting the different subnets of the core network are implemented as Open vSwitches that connect to the OpenFlow controller running on the physical host machine implementing the virtual core network. This final scenario, the focus of our work, is depicted in figure 3.

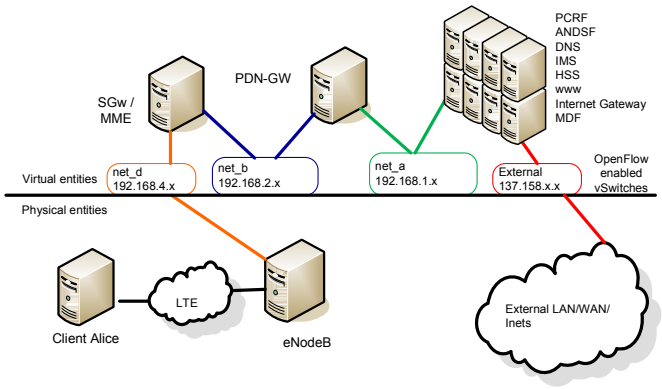



Figure 3. Testbed Layout

The hardware set for each evaluation scenario is described in full in the table below. In all scenarios core elements are hosted on personal computer (PC) machines

TABLE I. SPECIFICATIONS FOR PHYSICAL AND VIRTUAL MACHINES.

| | Physical PCs | EPC in-a-bottle physical host | VMware VMs | KVM VMs |
|---|---|---|---|---|
| Processor | Intel (R) Core TM i3-2120 | Intel(R) Core(TM) i7-3612 | Intel(R) Core(TM) i7-3612 | Intel(R) Core(TM) i7-3612 |
| CPU (GHz) | 3.3 | 2.10 | 2.10 | 2.10 |
| RAM (kB) | 4194304 | 8038284 | 507696 | 507696 |
| OS (Ubuntu) | 12.04.2 Precise | 12.10 Quantal | 12.04.2 Precise | 12.04.2 Precise |
| OS Kernel | 3.5.0-23 | 3.5.0-37 | 3.5.0-23 | 3.5.0-23 |

*2) Proof of Concept Tests*

The first test was to determine if the virtual network entities and the physical network entities could communicate and function as per normal function (i.e. when all the entities are either virtualized or physically implemented). In all scenarios, the user equipment was able to successfully attach to the EPC network, perform IMS registration and video delivery.

*3) Performance evaluation*

The realization of the testbed implementation has demonstrated that proof of concept has been achieved. Further testing included the evaluation of the proposed concepts in the practical environment, where performance can be studied. The next step was to determine the effect of network function virtualization on normal user traffic. For this we attempted to measure maximum throughput data rate on user's connection with the network and compare this with the first two scenario's performance .The results in show that the first two scenarios achieve similar average connection throughputs. Whereas the third scenario, where the VMs are deployed and interconnected with OpenFlow enabled vSwitches, higher throughput is achieved.

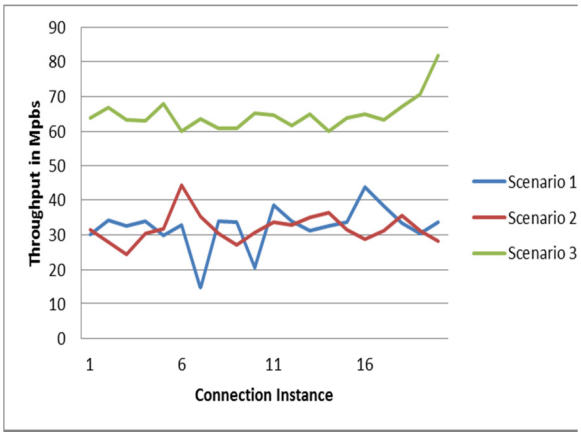



Figure 4. Comparative Throughput as per connection instant

*4) Discussions*

In section II, requirements for network sharing were presented. From the prototype implementation we fulfill the requirement that the user connection experiences better performance than the reference cases, which satisfies the requirement that connections bare no noticeable difference. The network operator is able to still offer and differentiate services as in the reference implementation. This provides us with the basis to further develop and perform tests relating to other issues. Not yet tested is the implementation of more than one network operator, or connections will be affected with user performs handover or is roaming. Scalability, reliability and security are also not yet provisioned for and tested however this can be expected in future works.

In the throughput test, scenario 3 outperformed the first two scenarios. This can be attributed to the fact that Open vSwitch is kept "in-kernel" code, which boosts the forwarding performance compared to normal, non OpenFlow switching. This further highlights the benefits of separating control and data functions from networking equipment.

## VI. CONCLUSION

Virtualization of mobile carrier networks is an exciting and beneficial way for network operators to handle the forecasted "data explosion" while at the same time increasing the return on investment in CAPEX and OPEX infrastructure costs. In this paper, an implementation of a mobile virtual mobile network is proposed. This implementation leverages the high cost of special purpose network infrastructure by deploying these on general purpose virtual network functions. This opens the door for infrastructure sharing, that achieves optimal resource utilization, isolation and performance guarantees be achieved in LTE/EPC deployments. By taking advantage of the tools offered by the OpenFlow network protocol and the virtual Open vSwitch the network scalability and elasticity becomes easier to achieve without modifying or increasing complexity in the mobile network architecture. These tools can facilitate on demand instantiation and tear down of network resources to ensure that over provisioning does not negatively affect the budget of the network. Future work includes the study of two or more virtual core mobile networks over the same physical infrastructure.

## ACKNOWLEDGEMENTS

This research is supported by Telkom South Africa, Jasco / TeleSciences, and the Department of Trade and Industry / National Research Foundation / Technology and Human Resources Programme (DTI/NFR/THRIP).

## REFERENCES

[1] M. Olsson, S. Sultan, S. Rommer, L. Frid and C. Mulligan, SAE and the Evolved Packet Core, Elsevier Ltd, 2009.

[2] X. Costa-Perez, J. Swetina, T. Guo, R. Mahindra and S. Rangarajan, "Radio access network virtualization for future mobile carrier networks," *Communications Magazine, IEEE,* vol. 51, no. 7, July 2013.

[3] P. Minoves, O. Frendved, B. Peng, A. Mackarel and D. Wilson, "Virtual CPE: Enhancing CPE's deployment and operations through virtualization," in *ClouComputing Technology and Science (CloudCom), 2012 IEEE 4th International Conference on*, Taipei, Taiwan, 2012.

[4] S. Sezer, S. Scott-Hayward, P. Chouhan, B. Fraser, D. Lake, J. Finnegan, N. Viljoen, M. Miller and N. Rao, "Are we ready for SDN? Implementation challenges for software-defined networks," *Communications Magazine, IEEE,* vol. 51, no. 7, July 2013.

[5] H. Shimonishi and S. Ishii, "Virtualized network infrastructure using OpenFlow," in *Network Operations and Management Symposium Workshops (NOMS Wksps), 2010 IEEE/IFIP* , Osaka, Japan, 2010.

[6] A. Khan, W. Kellerer, K. Kozu and M. Yabusaki, "Network sharing in the next mobile network: TCO reduction, management flexibility, and operational independence," *Communications Magazine, IEEE,* vol. 49, no. 10, pp. 134-142, October 2011.

[7] Third Generation Partnership Project (3GPP), "Service aspects and requirements for network sharing," 2011.

[8] E. Salvadori, R. Corin, A. Broglio and M. Gerola, "Generalizing Virtual Network Topologies in OpenFlow-Based Networks," in *Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE*, Houston, TX, USA, 2011.

[9] R. Esteves, L. Granville and R. Boutaba, "On the management of virtual networks," *Communications Magazine, IEEE,* vol. 51, no. 7, July 2013.

[10] J. Kempf, B. Johansson, S. Pettersson, H. Luning and T. Nilsson, "Moving the mobile Evolved Packet Core to the cloud," in *Wireless and Mobile Computing, Networking and Communications (WiMob), 2012 IEEE 8th International Conference on*, Barcelona, Spain, 2012.

[11] H. Woesner and D. Fritzsche, "SDN and OpenFlow for converged access/aggregation networks," in *Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference (OFC/NFOEC)*, Anaheim, CA, USA, 2012.

[12] K. Pentikousis, Y. Wang and W. Hu, "Mobileflow: Toward software-defined mobile networks," *Communications Magazine, IEEE,* vol. 51, no. 7, 2013.

[13] D. Drutskoy, E. Keller and J. Rexford, "Scalable Network Virtualization in Software-Defined Networks," *Internet Computing, IEEE,* vol. 17, no. 2, pp. 20-27, 27 March 2013.

[14] R. Corin, M. Gerola, R. Riggio, F. De Pellegrini and E. Salvadori, "VeRTIGO: Network Virtualization and Beyond," in *Software Defined Networking (EWSDN), 2012 European Workshop on*, Darmstadt, Germany, 2012.

[15] S. Min, S. Kim, J. Lee, B. Kim, W. Hong and J. Kong, "Implementation of an OpenFlow network virtualization for multi-controller environment,," in *Advanced Communication Technology (ICACT), 2012 14th International Conference on*, PyeongChang, South Korea, 2012.

**Joyce Mwangama** completed her BSc Eng. and MSc Eng. at the University of Cape Town in 2008 and 2011 respectively. She is currently working towards her PhD in Electrical Engineering in the Centre for Broadband Networks at the same institution. She is currently a Research and Teaching assistant within the electrical engineering department at UCT.

# Design of a Network Packet Processing platform

Sean Pennefather and Barry Irwin
Department of Computer Science
Rhodes University
Grahamstown 6140
Email: g10p0016@campus.ru.ac.za, b.irwin@ru.ac.za

***Abstract*—This paper describes the design considerations investigated in the implementation of a prototype embedded network packet processing platform. The purpose of this system is to provide a means for researchers to process, and manipulate network traffic using an embedded standalone hardware platform, with the provision this be soft-configurable and flexible in its functionality. The performance of the Ethernet layer subsystem implemented using XMOS MCU's is investigated. Future applications of this prototype are discussed.**

## I. Introduction

Research into Telecommunications security and Cyber Defence is a growing field, particularly with regards to incident identification and remediation. In order to assist researchers in this area, appropriate testbed environments need to be set up that are capable of simulating network environments. Unavailable to most researchers is an easy way to create non-ideal network configurations, with conditions such as corruption and packet loss, in addition to being able to log and manipulate traffic during tests.

### A. Motivation

Commercial systems such as Packetstorm and Breaking-Point do exist that are capable of simulating such environments but are large, propriety, have limited extensibility, and typically cost in excess of a million Rand. As different approaches to identifying security threats and the techniques to resolve them are developed, the testing environment should be flexible enough to allow for modifications and extensions by the researchers themselves. By implementing this flexibility, researchers can quickly modify the packet processing platform in order to operate in the desired manner for an environment in order to support the research being conducted. This is particularly important for systems employing novel approaches to be tested. A flexibility realised though a soft-programmable hardware platform allows for a rapid turnaround in the test phase of systems development life cycle, along with rapid prototyping. While some of this could be done in software using traditional server platforms researchers often need to introduce additional latency, and have to deal with issues specific to network card drivers. In order to facilitate this flexibility, what is required is a network platform capable of generic network frame processing. This is achieved though a combination of a dedicated hardware module, with the flexibility being brought though the software running on it. The software interface is intended to be editable and available to both researchers and security professionals to allow them to quickly configure or define the functionality of the platform. Researchers modifying the board should not need to be concerned with programming specific components of the platform, and particularly with the hardware interfacing as such details should be suitably abstracted into a Domain Specific Language (DSL).

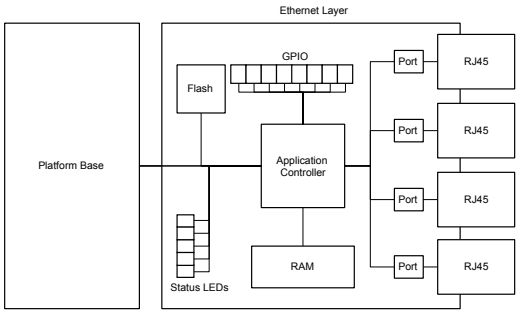


Fig. 1: Hardware Overview

Similar approaches to developing a programmable network packet processing platform have been successfully carried out with prototypes such as the Scalable Programmable Packet Processing Platform (SP4) [1]. The SP4 uses a custom language based on Declarative Networking methodology [2] with high level programming abstractions to provide the functionality to reprogram the system. Our research differs from this approach by developing a language that focuses on general frame manipulation as well as routing and delay withing the platform rather than implementing rule sets.

### B. Structure

The remainder of this paper is structured as follows. Section II explores the hardware design and system components, and subsequent selection of the hardware on which the prototype is implemented. The inter processor communication facilitated by the XMOS platform is discussed in sections III and IV. Performance testing and considerations for the implemented prototype are presented in section V. The DSL developed is covered in section VI. Section VII concludes the paper and presents future development goals.

## II. Hardware Design Overview

As the platform is to operate as an independent device in the research environment, it is necessary to design and develop hardware dedicated to its functionality. To this end, the platform can be broken into two major components; the Ethernet layer and the platform base. Figure 1 shows a general overview of this architecture with the two major components being connected using a Serial Peripheral Interface (SPI) bus, a common protocol for interfacing between a Microcontroller (MCU) and peripherals using

the same data bus. These primary elements are discussed below.

### A. Ethernet Layer

The Ethernet layer is designed to handle network communication that can operate at speeds comparable with 100Mbit/s while applying constraints and modifications to frames transmitted according to the user defined application. Logic not directly relating to frame passing, filtering, and modification should rather be moved to the platform base, allowing this layer to focus on operating with minimal latency (unless requested by user).

Considering these requirements, it was necessary to first investigate potential architectures that could be used to implement this layer. After consideration, the xCORE XS1 and the Xilinx Spartan 6 field-programmable gate array (FPGA) devices were initially investigated further. Both devices were chosen due to their low cost and capabilities to operate at speeds suitable for network communication.

*1) FPGA:* The FPGA architecture allows for integrated circuits to be designed and implemented after the device has been manufactured. This is due to the design of an FPGA consisting of large collection of logic and memory blocks. These blocks are then configured and connected to represent the designed circuit [3]. The Spartan 6 LX25 architecture allows for up to 358 User I/O pins, grouped into 4 I/O banks [4].

*2) XMOS:* The XMOS Xcore XS1 is a microcontroller designed to provide parallel processing as well as determinism and bounded latency for embedded systems. The XS1 provides this functionality by including one or more XCore tiles. Each tile contains up to eight logical processors with their own dedicated register sets. Memory is shared between each logical processor but not across tiles. To allow communication between tiles, the XMOS provides an interconnect to route data and control messages between tiles [5], [6].

*3) Development Toolchain:* Both the Xilinx Spartan 6 and the XMOS XS1 architecture provide functionality that make them capable of achieving the requirements necessary to be suitable for the Ethernet layer. Both architectures can operate at the necessary speed and both architectures come with previously developed software/circuits relating to Ethernet communication. The development tools in both cases are extensive and maintained which limits the motivations of the decision to the actual languages used to develop on each. Both Verilog and VHSIC Hardware Description Language (VHDL) are hardware description languages that are used to describe hardware to be simulated on the FPGA [7]. To generate a flash image from circuits described in these languages propriety software is required such as the tool set provided by Xilinx [7]. Furthermore, experience has revealed that this is not a quick process, nor are the tools relating to this process lightweight or small.

XC on the other hand is the C programming language with mutlicore extensions to take advantage of some functionality offered by the XMOS architecture. More importantly, because it is an extension of the C language, developing a domain-specific language (DSL) which will then translate into XC is simpler than developing a DSL for Verilog or VHDL. Furthermore, the XC tool-chain is light weight and can be implemented on either of the platform base architectures proposed below [8]. Based on the above discussion, the controller selected for implementation of the Ethernet layer of this prototype was the XMOS XS1 microcontroller.

### B. Platform Base

The platform layer provides the base, and user interface for the resulting device. This layer handles all the interaction with other devices such as a PC and storage devices as well as compiling the user applications and generating the appropriate flash images. Retrieving or adding data to the platform such as interrupt commands and log printouts is done through this layer. Due to the scale and complexity this layer would introduce to the development of the platform, it has been decided to rather implement this layer using an existing device. Initial investigations identified two possible candidates; the BeagleBone Black and the Raspberry Pi. Both of these have existing, established Linux operating system support.

*1) BeagleBone:* The BeagleBone white was initially considered as well but decided against due to cost limitations imposed on the resulting platform. The BeagleBone Black is an embedded platform designed by Gerald Coley and manufactured by Circuitco LLC. It is a board developed as open source to allow for the development of systems and applications using a ARM Cortex A8 based processor [9]. While the BeagleBone Black provides a superior offering in terms of hardware it does so at a greater price. Furthermore the BeagleBone Black has very limited availability due to a limited number being produced and high demand.

*2) Raspberry Pi:* The alternative is the Raspberry Pi which has a lower cost and is more readily available than the BeagleBone Black. Initial investigations determined that the speeds attainable using the GPIO interface available on this device was limited to the KHz range (and could potentially be improved to approximately 2 MHz). However, the Raspberry Pi uses a BCM2835 Broadcom processor which allows for the ARM peripherals to be mapped into the kernel virtual address space. This allows for the pins to operate in the region of 20MHz toggle speeds by addressing the pin registers directly [10]. Furthermore, having direct access to the pin registers allows us to set the pins in different modes such as SPI. The Raspberry Pi uses a 700MHz Broadcom ARM processor with 512MB RAM and a separate graphics processor. The board contains two host USB ports and an 100MBit/s Ethernet port [11].

The Raspberry Pi is cheaper but supports all the requirements to function as the Platform base. Access to the processor peripherals enables the setup of two dedicated SPI buses between the Raspberry Pi and the Ethernet layer. As a result, the latter is being designed to use the Raspberry Pi as the base platform.

## III. XMOS Communication

To test the feasibility of using the XMOS architecture for passing network frames between ports, the throughput and latency of the connections between components needs to be tested. As stated in the architecture overview, the Ethernet

layer will consist of multiple XMOS devices, one for each network port, and a larger XMOS L16 to execute the user application. The XMOS devices dedicated to each network port will be minimal in size, currently the XS1-L4 is the smallest device provided by XMOS and only contains four logical cores. This however is suitable to act as a dedicated network port which has been designed to operate on three logical cores, leaving one core free for additional processing.

As each XMOS tile is segmented into multiple logical processors, a core feature of this architecture is to allow applications to execute multiple concurrent tasks in parallel. To allow these individual tasks to communicate during execution, each tile within the device supports a communication infrastructure called a switch to allow data and control messages to be passed between each processor [6].

This communication infrastructure is represented programmaticly as a channel which can be declared and passed to concurrent tasks prior to execution. Inter-tile communication can then happen over a peripheral interface called an XMOS link. Currently, there are two types of links that the switch can use to communicate over; the serial XMOS link and the fast XMOS link. The serial XMOS link is also referred to as a two wire link and consists of four wires allowing for fully duplex operation with a theoretical throughput of 160Mbit/s. The XMOS fast link is made up of 10 wires and has a theoretical throughput of 400Mbit/s. In reality, achieving speeds in excess of 100Mbit/s in the serial XMOS link interface and 250Mbit/s on the fast XMOS link interface is difficult due to physical limitations of the connections and additional data being transmitted [12].

To manage communications between connections, each XMOS device contains a switch which handles the routing of channels relating to that tile. The switch supports up to four intratile channels and up to eight intertile channels, all duplex [12]. The actual number of links present on a device however is dependent on the type package used. In multi-tile packages, at least one fast XMOS link is reserved for communication within the package per tile while on smaller packages such as the XS1-L4 [13], only two links are available due to the number of pins in the package.

Unfortunately, due to the number of pins required to communicate between the XS1-L4 and the PHY, communication to the larger XMOS device must occur over a serial XMOS link. Thus it was important to insure that it is possible to achieve throughput over the serial XOMS link comparable with the transmission speeds of the PHY.

## IV. Inter-tile Delay

To account for wire impedance and track length between the communicating tiles, the developer can set delays which are associated with the link between two tiles. As discussed in the XN Specification document by XMOS [14], a link delay has two values associated with it. The fist value represents the intertile delay value while the second represents intratile delay. These delays indicate the number of cycles the switch should delay between signal transmissions so as to maintain a reliable link between the two tiles. Restrictions on these values are that the delays must integer values and cannot be negative. Furthermore the intertile delay must be less or equal to the intratile delay. During

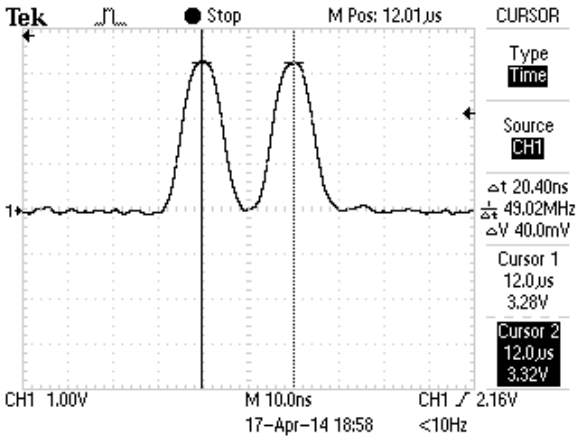


Fig. 2: Pin toggle Speed

hardware design, it is important to minimize the length of wires used as a link so that the specified delay can be set as low as possible while still being reliable. The number of cycles the switch must delay data transmission by will have a direct effect on the maximum throughput of that link.

To investigate actual speeds observed in the application, an oscilloscope was used to measure the duration of a bulk data transfer over the link. One of the hopes of this research was to show that it would be possible to implement an XMOS network that could handle transmission at 100Mbit/s while using the serial XMOS link rather than the fast XMOS link. The primary motivation behind this is that a serial XMOS link requires less reserved pins and is more readily available on less costly XMOS devices such as the XS1-L4. To investigate this, all tests relating to intertile communication are performed using two XMOS Startkit development boards [15].

The time taken to transmit 1526 bytes between the two Startkits was recorded by using an additional pin which had its value toggled after each bulk transfer was completed. It is important to be aware that the recorded value will also include the time taken to set the value of the pin. Figure 2 graphs the time taken for the pin used to transition between its high and low states twice. From this the duration of a full cycle is measured to be approximately 20ns which implies that toggling the pin will include 20ns of latency in the recorded transfer times.

Two different approaches to communication were tested for passing frames between the Startkit development boards; interfaces and transactions. Both approaches use a channel to send the frame body as a stream of data over the link. To keep tests uniform, the frame to be transmitted is simulated to be an array of 1522 bytes and an unsigned integer to indicate the length. The application used to perform this test continually transmitted a frame in one direction across the link. The receiving Startkit board would set the monitoring pin low when the transfer began and then set it high again as soon as the transfer completed.

Figure 3 graphs the time taken to transmit a single frame over the link using the interface approach. This approach worked by sending a reference to the array to the frame from the source tile to the destination tile over the link. This reference was then used in a *memcpy* function which

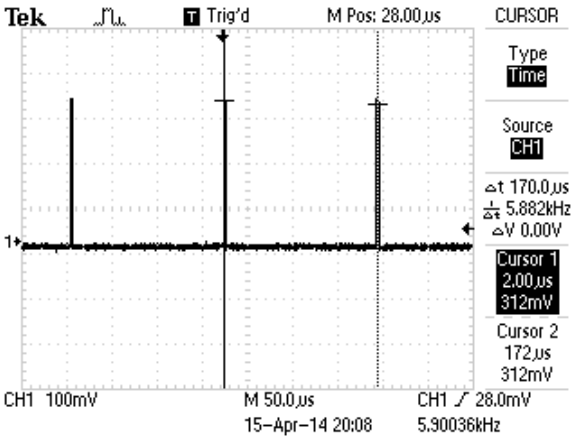



Fig. 3: Pin toggle Speed

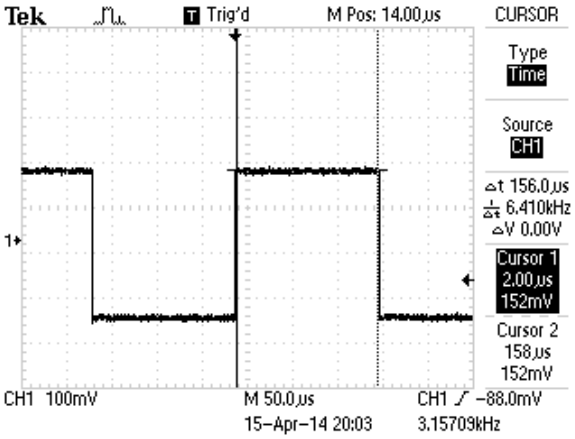



Fig. 4: Pin toggle Speed

handled the passing of the frame content over the link from the remote memory to local memory. This process took 170us which approximates into a transmission speed of 68.3Mbit/s.

The transaction approach works by cycling through the frame data and passing it over the channel in sets of four bytes at a time. This is done in a master slave configuration with the sender being the master and the receiver being the slave. Using a transaction insures that the data is transmitted as a stream rather than a sequence of packets over the link and so cannot be interrupted buy other tasks on the device. The results of using the transaction approach are shown in figure 4 which records the time taken to send a single frame to be approximately 156us. This implies a transmission speed of about 74.44Mbit/s.

Using the transaction approach however means that developer is responsible for insuring that the data relating to the frame is transmitted per attribute rather than simply as a body of data. It is also important to maintain that the receiving component of the transaction is accepting frame attributes in the correct order.

The next component of the transmission to consider is the delay applied to the link between XMOS devices. For both of the above tests, this delay was set to four cycles for both intertile and intratile communication. Modifying this communication to operate at two cycles for intratile communication and three cycles of intertile communication,

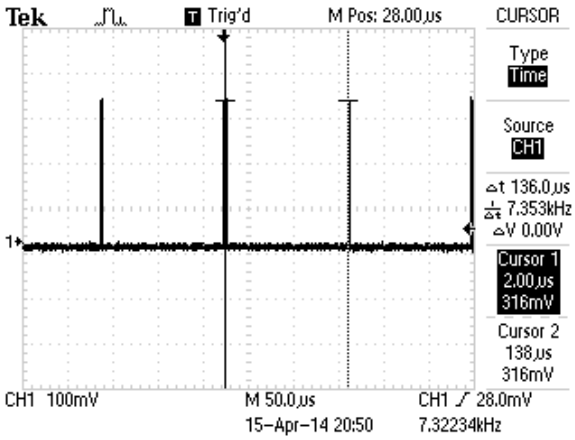



Fig. 5: Pin toggle Speed

it was possible to record a reduced transmission delay for the interface approach. Figure 5 records this reduced delay to be 136us which implies a transmission speed of 85.38Mbit/s. Unfortunately, reducing the link delays caused the transaction communication to break down and it is suspected to be due to the length and quality of the physical connection between the tile switches. Attempting further reductions in delay resulted in the same communication breakdown over the link.

## V. Throughput Analysis

Assuming that we are limited to a two cycle intratile and three cycle intertile delay, then the time taken to transmit a frame of 1522+4 bytes over a link is approximately 0.136ms. Combining this with the calculate intratile delay of 0.274ms, this gives us:

$$0.274[ms] + 0.136[ms] = 0.40[ms]$$

Frame delay over the XMOS network. It is important to note that this delay is not fully duplex as there is additional overhead that comes into play with transmitting and receiving over the same XMOS link simultaneously. Unfortunately, during tests, running the application with an intertile delay of three cycles proved to be too unstable and so the application was switched back to using four cycles. As a result, the revised inter-tile latency becomes:

$$0.274[ms] + 0.170[ms] = 0.44[ms]$$

### A. ICMP

Initial tests of the system were done using a simple ICMP echo request between the two end hosts through the XMOS network. Though not accurate enough to be meaningful, the Round Trip Time (RTT) recorded for these requests was less than 1ms which supports the inter-tile latency of 0.88ms.

### B. TCP

For a more meaningful analysis, TCP throughput can be evaluated to give a better understanding of the bandwidth limitations relating the implemented application. In order to perform this test, an application called iperf[1] was used to seed and transmit TCP traffic through the XMOS network.

[1] http://iperf.fr/

Iperf allows the user to setup a server on a host to accept TCP or UDP traffic and calculate throughput for each stream. A connected host can then run iperf in client mode and specify the host running the iperf server.

This experiment was run on three different platforms: Two XMOS Startkit development boards, an XMOS Slicekit development board [16], and a 100Mbit router. The Startkit boards were connected using a serial XMOS link while the Slicekit board contains a XS1-L16 which contains two tiles connected internally via a fast XMOS link with 1 cycle inter-tile delay [17]. The 100Mbit router used in these tests was the Billion Bipac 7300W [18].

Over a series of tests it was shown that both the Slicekit broad and the router achieve the same throughput at approximately 84.3Mbit/s for a TCP stream. The Startkit configuration however, was only able to achieve an average of 64.3Mbit/s throughput. This shows that the main bottleneck to be the serial XMOS link as expected.

Later we were able to modify the physical connection between the two Startkit development boards such that it became possible to have the serial XMOS link operate with a three cycle inter-tile delay. The above tests were again performed and TCP throughput was recorded by iperf to be 76.5Mbit/s, showing a 13Mbit/s improvement. This implies that by reducing the inter-tile delay from three cycles to two cycles, the serial XMOS link may be suitable for performing frame transmission with a throughput comparable with retail products.

## VI. Domain Specific Language

Programming languages can be grouped according to the level of generality of the applications they are designed to create. For languages such as C, C++, Java and C#, the range of possible implementations is very large. As a result, these languages are all examples of General Purpose Languages (GPL) [19]. A GPL boasts a large set of features allowing for high functionality and application development in a large range of environments. Alternatively, a DSL is more limited in the features available which in turn limits the range of possible implementations of the language [20]. In exchange, a DSL can be optimized to suit a particular domain and becomes very expressive for applications being developed exclusively for that domain [21]. As the goal is to have the researcher develop applications specific to platform, a DSL will be developed.

### A. Specification

The underlying architecture of the platform is XMOS multicore devices which allow for multiple concurrent tasks to be executed in parallel by supporting multiple logical cores per device. To take advantage of this, the developed language should allow for applications developed to be inherently parallel without the user specifying such.

The developed language should also abstract away from all hardware related requirements beyond network port specification. Concerns relating to memory allocation should not the responsibility of the user and communication between logical cores should only be at an abstracted level.

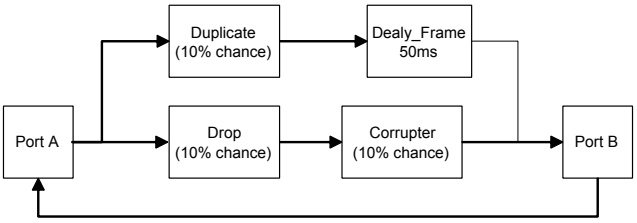


Fig. 6: Data Flow of Application Example

These constraints will allow the user to focus on identifying the requirements of the application and minimize the programming knowledge required to operate the platform.

### B. Application Example

Once completed, the platform will be expected to provide functionality to execute applications that can be of relevant use to the user. In order to show the type of functionality that will be available, we describe an example application that can be developed using the current revision of the DSL and executed on the platform.

To goal of this example application is to introduce errors into a network connection between two hosts. This system will have a probability to either drop, corrupt, or duplicate any packet routed through it. If duplicated, the duplication will be transmitted 50ms after the packed was received. If corrupted, the packet will have a random byte set to 1. Figure 6 describes this scenario as a data flow diagram.

The source code for this application, written in the DSL would be:

```
//-----Declaration Phase-----

Switch Drp //DropSwitch
{
  on Random(10%) : FAIL;
  Default:PASS;
}

Switch Dup //Duplicator
{
  on Random(10%):PASS;
  Default:FAIL;
}

Modifier Cpt //Corruptor
{
  on Random(10%)
  {
    frame_data[rand_range(0,length)] = 1;
  }
  on Random(10%) : Log(length);
}

//-----Linking Phase-----

PortA->Drp<on PASS>->Cpt->PortD;
PortA->Dup<on PASS>->Delay(50ms)->PortD;
PortD->PortA;
```

If the user wished to change the source IP address of

all frames with the IPv4 ethertype to 192.168.123.123, this could be done by simply including the line *'source_ip = 192.168.123.123'* in the modifier link. The DSL allows the user to declare multiple modifier and Filter links which can then be used to build a chain in the linking phase. As the research progresses we intend to add more link types and greater functionality to the DSL.

## VII. Conclusion

This platform prototype is intended to be developed further as a dedicated hardware device to allow for traffic manipulation at the frame level while supporting transmission speeds comparable with retail routing devices. This platform comes with a DSL design to simplify the development of target applications which can be uploaded and run on the device without any specialized programmers or additional hardware. The prototype was found to be successful in verifying the functionality and performance of the selected processors from XMOS and the associated inter processor communication links.

While it is recognized that the prototype described currently lacks full functionality, the intention is to extend the work to offer the following core functionality. Future work on the prototype will include the implementation of a web interface allowing code to be uploaded to the device for compilation. It is our intention to further extend this implementation to support visual representations of the data returned from the Ethernet platform in real time.

We also intend to extend the DSL to allow for more complex applications to be developed and provide better constructs for interacting with frame data. Once extended a series of customizable code modules will be developed to aid developers in quickly building custom applications to run on the platform.

It is also important to note that the above analysis of transmission speeds was performed with minimal overhead. Once further development on the DSL as been performed, further testing will be performed to include user applications. Form this we will be able to determine the performance impact applications could have on platform transmission speeds. It may become apparent that the platform is not reliably capable of transmitting frames at speeds comparable with retail routers and may have to be advertised as operating at a slower speed.

## Acknowledgments

This work was undertaken in the Distributed Multimedia CoE at Rhodes University, with financial support from Telkom SA, Tellabs, Genband, Easttel, Bright Ideas 39, THRIP and NRF SA (TP13070820716). The authors acknowledge that opinions, findings and conclusions or recommendations expressed here are those of the author(s) and that none of the above mentioned sponsors accept liability whatsoever in this regard.

## References

[1] H. Gill, D. Lin, and L. Sarna, "SP4: Scalable Programmable Packet Processing Platform," *Sigcomm*, pp. 75,76, August 2012, accessed on: 24 July 2014.

[2] B. T. Loo, T. Condie, M. Garofalakis, D. E. Gay, J. M. Hellerstein, P. Maniatis, R. Ramakrishnan, T. Roscoe, and I. Stoica, "Declarative networking," *Commun. ACM*, vol. 52, no. 11, pp. 87–95, Novemeber 2009, accessed on: 24 July 2014. [Online]. Available: http://doi.acm.org/10.1145/1592761.1592785

[3] R. Wain, I. Bush, M. Guest, M. Deegan, I. Kozin, and C. Kitchen, "An overview of FPGAs and FPGA programming at Daresbury," CCLRC, Tech. Rep., November 2006, accessed on: 20 February 2014. [Online]. Available: http://www.mecatronica.eesc.usp.br/wiki/upload/b/bd/An_overview_of_FPGAs_and_FPGA_programming.pdf

[4] Xilinx. (2011, October) Spartan-6 Family Overview. Xilinx. Accessed On: 20 February 2014. [Online]. Available: http://www.xilinx.com/support/documentation/data_sheets/ds160.pdf

[5] XMOS. (2012, October) XS1-L01A-TQ128 Datasheet. XMOS. Accessed on :13 February 2014. [Online]. Available: https://www.xmos.com/download/public/XS1-L01A-TQ128-Datasheet(X1154A).pdf

[6] D. May. (2008, July) XMOS XS1 Architecture. XMOS. Acessed on: 10 February 2014. [Online]. Available: http://www.xmos.com/published/xmos-xs1-architecture

[7] Xilinx. (2010, September) ISE In-Depth Tutorial. Xilinx. Accessed on: 21 February 2014. [Online]. Available: http://www.xilinx.com/support/documentation/sw_manuals/xilinx12_4/ise_tutorial_ug695.pdf

[8] XMOS. xTOOLS. XMOS. Accessed on: 14 February 2014. [Online]. Available: https://www.xmos.com/en/download/public/xTIMEcomposer-Flyer(1.0)pdf

[9] G. Coley, *BeagleBone Black System Reference Manual*, a5.2 ed., Circuitco, 1380 Presidential Dr. 100 Richardson, TX 75081 U.S.A., April 2013, accessed on: 18 February 2014. [Online]. Available: http://embeddedcomputer.nl/lanotattachments/download/file/id/3/store/1/

[10] Broadcom, *Boardcom BCM2835 ARM Peripherals* , Broadcom Corporation, Broadcom Europe Ltd. 406 Science Park Milton Road Cambridge CB4 0WW, February 2012, accessed on 14 May 2014. [Online]. Available: http://www.raspberrypi.org/wp-content/uploads/2012/02/BCM2835-ARM-Peripherals.pdf

[11] G. Upton, Eben; Hlfacree, *Rasberry Pi User Guide*. Wiley, 2012, vol. 1, accessed on 18 February 2014. [Online]. Available: http://www.myraspberry-pi.org/wp-content/uploads/2013/02/Raspberry.Pi_.User_.Guide_.pdf

[12] XMOS. (2013, November) xCONNECT Architecture. XMOS. Accessed on: 20 February 2014. [Online]. Available: https://www.xmos.com/en/download/public/xCONNECT-Architecture%281.0%29.pdf

[13] ——, *XS1-L4A-64-TQ48 Datasheet*, XMOS, December 2013, accessed on 14 April 2014. [Online]. Available: https://www.xmos.com/download/public/XS1-L4A-64-TQ48-Datasheet(X2612E).pdf

[14] ——. (2013, November) XN Specication. XMOS. Accessed on 14 April 2014. [Online]. Available: https://www.xmos.com//download/public/XN-Specification(X3944B).pdf

[15] ——. (2013, October) startKIT Hardware Manual. XMOS. Accessed on 15 April 2014. [Online]. Available: https://www.xmos.com/en/download/public/startKIT-Hardware-Manual(1.0).pdf

[16] ——. (2013, November) sliceKIT hardware manual. XMOS. Accessed on 15 April 2014. [Online]. Available: https://www.xmos.com/download/public/sliceKIT-Hardware-Manual(1.0).pdf

[17] ——, *XS1-L16A-128-QF124 Datasheet*, XMOS, December 2013, accessed on 14 April 2014. [Online]. Available: https://www.xmos.com/download/public/XS1-L16A-128-QF124-Datasheet(X8006A).pdf

[18] Billion, *BiPAC 7300W Wireless-N ADSL2+ Firewall Router User Manual* , v2.01.rc1 ed., Billion, April 2010, accessed on: 20 May 2014. [Online]. Available: http://www.seasonstelecom.co.za/documents/billion/Billion%207300W%20Wireless%20Router%20User%20Manual.pdf

[19] J. Bentley, "Programming pearls: Little languages," *Commun. ACM*, vol. 29, no. 8, pp. 711–721, August 1986, accessed on 11 May 2014. [Online]. Available: http://doi.acm.org/10.1145/6424.315691

[20] M. Mernik, J. Heering, and A. M. Sloane, "When and how to develop domain-specific languages," *ACM Comput. Surv.*, vol. 37, no. 4, pp. 316–344, Dec. 2005. [Online]. Available: http://doi.acm.org/10.1145/1118890.1118892

[21] A. van Deursen, P. Klint, and J. Visser, "Domain-specific Languages: An Annotated Bibliography," *SIGPLAN Not.*, vol. 35, no. 6, pp. 26–36, June 2000, accessed on: 11 May 2014. [Online]. Available: http://doi.acm.org/10.1145/352029.352035

## Bio

Sean Pennefather is a Masters candidate in the Department of Computer Science at Rhodes University. He received his Honors degree in 2014 from Rhodes University.

# An Approach to Providing Quality of Service (QoS) for Over the Top (OTT) Voice in LTE Networks

N Nageshar, R Van Olst
School of Electrical and Information Engineering
University of the Witwatersrand, Private Bag 3, WITS, 2050
Johannesburg, South Africa
email: {nikesh.nageshar@gmail.com; rex.vanolst@wits.ac.za}

*Abstract*— **The following paper proposes the initiation of Quality of Service (QoS) for Over the Top (OTT) voice traversing an LTE network using the combination of a voice recognition scheme and QoS bearer initiation procedure. The prevalence of OTT voice in social media applications has resulted in erosion of traditional voice revenues for telecommunication operators. In order to sustain these revenues, the use of QoS becomes a viable option for operators to differentiate themselves in the voice application space. In the solution presented, OTT voice is proposed to be recognised via a Packet Inspection (PI) recognition algorithm, thereafter a QoS bearer is initiated at the Packet Data Network Gateway (P-GW) level interacting with the LTE Policy and Charging Rules Function (PCRF). Results of the tested solution are presented in terms of latency and jitter; illustrating superior network performance metrics for a QoS initiated Bearer as opposed to a non-QoS Bearer.**

*Index Terms*—**IP, LTE, OTT, VoIP**

## I. INTRODUCTION

THE traditional definition of OTT voice is highlighted as the distribution of voice without traversing the switching core of a mobile or fixed line operator; that is voice carried via an application over the public internet or a private network via IP. Voice over IP (VoIP) has an inherent advantage; that being its ability to easily integrate with numerous communication systems. This advantage has been leveraged by the likes of Skype [1] and Google Talk [2]. to provide voice services to social media customers who are in possession of basic internet connectivity.

In the mobile space one of the key catalysts for OTT voice is the increased penetration of smartphones as well as the emergent availability of attractive mobile data plans. Mobile consumers are showing greater interest in OTT voice as these applications have become easier to use, reduces call costs and can be integrated with consumers existing social media usage patterns.

OTT providers offer voice as a free or cheap service via their custom applications, these providers actually attract revenue from opportunities such as advertising, which is contrary to the traditional telecommunications operator revenue model. This loss in revenue has in-turn spurred traditional mobile and fixed operators to broaden their service portfolio into adopting social media applications and OTT voice services to combat the revenue outflow. For this reason operators have taken the strategic direction to explore services such as voice plugins catering for internet communities and social networks [3]. This affords operators the opportunity to compete with OTT providers, yet differentiate their offerings by providing QoS. For an operator to deliver OTT voice each of the following delivery approaches can be considered:

1. Delivery via a user defined application for the enablement of a VoIP session between users of the same application; and
2. Applications with integrated Session based call initiation using protocols such as Session Initiation Protocol (SIP) or H323 / H248 to initiate call flow between users and a centralised platform.

The goal of this paper is to highlight the QoS initiation procedure in the LTE Evolved Packet Core (EPC) [4] [5] for the deployment of a Guaranteed Bit Rate (GBR) Bearer for OTT voice [6]. A Packet Inspection based recognition scheme in conjunction with the LTE Guaranteed Bit Rate (GBR) Bearer initiation procedure is presented illustrating the QoS initiation methodology. The procedure embraces the Policy and Charging Rules Function (PCRF) as central to invoking a Quality of Service Bearer for voice traffic. The method is as follows; subsequent to voice recognition, the Rx interface on the PCRF is manipulated to trigger a GBR Bearer for voice traffic only. A network test-bed was set-up demonstrating the GBR Bearer initiation mechanism as well as the effectiveness of the GBR Bearer in carrying a Real Time Protocol (RTP) voice stream as compared to a Non-Guaranteed Bit Rate (Non-GBR) Bearer.

The paper is set out as follows: Section II provides an introduction to utilising Packet Inspection for voice recognition; Section III illustrates the LTE framework used for the voice GBR Bearer initiation; Section IV presents the experimental design in association with the voice QoS Bearer initiation procedure; in Section V the results of the experimental design are discussed and finally the conclusions are presented in Section VI.

## II. Packet Inspection Approach for Recognition and Admittance of Voice Traffic

Packet Inspection can be enunciated as an advanced method of packet filtering. The use of Packet Inspection makes it possible to find, identify, classify, reroute or block packets with specific data or code payloads that conventional packet filtering cannot detect [7]. With reference to the Open Systems Interconnection (OSI) model, shallow packet inspection inspects headers at layer 3 and ports at layer 4, while deep packet inspection examines headers and payload at layers 4 through 7 [7].

In the work presented a Packet Inspection engine is proposed for the ordered identification and admittance of voice traffic in an LTE network [6]. Using a Packet Inspection Engine, voice can be recognised in the following manner, depending on an equipment manufacturer's choice [7]:

1. IP header on the network layer.
2. Port on the transport layer.
3. Session Initiation Protocol (SIP), Real Time Protocol (RTP) on the session layer.
4. OTT voice application on the application layer.

There are various solutions conceivable to construct a trigger for voice traffic. One option would be for a voice application server to generate a trigger on registration of a voice session from a User Equipment (UE). An alternative innovative solution using PI is to mathematically model a Heuristic Objects and Axioms Algorithm for the recognition of voice traffic as per Nageshar and Van Olst [6]. The model proposed is based on the analysis of voice protocol set-up messages for the successful recognition of voice traffic. This scheme can further be combined with a voice specific variable Admission Control (AC) arrangement in order to provide a combined Heuristic Analysis and AC trigger [8].

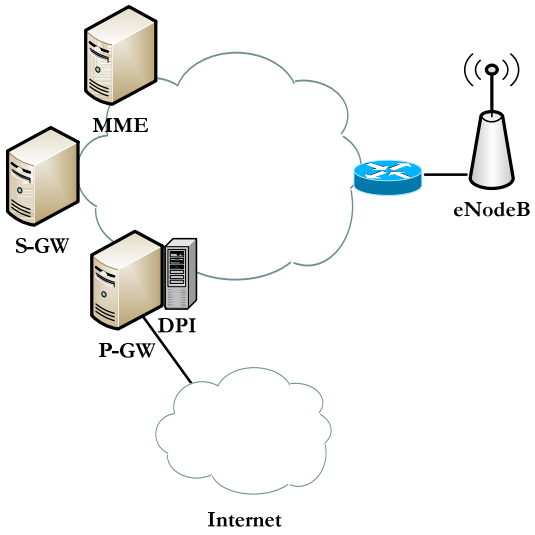



**Figure 1.** Proposed Location of the Packet Inspection Engine

The locality of the Packet Inspection Engine is proposed to be at the P-GW level. Dolgonow et al [9] specify a Packet Inspection Engine on the S-GW as this is a centralised point. All LTE bearers terminate at the P-GW, hence the positioning of the PI engine at the P-GW will result in optimal network expenditure as opposed to a decentralised model and this positioning does not compromise on mobility.

Figure 1 illustrates the addition of a Packet Inspection (PI) engine at the P-GW in relation to the LTE core [4] [5].

## III. LTE Framework for the Voice Guaranteed Bit Rate (GBR) Bearer Initiation

With regard to LTE, a GBR Bearer for QoS can be invoked by means of a Diameter Gx interface terminating between the P-GW and PCRF as illustrated in Figure 2 [10]. The Diameter protocol was derived from the Radius protocol and is the next generation Authentication, Authorization, and Accounting (AAA) protocol.

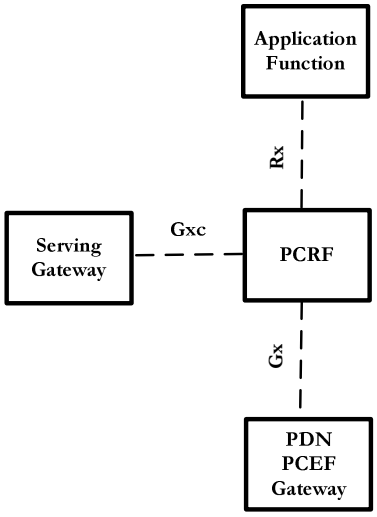



**Figure 2.** LTE Architecture with the Diameter Interfaces Associated [10]

In terms of the voice architecture, the Application Function represents a proxy server that interacts with the session layer or application layer protocol or service. Usually the application layer signalling passes through the Application Function or is terminated in the Application Function. The Application Function extracts session information from the application signalling and provides this to the PCRF over the Rx interface [10].

The PCRF receives session information over Rx interface as well as from the S-GW via the Gxc / Gxa interfaces. The PCRF takes the available information and creates service-session level policy decisions that are provided to the Policy and Charging Enforcement Function (PCEF) and the S-GW [10].

The PCEF is part of the P-GW and is the functional element that encompasses policy enforcement and flow based charging functionalities [10]. The PCEF enforces policy decisions such as maximum bit rate policing that are received from the PCRF and also provides the PCRF with user and access specific information over the Gx interface. The PCEF also performs measurements of user traffic such as traffic volume and / or user session duration [10].

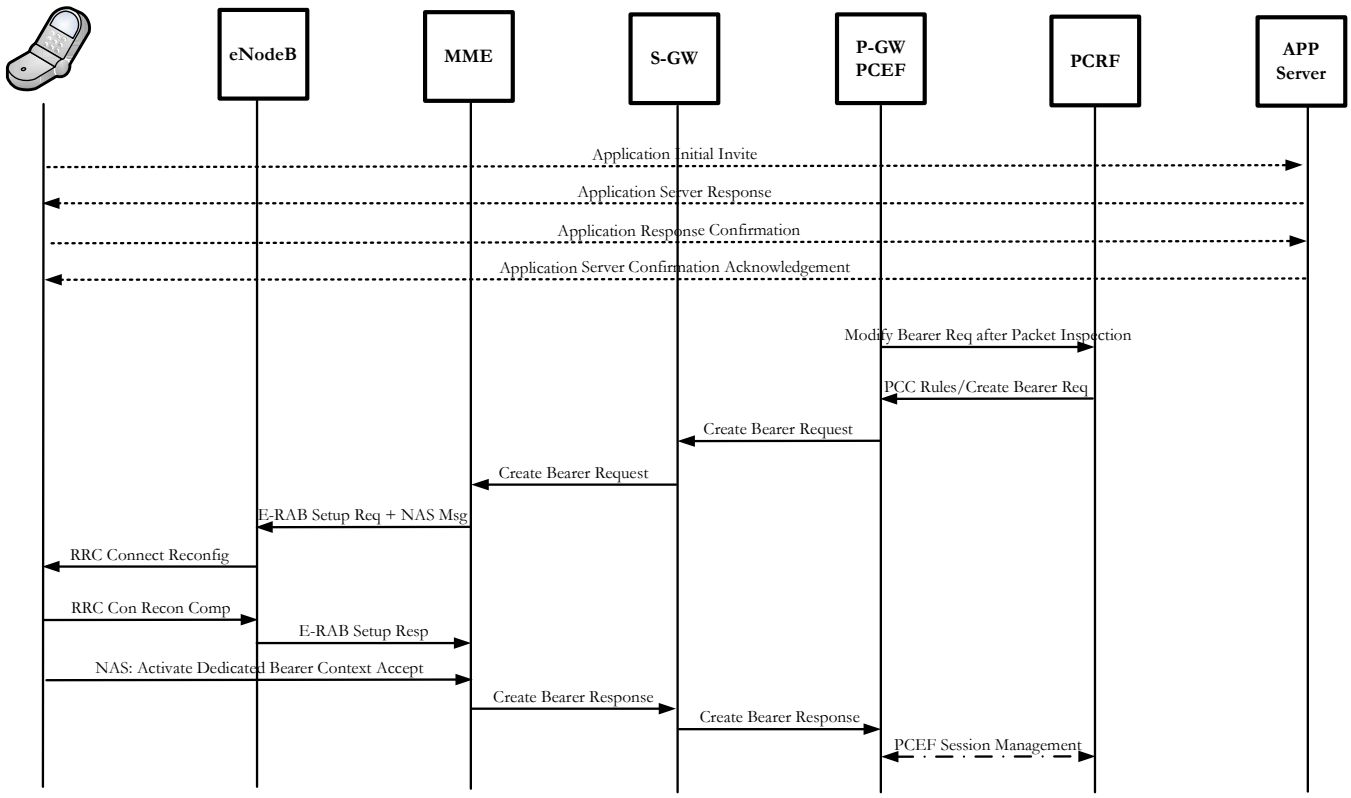



**Figure 3.** P-GW Packet Inspection (PI) Initiated Voice Bearer Setup Procedure

The philosophy depicted was to manipulate the Rx interface on the PCRF to trigger a GBR Bearer via its Gx interface for voice traffic only. On recognition of voice traffic the scheme can generate an attribute value, and then forward this attribute value to the Rx interface of the PCRF to generate a QoS Bearer for a voice traffic service flow.

### *A. Proposed Packet Inspection Voice Bearer Set-up Procedure*

The proposed procedure for voice Bearer setup using a Packet Inspection Engine located at the P-GW is summarised as follows and illustrated in Figure 3 [10] [11]:

1. All traffic flows through the default Bearer onto the P-GW as per the existing LTE attach and Bearer setup procedure [11].
2. In terms of the LTE QoS framework the Policy Charging and Control (PCC) rules for QoS Class Identifier (QCI) are set in the PCRF. The PCRF waits until prompted to download the PCC rules to the PCEF.
3. The Packet Inspection Engine located at P-GW recognises the voice trigger and requests (by forwarding an attribute value via the Rx interface) the necessary QoS resources.
4. The P-GW interacts with the PCRF and pulls down the PCC rules. The PCRF builds policy decisions and sends these QoS rules to the PCEF.
5. The PCEF installs the QoS rules and performs Bearer binding to ensure that the traffic for the voice service will receive appropriate QoS treatment. This will result in the establishment of a new Bearer (Create Bearer Req). The PCC rules enable the PCEF to perform gating, bit rate enforcement and service level charging for the traffic flow.
6. The S-GW processes the Create Bearer Request and forwards to the Mobility Management Entity (MME). The MME sends the EUTRAN Radio Access Bearer (E-RAB) setup request to the eNodeB for Bearer allocation between the eNodeB and the P-GW via the NAS Activate Dedicated EPS Bearer Context Request to the UE.
7. The eNodeB allocates the resources for the Radio Bearers using a Radio Resource Control (RRC) Connection Reconfig Request message to the UE after which the UE establishes the Radio Bearers and responds back with a Radio Resource Control (RRC) Connection Reconfiguration Complete message to the eNodeB. The GBR Bearer is established over the radio interface and the attach Accept is sent to the UE.
8. The voice call is now being transported across the network. The PCEF perform service data flow detection to detect the IP flow for the voice service. The UE utilises uplink packet filters to determine which Bearer shall carry uplink traffic whereas the P-GW forwards downlink traffic over the designated Bearer.

## IV. Experimental Design – LTE Quality of Service Bearer Initiation

The following section describes an experiment that was set-up to demonstrate the ability to successfully invoke a QoS Bearer for voice on an LTE network. The purpose of this experiment was to demonstrate that QoS Bearer initiation is possible, it can be integrated with a Packet Inspection scheme and find a practical solution to demonstrate quality voice over an LTE network without the deployment of composite solution such as the addition of IP Multimedia Subsystem (IMS) [12].

The experimental variables consist of: the PCRF Rx interface; LTE Bearers with QoS Class Identifiers (QCI) and Traffic Flow Templates (TFT) [13] that are associated with the LTE Bearers.

As per the procedure listed above the P-GW interacts with the PCRF after voice is recognised. This request creates a pull down of Policy Charging and Control (PCC) rules from the PCRF. The PCRF sends the Policy Charging and Control (PCC) rules to the Policy Control Enforcement Function (PCEF) for appropriate enforcement.

When a GBR Bearer is initiated; that initiated Bearer is associated with a Traffic Flow Template (TFT) at the eNodeB and the P-GW [13]. This Traffic Flow Template (TFT) consists of a source IP address, destination IP address, source port, destination port and if required a protocol filter applied at the eNodeB and the P-GW [13]. Inherently specified traffic is associated to a Traffic Flow Template (TFT), which is associated with a Quality of Service Class Identifier (QCI), which is then associated with the applicable Radio Link Control (RLC) and Media Access Control (MAC) attributes that overall manage QoS over an LTE Bearer.

### A. *Testing Model and Set-up*

Testing of the LTE Bearer initiation procedure was conducted on an actual LTE system with a system model consisting of the following components:

1. LTE User Equipment (UE).
2. An LTE eNodeB.
3. LTE network management system.
4. A Mobility Management Entity (MME).
5. A Serving Gateway (S-GW).
6. A Packet Data Network Gateway (P-GW).
7. Policy Control Enforcement Function (PCEF) which is part of the Packet Data Network Gateway (P-GW).
8. Policy and Charging Rules Function (PCRF).
9. Connectivity to an Ethernet backhaul network.
10. Laptop connected at the LTE User Equipment (UE).
11. Session Initiation Protocol (SIP) User Agent.
12. Softswitch.

An illustration of the network that was set-up is presented in Figure 4. The system model was configured in the following manner; the eNodeB, S-GW, MME, P-GW and PCRF were connected to each other via a singular Virtual Local Area Network (VLAN) subnet. An Ethernet backhaul network interlinked the EPC components with the Access Network via a fibre connection originating from a switch that fronted the LTE Evolved Packet Core (EPC) components and terminating at the eNodeB. Signalling and traffic paths between the EPC and eNodeBs were separated, with the S1 and X2 interfaces occupying different VLAN subnets. The eNodeB on the Access Network was located about 2Km line of sight from the EPC.

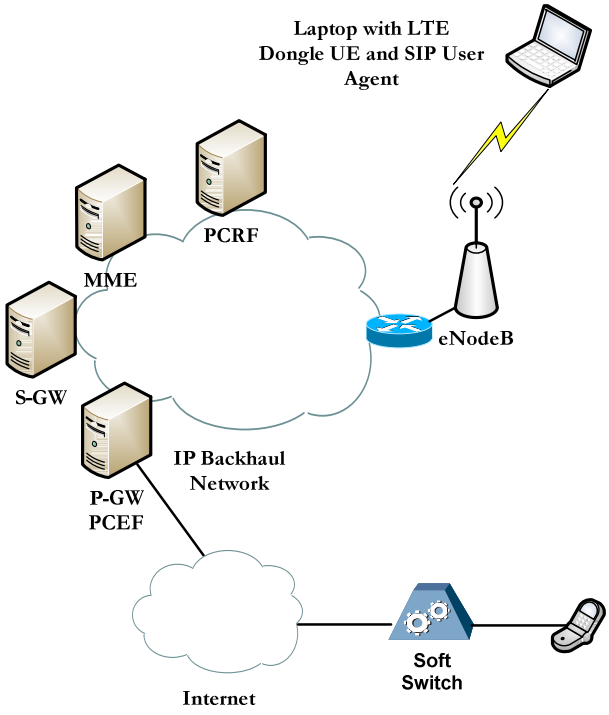



**Figure 4.** Experimental Design Network Set-up

The test laptop which also consisted of the SIP User Agent was connected to the eNodeB via an LTE dongle. An external network connection was provided to the P-GW thereby allowing a direct path from any of the UEs to this external network. The external connection consisted of traffic routing to a Softswitch. The Softswitch enabled voice calls to any Public Switched Telephone Network (PSTN) or SIP Agent reachable.

On the core of the LTE network, the PCRF was configured to consist of a profile that had a Non-GBR Bearer with QoS QCI 9 as well as a profile that had a GBR with QCI 1 for the sole purpose of voice traffic. The Non-GBR Bearer was to be set during the UE attach procedure and the GBR Bearer after the voice recognition trigger. The following was included as part of the procedure for testing:

1. The voice QoS Bearer shall be triggered via an external interface, thereby confirming that the UE Application procedure and Packet Inspection trigger procedure is operational.
2. Only voice traffic shall be allowed to flow through the GBR Bearer and no other traffic.
3. Another component of the testing included proposing the assignment of an attribute value by the voice recognition algorithm to the PCRF via its Rx interface to initiate the GBR Bearer with QCI 1 for voice traffic.

Representing the Packet Inspection initiating element, the Softswich was used as a trigger for the voice GBR Bearer via an external interface of the LTE network. On recognition of the Softswitch IP address an attribute value would be passed through the Rx interface to initiate the GBR Bearer with QCI 1 for voice traffic.

Theory that was also tested included the manipulation of the Traffic Flow Template (TFT) for the voice GBR Bearer such that the filtering mechanism used to classify traffic into the voice GBR Bearer, was allocated only the Softswitch IP during the voice GBR Bearer attach process. This was done by assigning the Softswitch IP as part of the Policy Charging and Control (PCC) rules assigned by the PCRF. When the Policy Charging and Control (PCC) rules were applied to the TFT at the eNodeB and P-GW they possessed a fixed destination IP at the eNodeB and a fixed source IP address at the P-GW that was the Softswitch IP address. This in turn meant that only traffic flowing to and from the Softswitch would traverse the GBR Bearer.

The LTE dongle was set-up with the SIP User Agent registered to the Softswitch. Spurious traffic was simultaneously inserted by the test laptop in conjunction with multiple voice calls being originated on the same laptop. An alternative Softswitch IP address that was then configured on the SIP User Agent to register once more to the Softswitch and multiple voice calls were made as per the previous test scenario. This was done so that a comparison can be made between traffic going to the Softswitch IP that was associated with the voice GBR Bearer and the other Softswitch IP which was associated with the default Bearer.

## V. RESULTS

### *A. Analysis of the Voice Guaranteed Bit Rate Bearer (GBR) Results*

The experiments were conducted in accordance with the experimental design. A trace of the attach message from the MME was used to confirm the initiation of the voice GBR Bearer. The trace confirmed that the GBR Bearer has been successfully created for voice traffic.

On recognition of the Softswitch IP address the PCRF was assigned an attribute value via its Rx interface. The PCRF forwarded Policy Charging and Control (PCC) rules associated with the attribute value which set a modified TFT such that the voice GBR Bearer was created and only traffic destined to the Softswitch traverses the GBR Bearer.

When this GBR Bearer was created with its associated TFT, this TFT consisting of a source IP address, destination IP address, source port and destination port applied at the eNodeB and the P-GW. Inherently traffic associated to the TFT was associated with QCI 1, which was then associated with the applicable RLC and MAC attributes that overall manage QoS of the LTE Bearer.

In line with measuring the results with voice quality network metrics it was deemed necessary to conduct latency and jitter measurement tests based on voice calls made to both Softswitch IP addresses. The analysis of Real Time Protocol (RTP) streams for either Softswitch IP is illustrated in Figure 5 and 6. Average latency and mean jitter results were recorded using the Wireshark Protocol Analyser [14] on both the uplink and downlink directions. Greater than 9 test-sets were recorded with each test-set consisting of hundreds of RTP packets. Individual test results that exceeded 10% of the average of the grouped test-set resulted in that test-set being discarded and the experiment repeated.

### *B. Latency and Jitter Results for Real Time Protocol (RTP) Packets across QoS and non-QoS LTE Bearers*

As is illustrated in Figure 5, the GBR Bearer on average outperformed the Non-GBR Bearer. The results for latency in the upload direction indicated a greater than 20% reduction in latency by the GBR Bearer as compared to the Non-GBR Bearer. The results for the latency in the download direction indicated a marginal reduction of 5-10% for the GBR Bearer over the Non-GBR Bearer.

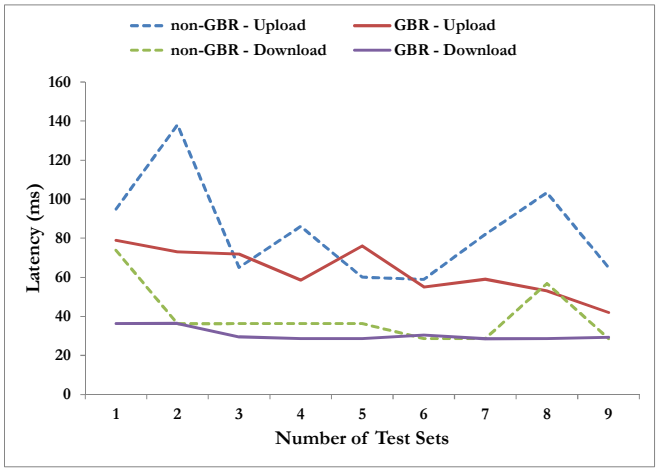



**Figure 5.** Average Latency for calls to a Non-GBR Bearer versus calls to a GBR Bearer

As is illustrated in Figure 6, the results for jitter in the upload direction were significant because this indicated a greater than 100% reduction in jitter for the GBR Bearer as compared to the Non-GBR Bearer. The results for the latency in the download direction were similar. It was initially expected that there be a greater difference in the download test results between the Bearers. Possible reasons for not validating this expectation include that not enough spurious traffic was being injected through the system to significantly affect the latency and jitter on the downlink. It has to be noted that the LTE system that the test-set were being conducted on, was dimensioned on an 82 Mb/s theoretical download limit.

The experiments presented demonstrated that QoS Bearer initiation for voice traffic is possible and available for integration with the Packet Inspection scheme. A practical solution was presented that exhibited quality voice over an LTE network without the addition of complex IMS. The solution provided could be construed as a competitor to IMS. In terms of voice traffic this assumption is correct, however it has to be noted that IMS is a larger framework for multimedia services and not only voice. Of greater importance is that such a solution may have an impact for providers that do not have IMS but have a Softswitch in their network. Such providers do

not need to focus on Circuit Switch Fallback (CSFB) [15], Voice over LTE Generic Access (VoLGA) [3] or similar options because they can now use their native LTE network for voice traffic. The major drawback in this approach is the current lack of handsets in the LTE space, especially handsets that consist of an integrated SIP agent.

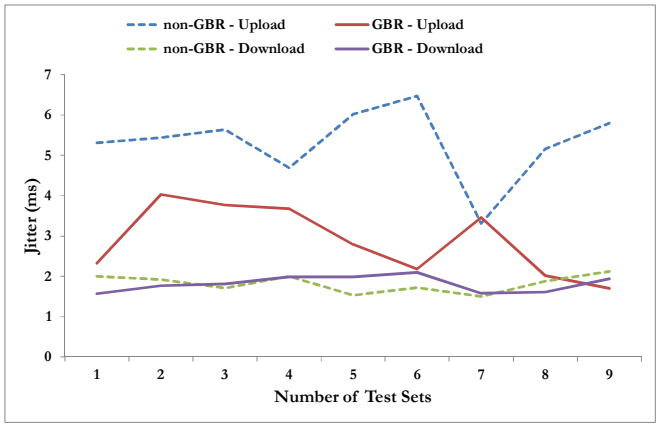



**Figure 6.** Average Jitter for calls to a Non-GBR Bit Rate Bearer versus calls to a GBR Bearer

## VI. CONCLUSION

In the research presented, the application of a Packet Inspection scheme in an LTE network for recognition of voice traffic and the initiation of a GBR Bearer for voice traffic was proposed. The notion was to recognise OTT or other voice and trigger a QoS Bearer via the proposed Admission Control scheme. The validity of the procedure was demonstrated via the Bearer initiation experiment which included proposing the assignment of an attribute value via an external interface by the voice recognition algorithm to the PCRF through its Rx interface. Theory that was tested included the manipulation of the TFT such that the filtering mechanism used to classify traffic into the voice GBR Bearer took the Softswitch IP address into consideration.

The testing proved that on recognition of the Softswitch IP address an attribute value would be assigned through the Rx interface to initiate the GBR Bearer with QCI 1 for voice traffic. The manipulation of the TFT profile in turn realised that only traffic flowing to and from the Softswitch IP address would traverse the voice GBR Bearer.

Latency and jitter results confirmed that QoS was being applied to the RTP voice stream. Two Bearers were tested; namely: the GBR Bearer and Non-GBR Bearer. The results for latency in the upload direction indicated a greater than 20% reduction in latency by the GBR Bearer as compared to the Non-GBR Bearer. The results for the latency in the download direction indicated a marginal reduction of 5-10% for the GBR Bearer over the Non-GBR Bearer. The jitter results in the upload direction provided an indication of a greater than 100% reduction in jitter for the GBR Bearer as compared to the Non-GBR Bearer. The results for jitter in the download direction were similar.

The experimental set-up demonstrated that QoS Bearer initiation for OTT voice is possible and viable in conjunction with a Packet Inspection voice Admission Control scheme.

## ACKNOWLEDGMENT

This work was supported in part by the University of Witwatersrand School of Electrical and Information Engineering. The authors acknowledge the support of the University of the Witwatersrand.

## REFERENCES

[1] The Skype Team. (2012, July) Skype.com. [Online]. http://www.skype.com/intl/en/home

[2] The Google Talk Team. (2012, July) Google Talk. [Online]. http://www.google.com/talk/

[3] Unstrung Insider, "Voice over LTE: Many Questions, No Easy Answers," *Light Reading*, vol. 8, no. 10, October 2009.

[4] 3rd Generation Partnership Project, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Stage 2," 3GPP, Technical Specification 3GPP TS 36.300 V9.3.0, 2010.

[5] 3rd Generation Partnership Project, "Evolved Universal Terrestrial Radio Access (E-UTRA); LTE physical layer; General description (Release 9)," 3GPP, Technical Specification 3GPP TS 36.201 V9.1.0, 2010.

[6] Nageshar N, Van Olst R, "A Heuristic Analysis Approach to Admission Control for Voice in Packet Switched Wireless Networks," in *IEEE Africon*, Livingstone, Zambia, 2011, pp. 1 - 6.

[7] Allot Communications, "Digging Deeper Into Deep Packet Inspection (DPI)," White Paper 04.2007 Allot Communications, 2007.

[8] Nageshar N, Van Olst R, "Regulation of Bearer / Service Flow Selection between Network Domains for Voice over Packet Switched Wireless Networks," in *ITU Kaleidoscope*, Cape Town, South Africa, 2011, pp. 175 - 180.

[9] Morin S, Allan K, Dolgonow A, "Application-Level Processing for Default LTE Bearer in S-GW," United States Patent Application Publication, Patent Pub. No. US 2010/0067400 A1, 2010.

[10] Sultana S, Rommer S, Frid L, Mulligan C, Olsson M, *SAE and the Evolved Packet Core*, 1st ed. Oxford, UK: Elsevier Ltd., 2009.

[11] Srinivasa Rao V. (Accessed March 2010) Signaling Procedures in LTE. Published in webbuyersguide.com. [Online]. http://www.slideshare.net/allabout4g/lte-signaling

[12] 3rd Generation Partnership Project, "IMS Multimedia Telephony Communication Service and Supplementary Services - Stage 3 (Release 9)," 3GPP, Technical Specification 3GPP TS 24.173 V9.1.0, 2010.

[13] Neekzad B, Hui J, Vannithamby R, Alasti M, "Quality of Service in WiMAX and LTE Networks," *IEEE Communications Magazine*, pp. 104 - 111, May 2010.

[14] Wireshark Foundation. (Accessed January 2011) Wireshark. [Online]. http://www.wireshark.org/

[15] 3rd Generation Partnership Project, "Circuit Switched (CS) Fallback in Evolved Packet System (EPS) Stage 2," 3GPP, Technical Specification 3GPP TS 23.272 V9.3.0, 2010.

**Nikesh Nageshar,** is a PhD graduate from the University of the Witwatersrand, Johannesburg and is currently employed in the telecommunications industry (e-mail: nikesh.nageshar@gmail.com).

**Rex Van Olst,** is an Associate Professor with the University of the Witwatersrand, Johannesburg. He heads the telecommunication postgraduate research arm at the School of Electrical and Information Engineering (e-mail: rex.vanolst@wits.ac.za).

# A Performance Analysis of the Phase Shift and Pulse Delay Techniques for Chromatic Dispersion Measurements and Compensation in Single Mode Fibre

S. Wassin, E. K. Rotich Kipnoo,R. R. G. Gamatham, A. W. R. Leitch and T. B. Gibbon
Physics Department
Nelson Mandela Metropolitan University, P. O. Box 77000, Summerstrand South,
Port Elizabeth-6031, South Africa.
Tel: +27 41 5042141
Email: Shukree.Wassin@live.nmmu.ac.za

**Abstract- Chromatic dispersion in an optical fibre arises from the wavelength dependence of the group velocity. This results in the broadening of a modulated optical pulse, pulse spreading and inter-symbol interference, ultimately increasing the bit errors. In this paper, the performances of the phase shift and the pulse delay measurement techniques for chromatic dispersion measurements were analyzed and compared. The phase shift method proved to be the more accurate measurement technique. Furthermore, a 40% decrease in the dispersion coefficient was obtained by designing a simple compensation system.**

**Index Terms—Chromatic dispersion, pulse delay, phase shift, VCSEL.**

## I. INTRODUCTION

Since the development of the Internet, a wave of developments with regards to international telecommunication systems has been observed [1]. Some of these developments include, increasing the bit rates and the bandwidth spectrum of transmission. As a result, the rise in the bandwidth usage has increased and the need for higher transmission rates grows daily. With optical fibre fostering the development of high-speed telecommunication networks, projects such as FTTx, e-Education, e-Health and e-Business are becoming more and more favorable. These types of systems are in the forefront of all leading telecommunication organizations, as well as governments all over the world. Chromatic dispersion is an integral property of the optical fibre that limits the transmission of data at high bit rates. The ability to measure and characterize the chromatic dispersion in single mode fibre now becomes extremely important. In this article, an in-depth discussion is given on two possible techniques for characterizing the chromatic dispersion in different optical networks. Furthermore, a method of compensating for chromatic dispersion by using non-zero dispersion shifted fibre with opposite sign, in any optical network is described and experimentally illustrated.

## II. THEORY

Consider a modulated optical signal, propagating along a single mode fibre having length *l*, as shown in Figure 1(a). The individual spectral components of the signal with its associated frequency $\omega$, travels with a group delay mathematically defined as,

$$\tau_g = \frac{l}{v_g} \tag{1}$$

where $v_g$ is known as the group velocity expressed [2, 3] as,

$$v_g = \left(\frac{d\beta}{d\omega}\right)^{-1} \tag{2}$$

Furthermore, from $\beta = nk = n\frac{\omega}{c}$, the group velocity can now be expressed as

$$v_g = \frac{c}{n} \tag{3}$$

by applying equation 2 [2]. From equation 2, it follows that the frequency of each spectral component is dependent on the group velocity $v_g$ of the modulated signal. Therefore, different spectral components of the modulated optical signal will propagate at different velocities. Hence, the pulses formed by the modulation are now broadened. The phenomenon described above is called chromatic dispersion, intra-modal dispersion or group velocity dispersion (GVD) [2].

Suppose the quantity $\Delta\omega$ of the pulse, is related to the pulse width, it now follows that the expression describing the amount by which the pulse broadens for a fibre with a given length fibre *l* is given by,

$$\begin{aligned} \Delta\tau_g &= \frac{d\tau_g}{d\omega}\Delta\omega \\ &= \frac{d}{d\omega}\left(\frac{l}{v_g}\right)\Delta\omega \\ &= l\frac{d^2\beta}{d\omega^2}\Delta\omega \\ \therefore \Delta\tau_g &= l\beta_2\Delta\omega \end{aligned} \tag{4}$$

where $\beta_2 = \frac{d^2\beta}{d\omega^2}$ is known as the group velocity dispersion (GVD) parameter or the second order dispersion parameter

[2]. The GVD parameter is an indication of the extent of the broadening experienced by the pulse, upon propagation along the fibre [2]. The value $\Delta\omega$, can be obtained from the wavelength spectrum $\Delta\lambda$ of the laser source. By relating,

$$\omega = \frac{2\pi c}{\lambda} \text{ and } \Delta\omega = \left(-\frac{2\pi c}{\lambda^2}\right)\Delta\lambda \quad (5)$$

equation 4 can now be expressed,

$$\Delta\tau_g = \frac{d}{d\lambda}\left(\frac{l}{v_g}\right)\Delta\lambda \quad (6)$$

$$= Dl\Delta\lambda$$

Equation 6 is now rewritten as,

$$D(\lambda) = \frac{d}{d\lambda}\left(\frac{1}{v_g}\right) = -\frac{2\pi c}{\lambda^2}\beta_2 = \frac{\Delta\tau_g}{l\Delta\lambda} \quad (7)$$

Equation 7 is defined as the *chromatic dispersion coefficient*, which describes the spreading of an optical pulse with respect to the wavelength of operation [2, 3].

Chromatic dispersion is given in ps/nm.km and is the sum of material $D_M$ and waveguide $D_W$ dispersion, mathematically expressed as $D = D_M + D_W$ [3, 4, 5]. Material dispersion arises since the refractive index of the optical fibre fluctuates with respect to the wavelength. Subject to the wavelength under consideration, the different spectral components will be transmitted at distinct velocities [3]. In standard single mode fibre, material dispersion is negative for wavelengths below 1300 nm and positive for wavelengths above 1300 nm, therefore, material dispersion can either be positive or negative in sign [5, 6]. Waveguide dispersion occurs as a result of the physical structure and shape of the fibre and is dependent on factors such as the core diameter, cladding diameter and also the difference in refractive indices between the fibre core and cladding [2, 6].

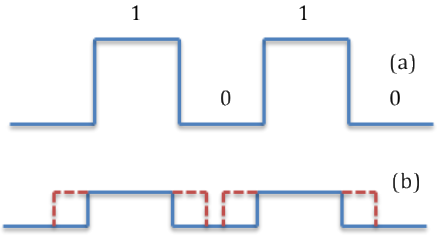



Figure 1: (a) Optically modulated "on-off" signal unaffected by chromatic dispersion. The pulse represented by a "1" can be considered to be a wavepacket containing a finite spectrum of wavelengths. (b) Pulse broadening, spreading and inter symbol interference shown as a result of chromatic dispersion.

With that being said, different spectral components with reference to a specific frequency $\omega$, will travel faster in the cladding and slower in the core of the optical fibre. This leads to pulse broadening, spreading and inter-symbol interference as shown in Figure 1(b). In single mode fibre, waveguide dispersion is negative in the 0 nm to 1600 nm wavelength spectrum [2, 6]. Chromatic dispersion is a deterministic and linear property of an optical fibre [4].

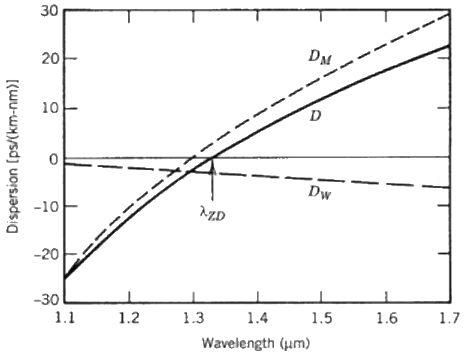



Figure 2: Classical chromatic dispersion $D$, material dispersion $D_M$ and waveguide dispersion $D_W$ curves [2].

It is possible to compensate for, or reduced the chromatic dispersion in a particular fibre link, depending on how a particular optical network is designed and implemented. One way of achieving this is to make use of the material and waveguide dispersion properties of a fibre. During the manufacturing stages, refractive index tailoring produces fibres with specific chromatic dispersion value at a discrete wavelength, as illustrated in Figure 3. Therefore, a standard G.652 single mode fibre or the G.655 non-zero dispersion-shifted fibre (+) can be added to the G.655 non-zero dispersion-shifted fibre (-). As a result, a low dispersion fibre network can be created.

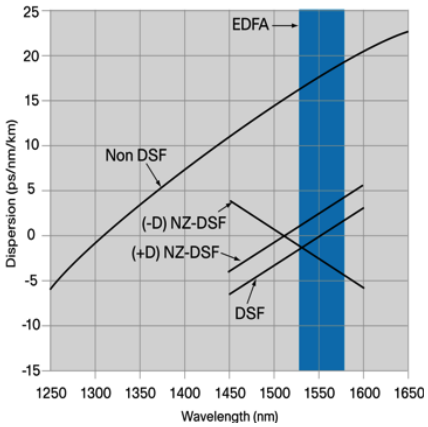



Figure 3: Chromatic dispersion curve related to the various single mode fibre [4].

There are a number of techniques available for measuring and characterizing the chromatic dispersion in single mode fibres. The pulse delay or time of flight technique determines the time delay between pulses transmitted at distinct wavelengths. This technique requires a high-powered laser and covers a larger bandwidth [8]. It also provides a measurement resolution of between 50 ps to 100 ps [8]. The pulse delay technique can be used to characterize the chromatic dispersion in optical fibres as long as 95 km [6, 8]. The phase shift technique, measures the phase variation between sinusoidally modulated signals propagating along a fibre at different wavelengths. The phase shift technique is reported to have a measurement resolution between 10 ps and 20 ps [8]. For the improvement of the measurement accuracy, a high signal to noise ratio is required, greater than 15 dB. Similarly, the accuracy increases with higher modulation frequency and when longer

fibres are measured [8, 9]. The technique having the best resolution for chromatic dispersion characterization in single mode fibre is the interferometic technique. This technique yields a 0.1 ps measurement resolution and is suitable for fibres shorter than five meters in length [8]. A novel technique developed for characterizing the chromatic dispersion in single mode fibre is discussed in [10].

In the sections to follow, we discuss and present the pulse delay and phase shift measurements techniques and the result obtained.

## III. EXPERIMENTAL PROCEDURE

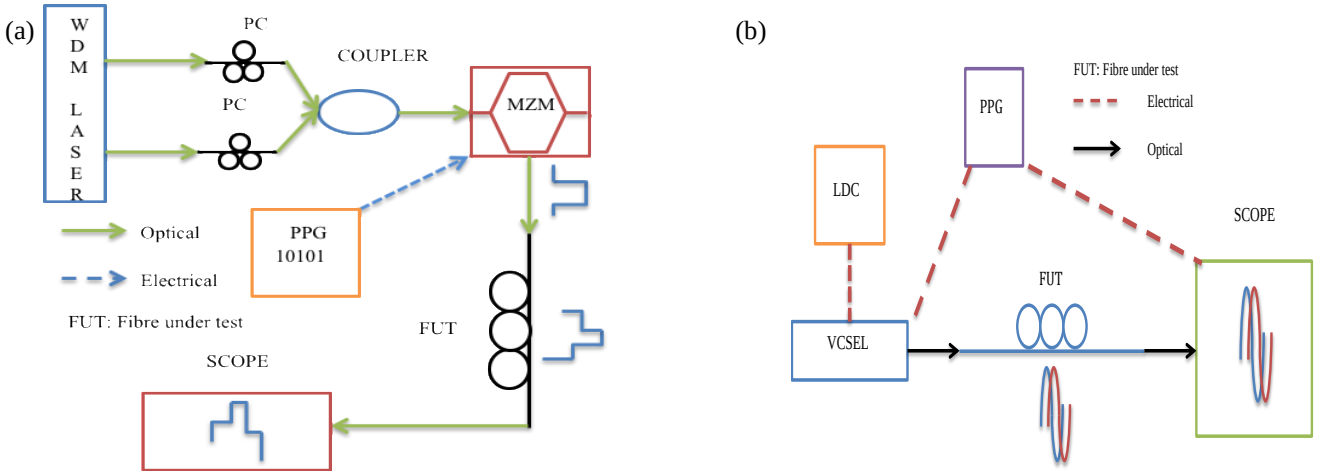



Figure 4: Measurement arrangements for the (a) pulse delay and (b) phase shift techniques for chromatic dispersion characterization.

Two aspects related to chromatic dispersion were experimentally investigated. Firstly, the measurement accuracy of the pulse delay and phase shift techniques was examined. The experimental arrangements are illustrated in Figures 4(a) and 4(b) respectively. Secondly, a chromatic dispersion compensation strategy was designed and tested utilizing the phase shift measurement technique. The experiments conducted were self proposed and constructed in controlled laboratory conditions by utilizing high tech equipment.

### A. *Pulse delay measurement technique*

Chromatic dispersion in 6.1 km, 12.2 km and 18.3 km G.652 single mode fibre was measured, respectively. Likewise, the chromatic dispersion of G.655 non-zero dispersion shifted fibre (+) was characterized for 26.6 km, 51.4 km and 76.7 km respectively. A high-powered wavelength division multiplexer (WDM) laser, with a wavelength spectral range of 1550 nm to 1554 nm was used as the optical source. The WDM laser is an array of eight laser modules, with a wavelength channel spacing of 50 GHz or 100 GHz. The continuous light, ejected from the WDM laser was then modulated by a Mach-Zehnder modulator (MZM) forming the pulses illustrated in Figure 1(a). A programmable pattern generator (PPG) defined the arrangement of bits (1's and 0's). The time delay $\Delta\tau$, was now measured between a pulse with a fixed reference wavelength $\lambda_1$ and that of wavelengths $\lambda_{2\text{-}8}$. The dispersion $D(\lambda)$ was then determined by substituting $\Delta\tau_g$ in equation 7.

An oscilloscope containing a photo detector of optical sensitivity -21 dBm and an 8.5 GHz optical bandwidth was used for measuring the time delay $\Delta\tau_g$. Polarization controllers (PC) were used, as the MZM is polarization sensitive.

### B. *Phase shift measurement technique*

For this experimental arrangement, the same lengths of G.652 and G.655 fibre characterized by the pulse delay measurement technique were analyzed. A directly modulated Vertical Cavity Surface Emitting Laser (VCSEL) with a 4.25 Gbps data rate was used, operating within a 1545 nm to 1550 nm wavelength spectrum.

A programmable pattern generator (PPG) inscribed the sinusoidal data onto the light emitted from the VCSEL. The modulation frequency, $f_m$ was set to 8.5 GHz. The same oscilloscope used in the pulse delay measurement technique was utilized for the phase shift technique. The bias current of the VCSEL was varied with a Laser Diode Controller (LDC), hence shifting the wavelength of operation of the VCSEL. A phase change occurs because of the correlation between the wavelength and the group velocity, in agreement with

$$\Delta\varphi = 360^\circ \Delta\tau_g f_m$$
$$\therefore \Delta\tau_g = \frac{\Delta\varphi}{360^\circ f_m} \quad (8)$$

where $\Delta\tau_g$ is the group delay and $f_m$ is the modulation frequency [11]. Since the oscilloscope was not able to directly measure the phase difference between two sinusoidal signals, a time delay value, $t_i$ between a reference and test signal was recorded from the oscilloscope. This time delay, $t_i$, was converted to a phase difference by using

$$\Delta\varphi = \frac{360^\circ t_i}{t_{per}} \quad (9)$$

where $t_{per}$ is period of the clock signal. Once $\Delta\phi$ was determined, equation 8 was inserted into equation 7 and the dispersion, $D(\lambda)$ was calculated.

### C. Compensation measurement schematic

Chromatic dispersion compensation was achieved by merging 51.4 km, G.655 non-zero dispersion shifted fibre (+) and 25.5 km of G.655 non-zero dispersion shifted fibre (-). Furthermore, the phase shift measurement technique, discussed in section III B and illustrated in Figure 4(b), was used for characterizing the chromatic dispersion for this compensation system.

## IV. Results and Discussions

### A. Technique performance analysis

The chromatic dispersion of 18.3 km G.652 and 51.4 km G.655 (+) single mode fibres was measured using the measuring system discussed in Figure 4(b).

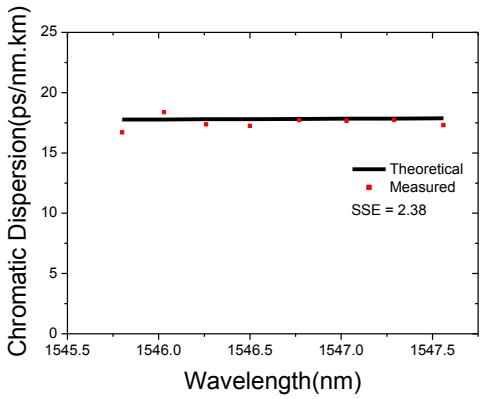



Figure 5: Experimental chromatic dispersion results obtained by the phase shift measurement technique of 18.3 km G.652 fibre.

A sum of squares of error (SSE) test was established in order to determine how well the phase shift measurement technique performed.

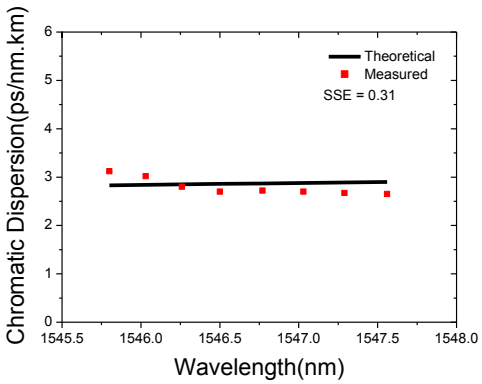



Figure 6: Experimental chromatic dispersion results obtained by the phase shift measurement technique of 51.4 km G.655 (+) fibre.

A fit of the theoretical dispersion curve was done onto the experimental data, a SSE=2.38 for the G.652 and a SSE=0.31 for the G.655 (+) single mode fibre was calculated respectively. This is illustrates in Figures 5 and 6. The pulse delay measurement technique was used for characterizing the chromatic dispersion in 18.3 km G.652 and 51.4 km G.655 (+) single mode fibres, as described in Figure 4(a).

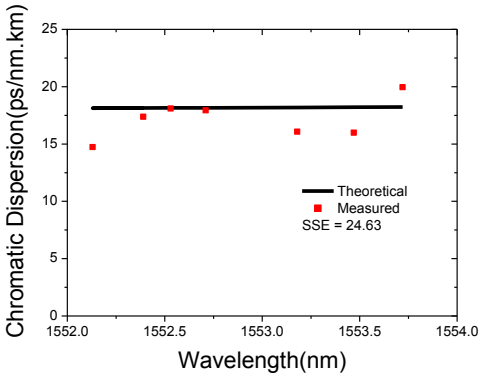



Figure 7: Experimental chromatic dispersion results obtained by the pulse delay measurement technique of 18.3 km G.652 fibre.

Again, a SSE test was done and a SSE=24.63 and SSE=9.96 was determined for the 18.3 km G.652 and 51.4 km G.655 single mode fibres respectively. An illustration of this is shown in Figures 7 and 8 respectively.

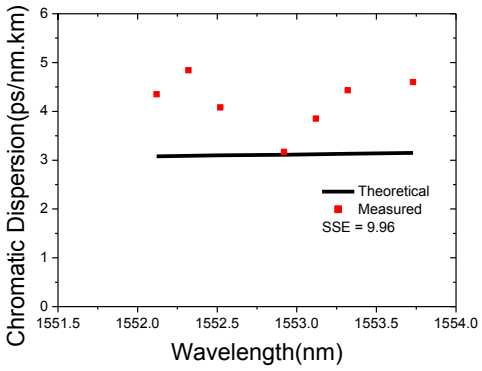



Figure 8: Experimental chromatic dispersion results of obtained by the pulse delay measurement technique of 51.4 km G.655 (+) fibre.

From the findings of the SSE tests, it can be inferred that the phase shift measurement technique performed better than the pulse delay measurement technique. As the SSE values determined for the results obtained by the phase shift technique are smaller than those measured by the pulse delay technique. This agrees well with the findings of literature, which reported that the phase shift technique has superior measurement accuracy as that of the pulse delay technique [8].

### B. Chromatic dispersion compensation measurements

The measured compensation results of a 77 km G.655 fibre link referred to in section III C, was compared to that of a 77 km long G.655 (+) fibre. The time delay measurement with respect to wavelength is illustrated in

Figure 9. With reference to 1546.70 nm, an average time delay $\tau$, of 253.54 ps was measured for the 77 km long G.655 (+) fibre. This result was compared to an experimentally measured time delay, $\tau$ of 102.01 ps for the 77 km compensation link.

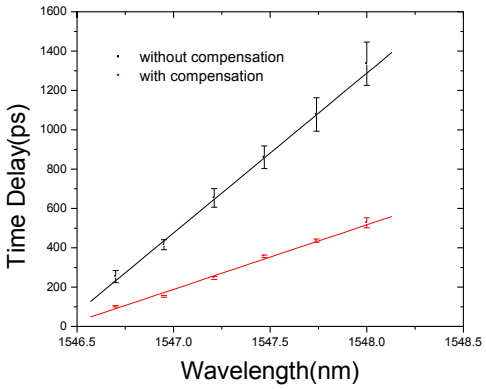



Figure 9: Measured time delay curves for 77 km fibre link. The red dotted curve represents the chromatic dispersion compensation arrangement.

Figure 10 shows a comparison between the dispersion curves of the 77 km chromatic dispersion compensation arrangement and the 77 km G.655 (+) fibre. It is clear that there is a significant drop observed in the chromatic dispersion, due to the merging of the G.655 (+) and G.655 (-) fibre. With reference to 1546.70 nm, a dispersion $D(\lambda)$, of 3.56 ps/nm.km was experimentally measured for the 77 km G.655 (+) fibre and 1.43 ps/nm.km for the 77km compensation arrangement discussed in section III C.

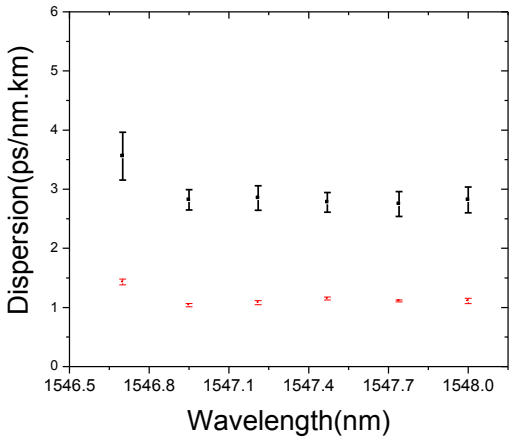



Figure 10: Chromatic dispersion curve for a 77 km long fibre link. The red dotted curve represents chromatic dispersion compensation system.

We were able to effectively demonstrate a 40% decrease in the chromatic dispersion by merging the 25.5 km long G.655 (-) fibre with a G.655 (+) fibre, 51.4 km in length. In doing so, we have illustrated the value of chromatic dispersion compensation when designing telecommunication network systems. Based on our ability to reduce the chromatic dispersion by 40%, it is possible to obtain an optical fibre network system with an overall dispersion of 0 ps/nm or any appropriate negligible value thereof. This can be achieved by matching the appropriate lengths of G.655 (+) or G.652 fibre with G.655 (-) fibre.

## V. Concluding Remarks

Two effective methodologies for characterizing the chromatic dispersion in any optical fibre link have been described. The phase shift technique demonstrated more accuracy than the pulse delay technique. A chromatic dispersion compensation strategy was illustrated and 40% chromatic dispersion compensation was achieved.

## VI. Acknowledgement

We are grateful for Research Funding and support from: Telkom, Dartcom, Infinera, Ingoma Communication Service, NLC, NRF, THRIP and the ALC.

## References

[1] L. G. Kazovsky, et al, *Broadband Optical Access Networks*, New Jersey: John Wiley & Sons Inc, 2011.
[2] G. P. Agrawal, *Fiber-Optic Communication Systems 3rd Edition*, New York: John Wiley & Sons Inc, 2002.
[3] G. Keiser, *Optical Fiber Communications 3rd Edition*, USA: McGraw-Hill Companies, 2000.
[4] Cisco White Paper, Fibre Types in Gigabit Optical Communication. USA: Cisco Systems, Available: http://www.cisco.com, [Accessed 01 May 2014], 2008.
[5] J. Hecht, *Understanding Fiber Optics 3rd Edition*. New Jersey: Prentice-Hall, 1999.
[6] N. Massa, *Fiber Optic Communication. Massachusetts: Springfield Technical Community College*, viewed 02 May 2014,
http://www.spie.org
[7] C. Lin, et al, "Pulse Delay Measurements in the Zero-Material-Dispersion Region for Germanium-and-Phosphorus-Doped Silica Fibres", *ELECTRONIC LETTERS*, vol.14, no. 6, pp. 170-172, 1976.
[8] L.G. Cohen, "Comparison of Single-Mode Fiber Dispersion Measurement Techniques", *Journal Of Lightwave Technology*, vol. LT-3, no. 5, pp. 958-966, 1985.
[9] S. Tanaka and Y. Kitayama, "Measurement Accuracy of Chromatic Dispersion by the Modulation Phase Technique", *Journal Of Lightwave Technology*, vol. LT-2, no. 6, pp. 1040-1044, 1984.
[10] L. Thevenaz, et al, *Review of Chromatic Dispersion Measurements Techniques*, viewed 24 July 2014,
http:www.infoscience.epfl.ch/record/173763/files/EFOC-LAN_1989_217.pdf
[11] B. Costa, et al, "Phase Shift Technique for the Measurement of Chromatic Dispersion in Optical Fibers Using LED's", *IEEE Journal of Quantum Electronics*, vol. QE-18, no. 10, pp. 1509-1515,1982.

## Biography

**Shukree Wassin** graduated with a BSc at Nelson Mandela Metropolitan University in 2012 and an Honours degree in Physics in 2013. He is currently registered for a Physics Masters degree at Nelson Mandela Metropolitan University.

# Analysis of Optical Signal to Noise Ratio in Modern Transmission Fibres during Raman Amplification

G. M. Isoe [1], K. M. Muguro [1], D. W. Waswa [1],
E. K. Rotich Kipnoo [2], T. B. Gibbon [2] and A.W. R. Leitch [2]
Optical Fibre and Laser Research Group, University of Eldoret[1], P.O Box 1125, Eldoret-30100 Kenya
Tel: +254-53-2063101; Fax: +254-053-43047
Optical Fibre Research Unit, Nelson Mandela Metropolitan University[2], P.O Box 77000,
Port Elizabeth, 6031, South Africa Tel: +27 41 504 2141 Fax: +27 41 504 2573

**Abstract— Optical Signal to Noise Ratio (OSNR) is one of the main performance indicators in Raman amplifiers. Detailed OSNR analysis can provide critical information about the performance of a Raman amplified optical network. In this work, co- and counter pumping schemes have been considered. A pump power of up to 23 dBm and -10 dBm signal power was used. An OSNR of 12.8 dB and 12.3 dB was achieved for co- and counter pumping schemes respectively for 25 km Single Mode Reach Fibre (SMF-Reach). For 50 km fibre, an OSNR of 10.0 dB and 9.3 dB was recorded for the two pumping schemes respectively. An on-off gain of 5.4 dB and 4.1 dB was achieved for co- and counter pumping schemes respectively for 25 km fibre. For a 50 km fibre, an on-off gain of 6.3 dB and 5.3 dB was obtained for the two pump configurations respectively. OSNR was observed to vary inversely with fibre length. Results of our study are important for the optimization of fibre Raman amplifiers in long haul signal transmission.**

**Index Terms— Fibre Raman amplifier, Optical Fibre, OSNR.**

## I. INTRODUCTION

Signal amplification in optical fibres during transmission has become one of the techniques in improving capacity and reach in telecommunication networks [1]. Fibre Raman amplifiers (FRAs) are among the best optical amplifier, both in long haul signal transmission and passive optical networks (PONs), for metropolitan applications [2,3]. The fact that signal amplification is distributed in FRAs is a major advantage of their application. In this case the signal is amplified along the fibre during propagation, a fact that drastically reduces the system nonlinearities [4].

OSNR is a vital parameter to be analyzed especially when carrying out impairment awareness and intelligent compensation in wavelength-division-multiplexed (WDM) optical networks, because different channels may route from different paths with varied amount of accumulated amplified spontaneous emission (ASE) noise [5]. ASE noise is generated as a result of spontaneous Raman scattering (SRS) in the process of Raman amplification. ASE noise of distributed FRA accumulates along the fibre resulting in the degradation of the system OSNR at the output. In this case the system performance is mostly limited by the OSNR rather than the optical power received. This noise also adds a wider band of background noise around the signal wavelength hence affecting the general performance of any optical communication system [6, 7].

Research in OSNR of optical fibres has so far been done using the non-modern fibre which is known to be limited by high level of impurities, high loss, as well as remarkable signal dispersion [8]. These effects have been found to enhance the optical noise in the system which in turn degrades the OSNR drastically. However modern fibres have been improved in quality and are of low polarization mode dispersion (PMD) [9]. As a result there is need to investigate and characterize the updated OSNR performance in modern fibre when used in Raman amplification as gain medium. In this work we investigate the OSNR performance of a Raman amplified optical system using co- and counter pumping schemes. We evaluate by experimental measurements the OSNR degradation with fibre length and pump power. Our discussion is limited to the Amplified Spontaneous Emission (ASE) noise and relative intensity noise (RIN) contribution of system OSNR degradation.

## II. THEORY

The process of Raman amplification involves the transfers of some of the pump power to the signal thus increasing the strength of the signal. The pump $p_p$ and signal $p_s$ power evolution over a longitudinal fibre axis $z$ are governed by the following coupled equations [10];

$$\frac{dp_s}{dz} = g_R p_p p_s - \alpha_s p_s, \tag{1}$$

and

$$\mp \frac{dp_p}{dz} = -\frac{\omega_p}{\omega_s} g_R p_p p_s - \alpha_p p_p. \tag{2}$$

where, $g_R$ is the Raman gain coefficient, $\omega_s$ and $\omega_p$ are the signal and pump angular frequencies respectively, while $\alpha_s$ and $\alpha_p$ are the signal and the pump attenuations respectively. The $\mp$ sign represents a backward and forward propagation of the pump wavelengths respectively.

Equations (1) and (2) imply that the signal is amplified by the pump at a certain proportion, with the constant of proportionality being determined by the Raman gain coefficient ($g_R$), and losses as a result of attenuations within the optical fibre. The pump power also reduces as a result of energy transfer to the signal as well as attenuations along the optical fibre. Assuming that there is no pump depletion, i.e. $g_R = 0$ in equation (2), the pump power

$p_p(z)$ as a function of fibre length $l$ in the counter pumping scheme is given as;

$$p_p(z) = p_0 e^{-\alpha_p(l-z)}, \quad (3a)$$

while in the co- pumping scheme is give as;

$$p_p(z) = p_0 e^{-\alpha_p(z)}. \quad (3b)$$

The OSNR at the receiver input of an optical amplifier can be calculated from the following differential equations [11, 12].

$$\frac{dp_s^+(z)}{dz} = T_r p_p(z)(p_s^+(z) + E_{pH} B_0) + rp_s^-(z) - \alpha_s p_s^+(z), \quad (4a)$$

$$\frac{dp_s^-(z)}{dz} = -T_r p_p(z)(p_s^-(z) + E_{pH} B_0) - rp_s^+(z) + \alpha_s p_s^-(z). \quad (4b)$$

Where, $B_0$ is the optical bandwidth, $r$ represents Rayleigh scattering, $E_{pH}$ is the photon energy of the signal. $p_s^+$ and $p_s^-$ represent the forward and backward travelling power levels of the signal respectively.
$\pm T_r p_p(z) B_0$, is the spontaneous Raman emission while $T_r p_p(z)$ describes the spontaneous Raman emission into one axial mode as the rate in photons per length. All spontaneous emitted photons propagate along the fibre hence experiences amplification. The terms $+rp_s^-(z)$ and $-rp_s^+(z)$ represent Rayleigh back scattering. Pump depletion has been neglected because the signal and noise remains at low powers compared to the pump power at any given point with significant Raman gain.
An approximate solution to equation (4) is obtained by interrelating with equation (3), so as to include only the single and double Rayleigh scattered signal and noise, neglecting light scattered more than twice. Solving equations (3) and (4) without the terms accounting for Rayleigh backscatter and/or the noise generating terms, we can derive the signal gain, the double Rayleigh scattered interfering signal, and the OSNR.

## III. EXPERIMENTAL SETUP

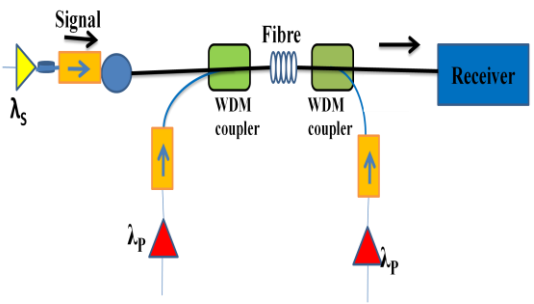



Fig 1: schematic diagram of a distributed Raman fibre amplifier.

Fig.1 shows the schematic diagram of a distributed RFA which was used in the experimental measurement. A -10 dBm source signal transmitted at a wavelength of 1550 nm was used. A typical externally modulated distributed feedback laser (DFB) was used. The typical output power of a DFB is about 12 dBm [13]. A pump power of up to 23 dBm and 1450 nm wavelength was used throughout the experiment. This ensured maximum amplification of the signal which in RFA occurs when the pump-signal detuning is 100 nm. The signal source and the pump were coupled using a filter based Wavelength Division Multiplexer (WDM) as the input coupler. Both were then propagated in a 25 km SMF-reach after which the pump wavelength was filtered out and the signal output analyzed. The same procedure was repeated for a 50 km fibre.

The signal power at the fibre output was measured using a GFHP-B power meter. The noise power at the output was also analyzed separately by the optical spectrum analyzer (OSA) and the resultant OSNR measured on the scope.

## IV. RESULTS AND DISCUSSION

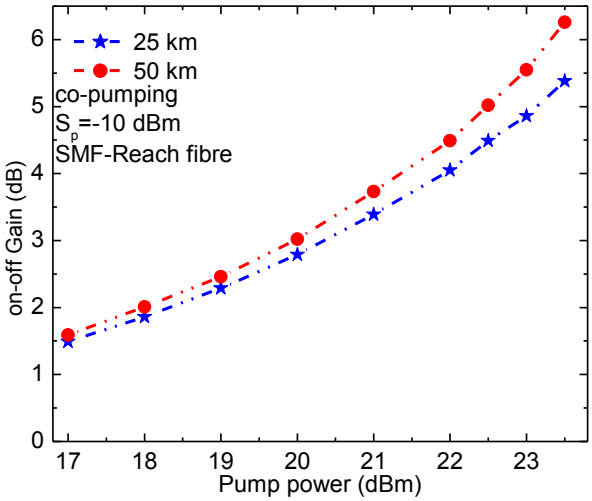



Fig 2: On-off gain evolution with pump power for co-pumping.

Fig 2 shows how the experimental on-off gain varies exponentially with pump power. For pump powers less than 17 dBm, a lower on-off gain of 1.5 dB and 1.6 dB was recorded for the 25 km and 50 km fibre. This is because this pump powers were still low to induce remarkable Stimulated Raman Scattering (SRS) and thus minimal power was transferred from the pump to the signal hence accounting for the small gain. However, as pump power increased above 17 dBm, the on–off gain increased exponentially up to a maximum of 5.4 dB and 6.3 dB for the two fibre lengths respectively. The increase in pump power induced remarkable SRS which in turn enabled pump to signal power transfer, hence compensating for the signal losses in the fibre as well as boosting the signal.

Fig 3 shows the experimental measurements for OSNR, for 25 km and 50 km fibre lengths for co-pumping scheme. From the results in Fig 3, the OSNR remained at lower level for pump powers less than 17 dBm. This is because the pump power over this region is very weak to introduce SRS thus accounting for the lower OSNR recorded.

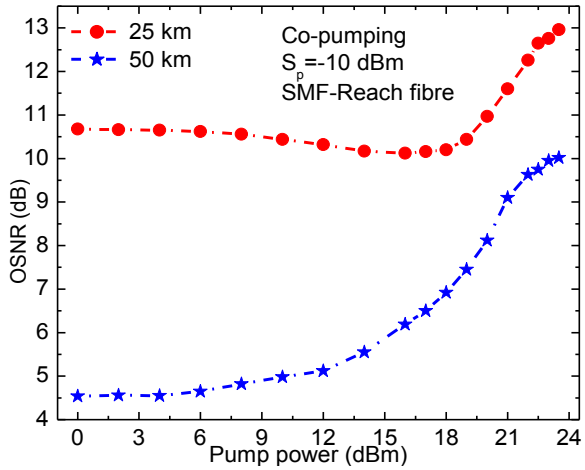



Fig 3: OSNR as a function of pump power at different fibre lengths for co-pumping scheme.

For pump powers beyond 17 dBm, a continuous increase in pump power resulted to a continuous increase in OSNR in both fibres. This is because as the pump power increases, the pump becomes more powerful to compensate for losses within the transmission fibre. A 25 km fibre recorded a higher OSNR of 12.8 dB compare to 10.3 dB recorded by a 50 km fibre at 23 dBm pump power. This is because an increase in fibre length also leads to an increase in ASE noise accumulation within the fibre thus accounting for the drop in OSNR at longer fibre lengths.

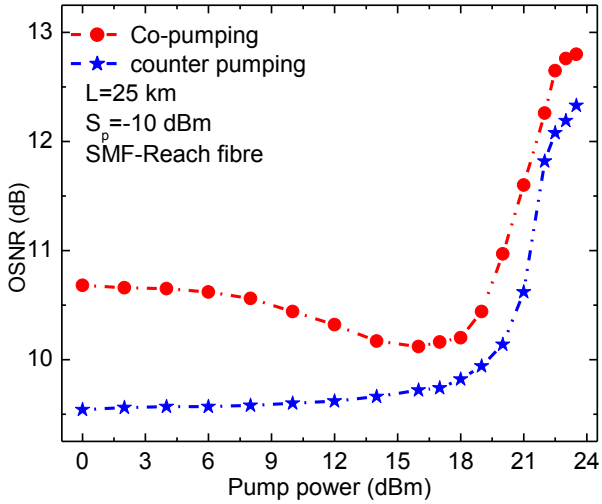



Fig 4: OSNR variation with pump power for co-pumping and counter pumping schemes.

From the experimental results in Fig 4, the OSNR for co – pumping reduces gradually for pump powers less than 17 dBm. This is because the pump power over this region is very weak to introduce remarkable SRS thus more noise is transferred from the pump to the signal thus a continuous reduction in OSNR. The RIN noise contribution from the pump is dominant over this low power region thus accounting for the OSNR drop especially as the pump power is increased between 8 dBm to 15 dBm. RIN noise accounts for the continuous drop in OSNR for co-pumping while the OSNR for counter pumping increases slightly over this low power range because counter pumping is more superior to RIN noise performance compared to co-pumping [14]. However, as the pump power is increased beyond 17 dBm, the OSNR of both the pumping schemes also increased. In fact, a substantial gain in OSNR was obtained beyond this pump power. This is because the pump power over this region was now strong enough to amplify the signal thus accounting for the continuous increase in the OSNR. However at pump power approaching 23 dBm, the OSNR of both pumping schemes started to degrade due to gain saturation. Co-pumping gave a better OSNR performance than counter pumping. This is because for the case of co-pumping, both the pump power and the signal power are propagated in the same direction. This ensures that the pump is well utilized by the signal thus more pump power will be transferred to the signal hence higher amplification levels are achieved. Also the ratio between the pump power and the signal power is maintained almost at the same level throughout the transmission span. For the case of counter pumping, the signal is propagated in a direction counter to the pump. This limits the pump to signal power transfer thus leaving much of the pump power unutilized hence accounting for the lower OSNR attained in this pumping scheme.

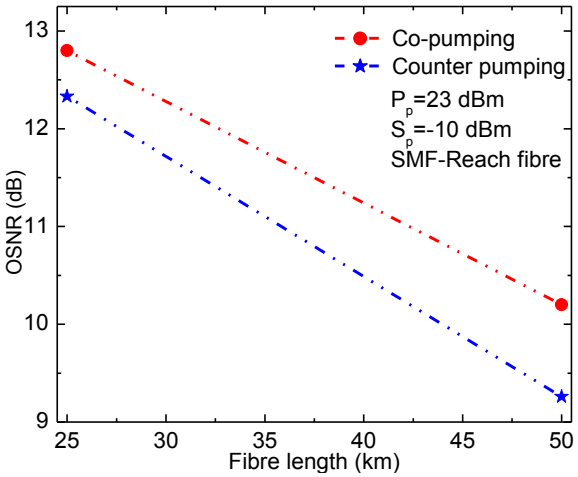



Fig 5: OSNR (dB) versus fibre length (km) at co- and counter pumping schemes.

Results in Fig 5 show the experimental measurements of the OSNR variations as a function of fibre length at different pump configuration. From Fig 5, the OSNR of both co- and counter pumping schemes reduces with an increase in fibre length. This is because the losses within the fibre increase with an increase in fibre length. As the fibre length is increased, the OSNR for counter pumping reduces greatly compared to that of co- pumping. This is because in the case of counter pumping, the pump power is not well utilized by the signal because the two are travelling in a direction opposite to each other. The ratio between the pump and the signal also reduced because the pump power reduces due to attenuations within the fibre as it approaches a stronger signal from the opposite end of the fibre. This in turn limits the pump to signal power transfer by a greater factor, which in turn increases the noise level within the fibre thus degrading the OSNR drastically. Co-pumping gave a better OSNR than counter pumping because co-pumping has a higher gain compared to counter pumping.

## V. CONCLUSION

Signal degradation in optical fibres during Raman amplification depends on the fibre length, pump power and the pumping scheme used. The OSNR decreases with an

increase in fibre length. This decrease is more in counter pumping than in co-pumping.

From our results, the ASE noise was minimal in the co-pumping thus accounting for the higher OSNR recorded in this pumping scheme. This makes co-pumping superior in noise performance compared with the counter pumping. We recommend this pumping configuration in long haul signal transmission.

## VI. ACKNOWLEDGEMENT

We are grateful for Research Funding and support from: Telkom, Dartcom, Infinera, Ingoma Communication Services, NLC, NRF, THRIP and the ALC.

## REFERENCES

1) K. Srivastava and Y. Sun. *Optical Fibre Telecommunications*, Chapter 4, pp. 174–212, Academic Press, 2002.
2) T. S. U. S. H. I. Mori, H. I. R. O. J. I. Masuda, K. Shikano, & M. Shimizu, "Ultra-wide-band tellurite-based fibre Raman amplifier," Journal of Lightwave Technology, pp. 1300-1306, (2003).
3) T. Monroy, R. Kjær, F. Öhman, K. Yvind, P. Jeppesen, "Distributed fibre Raman amplification in long reach PON bidirectional access links," Optical Fibre Technology (Elsevier) vol. 14, pp. 41-44, (2008).
4) C. Headley and G. P. Agrawal, *Raman Amplification in Fibre Optical Communication Systems*, Elsevier Academic Press, 2005.
5) J. Y. Yang, L. Zhang, Y. Yue, V. R. Arbab, A. Agarwal, L. Paraschis, and A. E. Willner, "Optical signal-to-noise ratio monitoring of an 80 Gbits/s polarization-multiplexed return-to-zero differential phase-shift keying channel," Opt. Lett., vol. 34, no. 7, pp. 1006–1008, (2009).
6) G. P. Agrawal, *Fibre Optical Communications Systems*, 3rd Edition, John Willey& Sons, inc., 2002.
7) R. G. Smith, "Optical power handling capacity of low loss optical fibre as determined by stimulated Raman and Brillouin scattering," Appl. Opt., vol. 11, no. 11, pp. 2489–2494, (1972).
8) G. Keiser, *Optical fibre communication*, 4th ed., New York, NY: McGraw-Hill, 2011.
9) H. R. Kogelnik, M. Jopson, L. E. Nelson, "Polarization mode dispersion in Optical fibre telecommunication," Academic press, San Diego CA, Chap 15, (2002).
10) M. L. Dakss, and P. Melman, "Amplified Spontaneous Raman Scattering and Gain Fibre Raman Amplifier," Journal of Lightwave Tech. Vol. 6, (1985).
11) J. Auyeung, and A. Yariv, "Spontaneous and stimulated Raman scattering in long low loss fibres," IEEE J. Quantum Electron., vol. QE-14, pp. 347–352, (1978).
12) E. Desurvire, M. J. F. Digonnet, and H. J. Shaw, "Theory and implementation of a Raman active fibre delay line," J. Lightwave Technol., vol. LT-4, pp. 426–443, (1986).
13) NEL, 1550 nm DFB Laser Diodes data sheet, 2005.
14) J. Bromage, "Raman Amplification for Fiber Communications Systems," Journal of Lightwave Technology, VOL. 22, NO. 1, pp. 88-90, (2004).

**G.M. Isoe** graduated with a B.Sc. Physics/Mathematics degree in 2010 from Moi University, Kenya. He is currently a M.Sc. Physics Student at the University of Eldoret, Kenya, working on Optical Noise Analysis in fibre Raman amplifiers.

# INTERNET SERVICES & END USER APLICATIONS

# Designing Novel Visualisation Techniques for Managing Personal Information across Multiple Devices

Simone Beets and Janet Wesson
Department of Computing Sciences
Nelson Mandela Metropolitan University, P. O. Box 77000, Port Elizabeth 6031
Tel: +27 41 5042323, Fax: +27 41 5042831
email: {Simone.Beets2, Janet.Wesson}@nmmu.ac.za

**Abstract-Research has shown that the information fragmentation problem has become exacerbated by the increasing number of devices used for storing personal information. Existing solutions do not sufficiently support personal information management (PIM) across multiple devices. This paper describes the design and evaluation of novel, interactive visualisation techniques to support PIM across different devices. A usability evaluation was conducted to determine the effectiveness of the proposed visualisation techniques. Results of the usability evaluation show that participants found that the proposed visualisation techniques provided a good overview of personal information across different devices within a single user interface. The participants also found that the visualisation techniques were easy to use and useful.**

**Index Terms—Personal Information Management, Information Visualisation, Multiple Devices, Design, Evaluation**

## I. INTRODUCTION

Personal information (PI) includes a user's information items such as documents and pictures. Personal information management (PIM) involves the operations that users need to perform with their PI [1]. The main problem relating to PIM is the known as the information fragmentation problem. This problem is becoming exacerbated by the increasing number of devices per user as well as the constantly growing volume of PI [2].

A personal space of information (PSI) refers to a user's entire PI collection on all devices. The hierarchical file/folder structure is currently the most common method used for PI organisation and visualisation. According to previous research [3], [4], the hierarchical file structure, as well as the indented list used to visualise the file structure, does not support accessing PI across multiple devices. The hierarchical file structure also possesses other limitations, including lack of visibility of files and folders and ineffective screen space usage [3]. Alternative solutions to the hierarchical file structure exist, such as Dropbox (www.dropbox.com) and TeamViewer (www.teamviewer.com), but these tools suffer from a number of limitations, such as problems associated with the hierarchical file structure and the visualisation thereof. A user is required to know beforehand what files need to be uploaded and no device information is available after PI is uploaded or transferred.

Novel visualisation techniques can be incorporated in a tool with the aim of effectively viewing PI [2]. Current systems that make use of visualisation techniques are focused on enhancing PIM, but suffer from several limitations [5]. Visualising PI across multiple devices requires multiple hierarchy visualisation. The PI distributed across multiple devices should preferably be displayed within a single user interface (UI) using an appropriate visualisation technique to view each device's hierarchy.

This paper describes the design and evaluation of the MyPSI prototype, which makes use of several novel visualisation techniques to support PIM across different devices. The next section describes related work regarding PIM and visualisation techniques. Section III details the design of the prototype, called MyPSI, and discusses the requirements verification. Section IV outlines the implementation of the prototype and the incorporation of the proposed visualisation techniques within MyPSI. The usability evaluation of the prototype is described in Section V. Section VI concludes the paper by identifying the contribution of the work and future work to be completed.

## II. RELATED WORK

### A. Personal Information Management

The main aim of PIM is to support a user in accessing his/her PI, while supporting the typical tasks of PIM [1]. Barreau [6] originally identified several PIM tasks including acquiring, organising, storing, maintaining, retrieving and producing PI. These PIM tasks were later simplified to include keeping (storing), managing (organising and maintaining) and finding (searching and browsing for information retrieval) [1]. Examples of PI manipulation tasks include creating, sorting, naming and copying PI.

### B. Information Visualisation Tools

Several tools that incorporate visualisation techniques for PIM were found in literature, including earlier tools such as LifeStreams [7] and Stuff I've Seen (SIS) [8], and more recent tools such as the Zoomable Object-oriented Information Landscape (ZOIL) [9] and PI Dashboard [2]. These tools provide support for enhancing PIM, but they suffer from several limitations, including concentrating on a particular PI type [5] or on limited PI aspects [10].

A number of the tools found in literature including Phlat [11] and SIS [8] do not use appropriate visualisation techniques to visualise PI. Most systems analysed omitted a hierarchy completely. Systems that made use of visualisation

techniques such as LifeStreams [7] and ZOIL [9], make use of a timeline and landscape metaphor to view PI. While the hierarchical file structure suffers from various limitations, it may be better to augment or extend the structure rather than replace it entirely, due to its familiarity to users [12].

In order to view PI across multiple devices and according to each PI type, multiple hierarchies need to be considered. Limited research was found relating to visualising multiple hierarchies to assist information retrieval. Graham and Kennedy [13] reviewed existing visualisation techniques for viewing multiple hierarchies. A graph-based visualisation technique and an innovative set-based technique to visualise a botanical taxonomy were reviewed in [14]. The comparison revealed that the set-based visualisation technique was the preferred technique to visualise the plant classification scheme as it was the more structured visualisation technique.

## III. Design

### A. Functionality

An interview study was used to obtain requirements for PIM across multiple devices by determining the existing problems in managing PI across different devices [4]. Functional requirements were then derived from these results to be supported by the MyPSI prototype. The functionality to be supported was described previously [15] and is summarised below:

1. *Manipulation:* The manipulation functionality supports the requisite to provide complete access to the PI items.
2. *Sorting:* Enabling sorting of the PSI facilitates providing different PI views and browsing.
3. *Intelligent Browsing:* Intelligent browsing within a PI collection type supports browsing across multiple devices.
4. *Intelligent Searching:* Intelligent searching supports searching across multiple devices.
5. *Filtering:* Filtering supports the searching facility or can be used independently of the search facility.
6. *Tagging:* Tagging allows the user to tag PI items.
7. *Linking:* A user can link related PI items within different PI collections.

### B. Visualisation Techniques

The selection of the set-based visualisation technique, hereafter referred to as the *Partition Layout*, was incorporated in the MyPSI tool as the technique facilitates a structured method to visualise multiple hierarchies, allowing a user to view on which device particular PI is stored, while minimising clutter (Section IIB). Three visualisation techniques were carefully chosen to support PIM in MyPSI. An *Overview* makes use of an interactive nested circles layout; a *Tag Cloud* represents the tags in the PSI; and the *Partition Layout* is used to visualise the folder structures for each user device:

#### 1) The Overview

The *Overview* consists of an enclosure diagram, which is displayed by means of a nested circles layout [16]. A general view of the PSI is provided by the *Overview* (Figure 1). The entire PSI is represented by the outer circle, while the PI types, including documents, pictures, music and videos, represent the following level of circles. The subsequent level of circles represent the user devices on which the different PI is kept. The circle size denotes the ratio of the size of the particular PI collection in relation to the entire PSI.

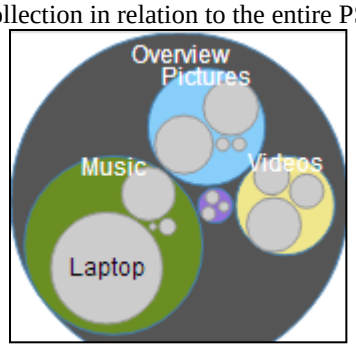



**Figure 1 The Overview of MyPSI**

#### 2) The Tag Cloud

The *Tag Cloud* represents existing tags in the PSI. A *Tag Cloud* is generally used for information-seeking tasks [17]. The *Tag Cloud* in MyPSI (Figure 2) represents existing tags assigned to PI items within the user's PSI.


**Figure 2 The Tag Cloud of MyPSI**

#### 3) The Partition Layout

The set-based visualisation technique was enhanced and implemented in the MyPSI prototype to support information retrieval. The *Partition Layout* displays the hierarchy of each device (Figure 3).

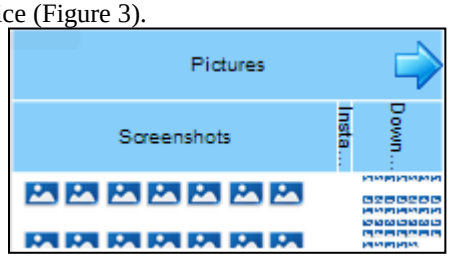



**Figure 3 A Folder Structure in the Partition Layout of MyPSI**

The *Partition Layout* represents a space-filling visualisation technique, similar to a treemap. The primary folders used for visualisation include the user's Libraries folder, which includes the Documents, Music, Pictures and Videos sub-folders. The *Partition Layout* represents the entire folder structure for each device, displaying these folder structures beneath one another. Within a device, each Libraries sub-folder is displayed parallel next to one another. Finally, subsequent sub-folders are located underneath parent folders, where PI items can be found below their direct parent folders.

### C. Requirements Verification

A cognitive walkthrough was conducted with a number of participants that took part in the interview study [4]. The participants of the walkthrough were provided with the requirements and the conceptual design of the MyPSI

prototype together with the proposed interaction, as discussed in a previous paper [15]. Results of the cognitive walkthrough were generally positive, verifying all the requirements as discussed in Section IIIA. The design of the MyPSI prototype and the suitability of the visualisation techniques, were also confirmed.

## IV. Implementation

### A. Implementation Tools

The MyPSI prototype requires that each device needs to have an Internet connection to enable access to the PI stored on the different devices. The MyPSI prototype was developed as a web-based application using a combination of the $D^3$ JavaScript visualisation library [18], HTML (Hypertext Markup Language), CSS (Cascading Style Sheet) and Bootstrap (http://getbootstrap.com/), a framework that works with HTML and CSS to provide a responsive layout for a web application. The $D^3$ visualisation toolkit, or Data-Driven Documents, provides an extensive and powerful means to visualise large datasets on the web using HTML, CSS and SVG (Scalable Vector Graphics) with a data-driven approach. A literature review was conducted to determine the most appropriate visualisation toolkit with which to implement the MyPSI prototype including Raphael (http://raphaeljs.com/) and the JavaScript InfoVis Toolkit (http://philogb.github.io/jit/). $D^3$ was selected as the preferred implementation tool as it provides more customisation possibilities and supports a wider range of visualisation techniques (http://d3js.org/).

### B. The MyPSI Prototype

The MyPSI prototype was implemented following the design described in Section III. The *Overview* (Figure 1), Dashboard and *Tag Cloud* (Figure 2) are displayed on the left of the screen within a sidebar that can be hidden depending on the available screen space. A toolbar is located at the topmost section of the screen, which allows for general operations such as undoing and redoing actions, searching, filtering and sorting. The *Partition Layout* occupies the remaining screen space.

The *Overview* is interactive and co-ordinated with the *Partition Layout*, so that, if a device is selected within the *Overview*, the *Partition Layout* will filter and collapse the other devices, only displaying the selected device's information. Additionally, if the *Overview* is reset, any devices collapsed within the *Partition Layout* will be expanded again. A user can also determine the total size regarding a PI type or PI on a specific device.

The *Tag Cloud* (Figure 2) provides a separate filtering facility and is co-ordinated with the *Partition Layout*. When a new tag is assigned to a PI item which does not exist within the *Tag Cloud* collection, the tag is added to the *Tag Cloud*. In addition to the search or used independently, a tag could be selected in the *Tag Cloud* to only display those PI items associated with that specific tag.

The *Partition Layout* is used to visualise the folder structures on the different devices (Figure 3). For the purpose of the prototype, four devices are displayed in the *Partition Layout*. Devices are represented by device name and each device's folder structure is represented by a window, which can be collapsed or expanded.

Colour-coding was used to distinguish between the different PI types. The colour-coding that was assigned to these PI types include purple for documents, blue for pictures, olive green for music and yellow for videos. This colour-coding addresses the red-green colour vision deficiency [19].

Sub-folders that have been most recently accessed within the Libraries folders are displayed by default, but no more than three sub-folders are displayed at a time. An arrow is displayed on the folder, which symbolises that the folder has hidden sub-folders. These arrow keys enable scrolling to browse for hidden sub-folders. PI items are initially represented by icons identifying their PI types. Hovering over or clicking on a file or folder will display a tooltip or popover respectively, to provide more information about the PI item. MyPSI also supports performing a search or filter, thereafter only displaying the relevant PI (Figure 4), de-saturating remaining PI items, as used in Shneiderman's SocialAction tool [20].

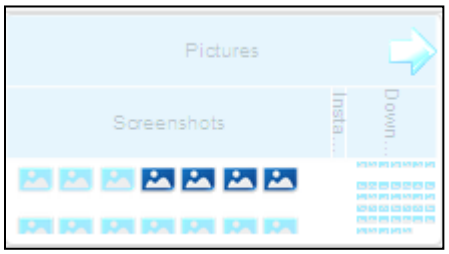



**Figure 4 Search and/or Filtering the Partition Layout**

The MyPSI prototype provides semantic zooming similar to the ZOIL collaborative tool [9]. When zooming into a specific PI item, the item enlarges, revealing the type of PI item and a thumbnail. Further zooming into the item will reveal a preview of the item, for example, the first few lines of content if the item is a document. The MyPSI tool supports all of the manipulation functions as stated in the required functionality.

## V. Evaluation

### A. Method

A user study was conducted to determine whether any usability problems existed within MyPSI. Usability metrics captured for this usability evaluation included effectiveness (task completion) and user satisfaction. Effectiveness was captured for each task using a task list together with participant observation. Satisfaction was captured using a post-test questionnaire. The task list for the usability evaluation consisted of seven main tasks with a training task included for each of these tasks. The evaluation made use of a within-subjects experimental design. Ten participants completed the usability evaluation of the MyPSI prototype.

#### 1) Evaluation Objectives

The aim of the usability evaluation was to determine the effectiveness of the visualisation techniques incorporated in the MyPSI prototype. The evaluation was also used to identify any usability problems with the MyPSI prototype.

#### 2) Participants

A convenience sample of ten postgraduate students from

the Department of Computing Sciences at the Nelson Mandela Metropolitan University (NMMU) was used for the usability evaluation. Participants were in the age range of 21-39 years. Seven participants had at least six years computer experience. The remaining three participants had 3-5 years computer experience. Most participants (n=7) managed their PI on a daily basis. The remaining participants managed their PI on a weekly (n=2) and monthly basis (n=1). Seven participants made use of at least three devices to manage PI. The main device used by participants to manage PI was either a desktop computer (PC) (n=5) or a laptop (n=3).

*3) Collection Methods and Evaluation Metrics*

Effectiveness was used to measure task success, i.e. whether a participant was able to complete a task or not. Satisfaction was captured using a post-test questionnaire, which made use of a combination of the NASA-TLX form [21] and the Computer Satisfaction Usability Questionnaire (CSUQ) [22]. The NASA-TLX form was used to measure cognitive load. The CSUQ was used to measure overall satisfaction and usability, and to capture general comments. Additional questions were asked specifically relating to the different visualisation techniques and included the following:

- *Technique X* provided a good overview of…;
- *Technique X* was useful;
- *Technique X* was easy to use;
- The zoom and filter of *Technique X* was easy to use;
- The zoom and filter of *Technique X* was useful.

A final question was added for the *Partition Layout*, namely "*The zoom feature of the Partition Layout enabled me to obtain more information on a file*". Questions in the post-test questionnaire were rated using a 7-point Likert scale.

*4) Tasks*

Seven main tasks were included in the task list. Three sub-tasks were included in the Manipulation task. The participants were also required to complete a training task before each actual task. The tasks that were included in the usability evaluation are listed in Table 1.

**Table 1 Tasks included in the Usability Evaluation**

| TASK # | DESCRIPTION |
|---|---|
| 1 | The Overview |
| 2 | Data Manipulation: Rename a File |
| 3 | Data Manipulation: Add a File |
| 4 | Data Manipulation: Delete a File |
| 5 | The Tag Cloud |
| 6 | Search |
| 7 | Filter |
| 8 | Semantic Zooming |
| 9 | Browsing |

*5) Procedure*

The usability evaluation took place in the PhD Lab of the NMMU Department of Computing Sciences using a desktop computer with the Windows 8.1 operating system. The prototype was run in the Google Chrome web browser. Prior to the commencement of the usability evaluation, each participant completed an informed consent form to take part in the evaluation. The participant then completed an electronic background questionnaire. The evaluation procedure was then explained and an introduction to the UI of MyPSI was provided. The participant then completed the tasks in the task list, followed by the electronic post-test questionnaire.

*B. Effectiveness Results*

The participants were required to answer a question relating to each task. Additionally, participant observation was used to identify any issues that the participants might have had during the completion of any of the tasks. A rating of 0-2 was assigned to each task with notes to determine with which interactions the participants had problems. The percentage of correct answers, together with the ratings from the observations, was used to determine the effectiveness performance metric.

Most tasks received 100% task success except for Task 3 (90%) and Task 5 (80%). Task 3 required participants to add a file to a specific folder and answer how many documents were contained in that folder. Participant observation revealed that one participant identified the documents in the folder but answered the question incorrectly. Task 5 involved clicking on a specific tag in the *Tag Cloud* and hovering over the file related to this tag to determine the number of tags related to this file. Both participants who answered this task answered incorrectly similarly to the training task, which required participants to determine how many files were assigned to a specific tag.

According to the participant observations ratings for each task, Task 3 and 5 received a similar task success rating of 90% and 80% respectively, with the addition of Task 2 (90%). Task 2 required participants to select the label of a folder to rename the folder. Although one of the participants answered correctly, s/he clicked on another folder's label and became confused thereafter.

*C. Satisfaction Results*

The satisfaction results were divided into several sub-sections, namely cognitive load, overall satisfaction, usability and the added sections relating to each visualisation technique. A section was also provided to identify the most positive and negative aspect(s) of the prototype, and also to allow any general comments.

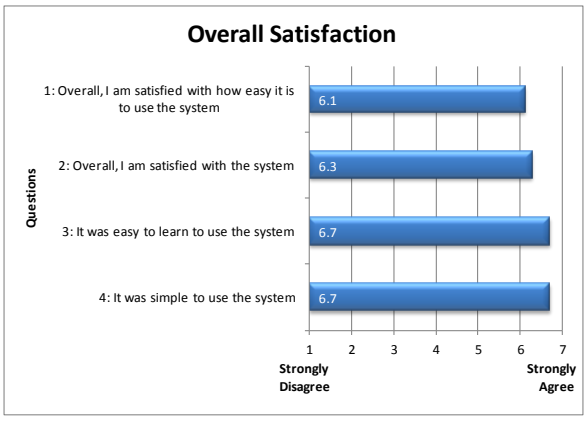



**Figure 5 Overall Satisfaction Ratings using a 7-point Likert Scale**

The mean ratings were all positive (< 2) for cognitive load, which indicated that the required mental workload was low. The mean overall satisfaction ratings were all rated

highly (> 6) as shown in Figure 5. The overall satisfaction received the highest ratings from the participants. Participants found the MyPSI prototype easy to use, easy to learn and simple.

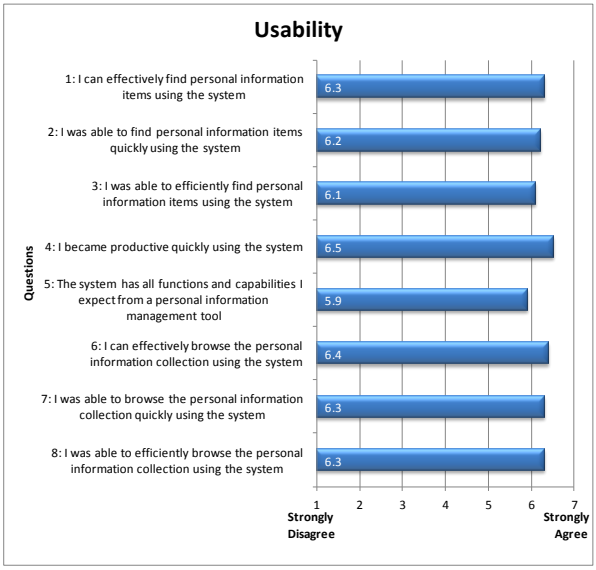



**Figure 6 Usability Ratings using a 7-point Likert Scale**

The mean usability ratings for the MyPSI prototype were also high (> 5.5) (Figure 6). Participants perceived that they became productive quickly using the MyPSI prototype and that they could easily find items and browse the system.

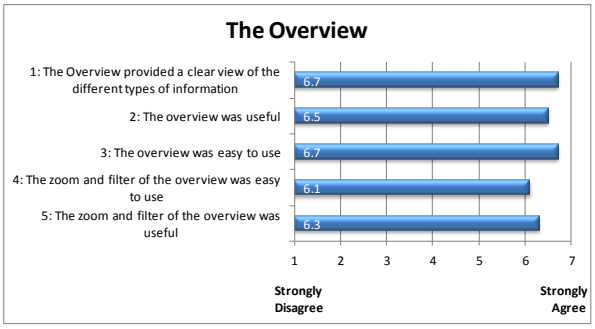



**Figure 7 Mean Ratings for the Overview using a 7-point Likert Scale**

The *Overview* visualisation technique also received high ratings (> 6) (Figure 7). Participants found that the *Overview* was useful, easy to use and provided a clear view of the different types of PI.

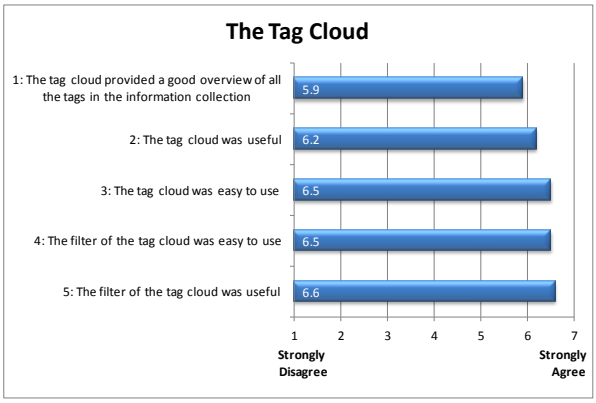



**Figure 8 Mean Ratings for the Tag Cloud using a 7-point Likert Scale**

The *Tag Cloud* also received high ratings for the questions relating to this visualisation technique (> 5.5), as shown in Figure 8. The participants found the *Tag Cloud* and *Tag Cloud* filter easy to use and useful.

The mean ratings for the *Partition Layout* also received high ratings (> 6) (Figure 9). Participants found the *Partition Layout* useful and easy to use. Participants also found that the layout provided a good view of the information in the different collections.

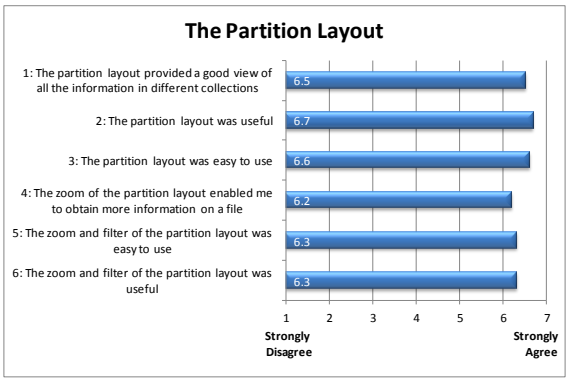



**Figure 9 Mean Ratings for the Partition Layout using a 7-point Likert Scale**

Participants were asked to note the most positive and negative aspect(s) of the MyPSI prototype. Several participants (n=5) indicated the *Overview* as the most positive aspect of MyPSI as it was intuitive and easy to use. Participants also had positive responses to the UI design layout (n=4) and found the features that were provided, such as filtering, to be useful (n=4). Three participants identified the *Partition Layout* as the most positive aspect.

The most negative aspect was related to the small file icons (n=4). A second negative aspect was that the *Tag Cloud* was difficult to use when there were many tags as these could overlap (n=2). Two participants suggested differentiating results that were part of a search or filter from the rest of the information items more obviously.

A section was also provided in the post-test questionnaire to allow participants to note any general comments or suggestions. Five participants noted that MyPSI was a good prototype, which they would consider using in the future. General suggestions related to the differentiation of results as described above and to allowing a user to maximize a folder to support easier browsing.

### *D. Discussion*

While MyPSI is in the early stages of development, participants agreed that the visualisation techniques provided a good view of PI across multiple devices. Participants also found the visualisation techniques to be effective, easy to use and useful. Although these results are promising, a few usability problems were identified. The problems that need to be addressed include the following:

- Modify the icon scaling according to the file's parent folder to account for small file icons;
- Differentiate more obviously between the files and folders that result from a filter, from those that are not part of the results;
- Improve on the layout, positioning and tags visualised by the *Tag Cloud* to make it easier to read;
- Look at an alternative method to maximise a folder when browsing.

Limitations of this evaluation include the convenience sample and the small sample size (n=10) that was used for

the study. An additional limitation was due to the sample PSI that was used. In future, the prototype will be tested using a field study with the users' own data.

## VI. Conclusions And Future Work

This paper identified that a need exists for novel visualisation techniques to support PIM across different devices. Three novel visualisation techniques were designed, namely an *Overview*, a *Tag Cloud* and a *Partition Layout*. The MyPSI prototype was developed to allow demonstration and evaluation of the proposed techniques.

The results of a usability study yielded highly positive results. Effectiveness ratings showed that participants could easily interact with the visualisation techniques. The prototype received positive ratings for cognitive load, overall satisfaction and usability. Participants perceived the system to be simple, easy to use and easy to learn. The participants also found the *Overview*, *Tag Cloud* and *Partition Layout* easy to use and useful. Only a few usability issues were found with the MyPSI prototype, which will be addressed. Future work will involve the design of a touch-based interface to support PIM on mobile devices, which will be evaluated using a more suitable representative sample.

## VII. Acknowledgements

Acknowledgements are due to the NMMU/Telkom Centre of Excellence for funding this research.

## VIII. References

[1] W. Jones and H. Bruce, "A Report on the NSF-Sponsored Workshop on Personal Information Management," Seattle, WA, 2005.

[2] J. Aires and D. Gonçalves, "Personal Information Dashboard - Me, At a Glance," in *Proceedings of the PIM 2012 Workshop*, 2012, pp. 1–8.

[3] M. Golemati, A. Katifori, E. G. Giannopoulou, I. Daradimos, and C. Vassilakis, "Evaluating the Significance of the Windows Explorer Visualization in Personal Information Management Browsing Tasks," in *Proceedings of IV'07*, 2007, pp. 93–100.

[4] S. Beets and J. Wesson, "Managing Personal Information across Multiple Devices: Challenges and Opportunities," in *Proceedings of IFIP INTERACT 2013*, 2013, pp. 1–8.

[5] M. Tungare, "Understanding the Evolution of Users' Personal Information Management Practices," in *Proceedings of INTERACT 2007 Doctoral Consortium*, 2007, pp. 586–591.

[6] D. K. Barreau, "Context as a Factor in Personal Information Management Systems," *Journal of the American Society for Information Science*, vol. 46, no. 5, pp. 327–339, 1995.

[7] S. Fertig, E. Freeman, and D. Gelernter, "LifeStreams: An Alternative to the Desktop Metaphor," in *Proceedings of CHI 1996*, 1996, pp. 410 – 411.

[8] S. Dumais, E. Cutrell, J. J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins, "Stuff I've Seen: A System for Personal Information Retrieval and Re-Use," in *Proceedings of SIGIR 2003*, 2003, pp. 72–79.

[9] H. C. Jetter, J. Gerken, W. A. König, and H. Reiterer, "ZOIL - A Cross-Platform User Interface Paradigm for Personal Information Management," in *Proceedings of PIM'08*, 2008, pp. 1–9.

[10] M. R. Al Nasar, M. Mohd, and N. M. Ali, "Personal Information Management Systems and Interfaces: An Overview," in *Proceedings of STAIR 2011*, 2011, pp. 197–202.

[11] E. Cutrell, D. C. Robbins, S. T. Dumais, and R. Sarin, "Fast, Flexible Filtering with Phlat — Personal Search and Organization Made Easy," in *Proceedings of CHI 2006*, 2006, pp. 261–270.

[12] J. Indratmo and J. Vassileva, "A Review of Organizational Structures of Personal Information Management," *Journal of Digital Information*, vol. 9, no. 1, pp. 1–19, 2008.

[13] M. Graham and J. Kennedy, "A Survey of Multiple Tree Visualisation," *Information Visualization*, vol. 9, no. 4, pp. 235–252, 2010.

[14] M. Graham, J. B. Kennedy, and C. Hand, "A Comparison of Set-Based and Graph-Based Visualisations of Overlapping Classification Hierarchies," in *Proceedings of AVI'00*, 2000, pp. 41–50.

[15] S. Beets and J. Wesson, "Designing a Tool to Support the Interactive Visualisation of Personal Information across Multiple Devices," in *Proceedings of SATNAC 2013*, 2013, pp. 1–6.

[16] J. Heer, M. Bostock, and V. Ogievetsky, "A Tour Through the Visualization Zoo," *Communications of the ACM*, vol. 53, no. 6, pp. 59–67, 2010.

[17] J. Sinclair and M. Cardew-Hall, "The Folksonomy Tag Cloud: When is it Useful?," *Journal of Information Science*, vol. 34, no. 1, pp. 15–29, 2008.

[18] M. Bostock, V. Ogievetsky, and J. Heer, "$D^3$ Data-Driven Documents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2301 – 2309, 2011.

[19] J. Bernhard and N. V. Kelso, "Color Design for the Color Vision Impaired," *Cartographic Perspectives*, no. 58, pp. 61–67, 2007.

[20] A. Perer and B. Shneiderman, "Balancing Systematic and Flexible Exploration of Social Networks," *IEEE Transactions on Visualization and Computer Graphics (InfoVis)*, vol. 12, no. 5, pp. 693–700, 2006.

[21] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Result of Empirical and Theoretical Research," *Human Mental Workload*, vol. 1, pp. 239–250, 1988.

[22] J. Lewis, "IBM Computer Usability Satisfaction Questionnaire: Psychometric Evaluation and Instructions for Use," *International Journal of Human-Computer Interaction*, vol. 7, no. 1, pp. 57–78, 1995.

**Simone Beets** received her Master of Science degree (Cum Laude) in 2011 from the Nelson Mandela Metropolitan University. She is presently studying towards her PhD degree at the same institution and is currently in her final year. Her research interests include personal information management and information visualisation.

# A Comparison of Machine Learning Techniques for Hand Shape Recognition

Roland G. Foster,[1] Mehrdad Ghaziasgar[1], James Connan[2] and Reg Dodds[1]
[1]Department of Computer Science
University of the Western Cape, Private Bag X17, 7535 Bellville, South Africa
Telephone: +(27) 21 959–3010, Fax: +(27) 21 959–3006
and [2]Department of Computer Science
Rhodes University, P O Box 94, Grahamstown, 6140
Tel: + (27) 46 603-8291, Fax: + (27) 46 636-1915
Email: [1]2916282@myuwc.ac.za; [1]mghaziasgar@uwc.ac.za; [2]j.connan@ru.ac.za; [1]rdodds@uwc.ac.za

***Abstract*—The SASL group of the University of the Western Cape aims to create a machine translation system that can translate full phrases and sentences between South African Sign Language (SASL) and English to bridge the gap in communication between the Deaf and hearing. There are five parameters—hand location, hand orientation, hand motion, hand shapes and facial expressions—which are fundamental to the recognition of sign language gestures. To date, the SASL group has developed systems to recognize these parameters at a high accuracy. Classification of data is carried out using Support Vector Machines by the majority of these systems.**

**This research compares three machine learning techniques—$k$-Nearest Neighbours, Support Vector Machines and Random Forests—in the context of SASL hand shape recognition. An experiment was carried out to compare the recognition accuracy of the three techniques for ten SASL hand shapes and yielded a 76% accuracy for Support Vector Machines, 72% for random forests and 64% for $k$-nearest neighbours.**

***Index Terms*—Hand Shape Recognition, Sign Language, Machine Learning**

## I. Introduction

The South African Sign Language (SASL) group at the University of the Western Cape is in the process of creating a system for automatic translation between SASL and English [1]. One aim of the system is to be able translate a recorded SASL video of a Deaf signer into spoken English. Another aim is to translate spoken English into SASL animation. These two processes are related but distinct since they require varied different processing techniques. Combining both processes enables communication between Deaf users and the English speaking people.

A system of this type can be invaluable in various contexts, such as the interaction between a doctor and Deaf patient. There is a communication gap between the hearing doctor and the Deaf patient, as the physician does not understand SASL and the Deaf patient does not understand English. One option is to have a translator present although, where health matters are concerned, the Deaf patient may prefer to keep information confidential. The SASL system can thus be used to bridge the gap in communication between the Deaf patient and English speaking doctor. Many other potential applications also exist.

This research concerns itself with a segment of the process of translation of a SASL video to spoken English. The initial task involved in this process is to recognize SASL gestures in a SASL video. There are five fundamental parameters which need to be extracted from SASL video to recognize a sign language gesture. These include hand location, hand motion, hand orientation, hand shape and facial expressions.

The SASL group has created systems to recognize hand shapes [2], hand location [3], hand motion [4], hand orientation [2] and facial expressions [5]. The majority of these systems use Support Vector Machines (SVM) to classify data. This machine learning technique has proven to achieve very accurate and reliable results in these systems [2]. It is, however, unknown how other machine learning techniques such as Random Forests (RF) and $k$-Nearest Neighbours (kNNs) compare to this technique in the context of sign language parameter recognition.

This research aims to compare the accuracy of SVMs, RFs and kNNs in the context of SASL hand shape recognition. Ultimately, the comparison of these machine learning techniques has to be carried out for all the SASL parameter recognition systems developed by the group. This research lays the foundation by starting with the hand shape recognition system. This research proposes to implement the hand shape feature extraction method developed by Li [2], and implement the above mentioned machine learning techniques on the extracted features.

This paper is organized as follows: Section II—*Related Work*—discusses the hand shape recognition framework developed by Li and analyses studies that have carried comparisons of machine learning techniques in the context of image processing. Section III—*Feature Extraction Procedure*—discusses the methods and image processing techniques used to extract the hand shape features in detail. Section IV—*Machine Learning Techniques*—briefly discusses the three machine learning techniques which are implemented and compared in this paper. Section V—*SASL Hand Shape Data Set*—describes the data set collected for use in this research. Section VI—*Experimental Setup and Analysis of Results* explains the experimental setup and results obtained, with an analysis of

the results. Section VII concludes the paper.

## II. Related Work

The first subsection describes the work of Li [2], a state-of-the-art SASL hand shape recognition system, which is the basis of the feature extraction procedure in this research. The second and third subsections are studies that compare the use of machine learning techniques in an image processing context.

### A. Li's Work

Li, a former student of the SASL group at the University of the Western Cape, developed a system to recognize ten SASL hand shapes in real-time [2]. The system takes in a live video stream of a signer's upper or entire body and continuously recognizes SASL hand shapes performed by the signer in real-time.

The system detects the face of the signer in the initial video frame using Haar-like features. Once the face has been detected, the position of the nose is determined by isolating the centre of the facial frame. The skin colour distribution of the detected nose is computed and used to highlight the skin pixels of the signer in every frame of the video sequence thereafter. In order to achieve skin highlighting, histogram back projection is applied using the skin colour distribution. Gaussian Mixture Model background subtraction is used to separate the background and foreground of the image. Doing this ensures that only the moving parts, in this case the signer's hands, are present in the image. The hand is located in the resulting image using Hierarchical Chamfer matching only once on the initial frame. This is used to initialize the CAMShift tracking algorithm, which continuously tracks the located hand.

Rotations of the hand are normalized by aligning the hand region to the vertical axis. Connected Components Analysis (CCA) is used to highlight the contour of the hand region in every frame, and the contour image is resized to a resolution of $20 \times 30$ pixels. The resulting image is used as a feature vector for the hand shape recognition process.

A SVM was used to classify the SASL hand shapes. The SVM was trained to recognize ten SASL hand shapes. The system proved to be very accurate, achieving an accuracy of 81% across all hand shapes.

### B. Kadous's Work

Kadous compared the classification accuracy of two machine learning techniques in the context of Australian sign language (Auslan) gesture recognition [6]. Tracking of the hand was accomplished by making use of the Nintendo PowerGlove. The PowerGlove shown in Figure 1 is a video game hardware accessory of the Nintendo gaming system. The glove provides a set of three attributes in order for feature extraction to take place: the relative coordinates of the glove, the degree of rotation of the wrist and the degree to which each finger is bent.

The $x$, $y$ and $z$ positions are given relative to the point of synchronization. The degree of rotation of the wrist is given in $30^\circ$ increments. The degree to which each of the first four fingers are bent is provided, as well as four possible values for each finger.


Fig. 1. The Nintendo PowerGlove from the Nintendo gaming system [6].

Two machine learning techniques were compared for the recognition of gestures: instance-based learning and decision tree building.

For the experimental setup, 95 Auslan signs were selected for five signers to test. Between 8 and 20 samples were obtained from each of the signers for each of the 95 signs. In total 6650 signs were collected and these were used to compare the two machine learning techniques. The collected data was divided into five equally sized sets. Four of the sets were used for training both machine learning techniques and the last set was used to test them. The Instance-based learning technique achieved an accuracy of 80% and the C4.5 implementation of a decision tree builder achieved a significantly lower accuracy of 55%. This clearly demonstrates that an investigation to determine the optimum machine learning technique is required.

### C. Nitze et al.'s Work

Nitze *et al.* compared four machine learning techniques in the classification and recognition of agricultural crop types [7]. The four machine learning techniques were: Artificial Neural Networks (ANNs), Maximum Likelihood (ML), Random Forests (RFs) and Support Vector Machines (SVMs). The study area, located in Indian Head, south-eastern

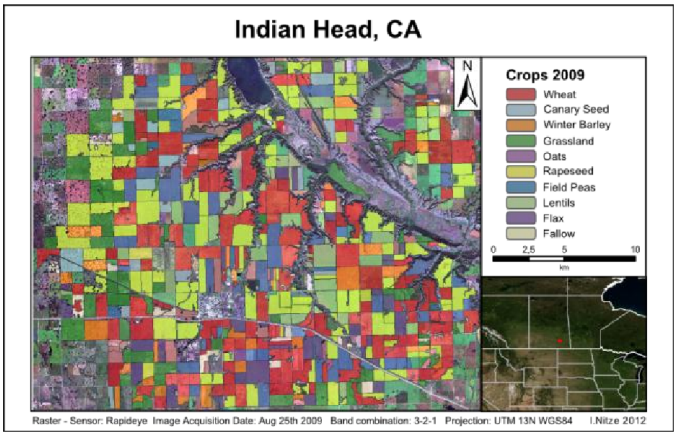



Fig. 2. Overview of the study area [7].

Saskatchewan in Canada, and the various types of crops which were classified are shown in Figure 2. Ten different types of crops were considered for recognition, namely: Wheat, Barley, Rapeseed, Oats, Field Peas, Lentils, Canary Seed, Flax,

Grassland and Fallow. A dataset of 500 crop fields were used as input. The four machine learning techniques were trained and tested using the WEKA Experimenter 3.6, a data mining library with a collection of machine learning techniques.

The mean accuracy achieved for the classification of crop types was 88.1% for SVMs, 87.4% for RFs, 87.1% for ANNs and 78.9% for ML. The average time taken to train each of the machine learning techniques on an image was recorded. ANN took an average of 15.1 seconds, RF took an average of 6 seconds, SVM took an average of 0.3 seconds and ML, which was the quickest to train, took an average of 0.004 seconds, to train on each image.

The time taken to classify a crop type for each machine learning technique was 0.003 seconds for ANN, 0.083 seconds for RF, 0.039 seconds for SVM and 0.017 seconds for ML.

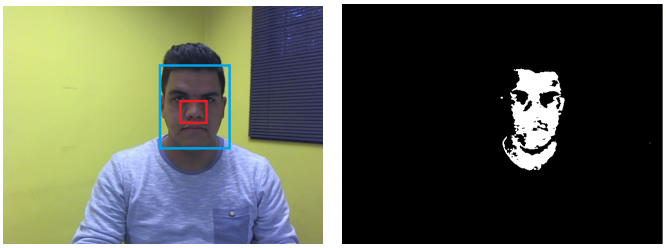
In this study, SVMs were shown to achieve the highest accuracy. It is unknown, however, whether this is also the case in a hand shape recognition context. Also, while a comparison of training time and classification time is also important, this is left for future work. This paper limits itself to a comparison of classification accuracy.

## III. Feature Extraction Procedure

This section describes the feature extraction procedure adapted from Li and used in this research. The image processing techniques that are used to extract the contour of the signer's hand are described in detail.

### A. Adaptive Skin Detection

In a South African context, a recognition system must be invariant or tolerant to variations in skin colour. This is achieved by locating the nose of the signer, and extracting a colour distribution of the nose, which has been shown to be well-representative of the signer's skin colour [3]. The Viola-Jones face detector [8] is used to locate the face of the signer. This region is set as the region of interest (ROI) in the image. The centre of the facial region is computed, and has been shown to lie on or very close to the tip of the nose [3]. The colour distribution of a $10 \times 10$-pixel region around the nose is computed and taken as the skin colour distribution of the signer.



(a) The detected face and nose of the signer

(b) The back-projected image

Fig. 3. The detected face and nose of the signer, and the corresponding back-projected image.

### B. Histogram Back-Projection

The skin colour distribution is represented as a histogram. The histogram is back-projected onto the original frame to highlight all pixels which match the skin colour distribution, as shown in Figure 3(b). This is known as the skin image.

### C. Background Subtraction

Gaussian Mixture Models [9] are used to subtract the background from the foreground, exposing only the moving parts in the image which, in this case, is the user's hand. This is known as the motion image.

### D. Combination of the Skin and Motion Images

To eliminate any extra noise in the image, the skin image and motion image are combined using a logical AND operation to form a single image. This image contains only skin pixels that are moving. In this case, this results in an image clearly highlighting the moving hand of the signer. Considering that the signer's face remains approximately stationary, this procedure also helps to eliminate tracking losses when the hand moves close to or over the face. The resulting image is called the moving skin image.

### E. Hierarchical Chamfer Matching

Hierarchical Chamfer matching [10] is performed using a template silhouette of an open hand to match with the combined skin and motion image. Once the hand is located, it is set as the ROI. Hierarchical Chamfer matching is only performed once to initialize the CAMShift tracking algorithm. This ensures that the system can operate at real-time processing speed.

### F. CAMShift tracking

CAMShift tracking is used to continuously track the hand as it moves in the video feed. Every frame is pre-processed to obtain the moving skin image, to which CAMShift is then applied. This ensures accurate tracking results and few tracking losses. Given that the moving hand is always the largest object in the moving skin image, it dominates the tracking window over other smaller sources of noise.

CAMSHift, or Continuously Adaptive Mean Shift, is an improved version of the Meanshift algorithm [11]. Meanshift uses a static region sizing step whereas its improved version—CAMShift—uses an adaptive one. The algorithm is dynamic and real-time and can be used to achieve very accurate tracking results if sources of noise are sufficiently reduced. Figure 4 depicts the CAMShift algorithm in operation.

### G. Feature Extraction

The contour of the hand is extracted and used as the feature vector to three machine learning techniques: SVM, RF and kNN. The contour of the hand is computed by applying Freeman's Connected Components Analysis (CCA) methodology [12]. While the algorithm detects many small-scale contour features, the largest contour is taken to represent the overall contour of the hand.

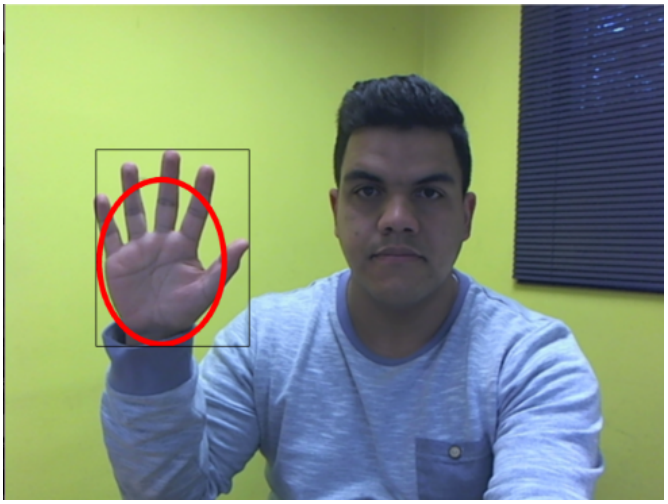


Fig. 4. An ellipse around the tracked hand using the CAMShift tracking algorithm.

This yields an image with the true contour of the signer's hand. Normalization is performed on the resulting image to avoid misalignment invariance by first forming an oriented minimum bounding box around the hand contour as shown in Figure 5(a). Subsequently, the box is aligned to the vertical axis to eliminate rotation variations [2] as shown in Figure 5(b). It is then scaled down to the resolution of $30 \times 40$ pixels shown in Figure 5(c) to reduce the dimensionality of the feature vector. Scaling down the image is crucial to reducing the computational cost of classification, and increases the overall recognition accuracy of the machine learning techniques. The

(a) (b) (c)

Fig. 5. The process of extraction and normalization of the hand contour. (a) Original tilted hand. (b) Contour of hand *after* rotation. (c) Hand contour *after* resizing to complete normalization

final extracted and normalized hand contour image shown in Figure 5(c) is written to a feature data file and used for training and testing.

## IV. Machine Learning Techniques

### A. $k$-Nearest Neighbours

$k$-nearest neighbours is a supervised, non-parametric machine learning technique. It is said to be non-parametric due to the fact that it does not make any assumptions about the data distribution. It caches all the training data samples and predicts the output of a new sample by analyzing $k$ of the nearest neighbours of the sample by means of voting or calculating the weighted sum [13]. The $k$ chosen for training can be optimized to achieve optimal classification accuracy. Trial and error showed that a value of $k = 1$ provided the highest accuracy.

### B. Random forests

Random Forests are a group of decision trees which form a forest [14]. Each group of trees votes for a class of some sample data and the class which receives the most votes from the groups of trees is said to be the predicted result. A maximum number of 230 trees was used for the random forest.

### C. Support vector machines

Support Vector Machines (SVMs) are a group of supervised learning models which possess learning algorithms that analyse data and recognize patterns for classification and regression analysis [14]. SVMs come with several choices of kernel function, each of which may be suited to specific applications. Li showed that the Radial Basis Kernel achieves the optimum accuracy for SASL-hand-shape recognition. Therefore, the Radial Basis kernel was chosen in this instance.

## V. SASL Hand Shape Data Set

This section describes the data set, as well as the chosen training and testing sets. Ten distinct SASL hand shapes were chosen for the training and testing of the three machine learning techniques. In Figure 6, these ten hand shapes are shown.

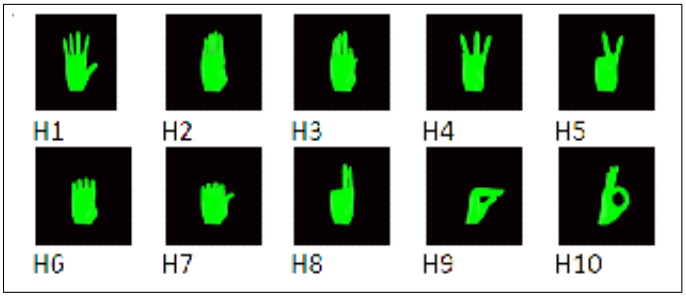



Fig. 6. The ten selected SASL hand shapes.

### A. Collection of Data

In order to collect training and testing data, six subjects of varied skin colour were selected. Each subject was asked to perform each hand shape, and hold it for no less than 500 frames—between 15 to 20 seconds. This process was repeated to gather ten videos, one per hand shape, for each subject. Of each video, 500 random frames were selected. In total, the dataset therefore consisted of 30000 randomly chosen samples: 3000 samples per hand shape and 5000 samples per subject.

A Logitech C920 web camera at a resolution of $640 \times 480$ pixels was used to gather the images for the dataset.

### B. The training set

A total of 400 randomly selected samples—40 samples per hand shape class—from three of the six subjects were used to train each of the three machine learning techniques.

### C. The testing set

A total of 540 randomly selected samples—54 samples per hand shape—from the three remaining subjects were used to test the recognition accuracy of the machine learning techniques.

## VI. Experimental Setup and Analysis of Results

Fig. 7. The scene of testing

The aim of the experiment was to determine and compare the recognition accuracy of the three machine learning techniques for the ten SASL hand shapes. Figure 7 depicts the scene in which the training and testing took place. The system used in experimentation was a desktop computer with an Intel Core i5-3570 3.4GHz Quad-Core Processor and 16GB RAM, running the Ubuntu Linux 13 Operating System.

Each machine learning technique was trained on labelled samples of the training set. Thereafter, unlabelled and unseen samples of the testing set were used as input to each trained classifier. In each case, it was recorded whether or not the sample was correctly classified. This made it possible to determine a classification accuracy per hand shape for each machine learning method, summarized in Table I. Maximum and minimum accuracies for each method have been highlighted in the table. The table also provides the mean accuracy and the standard deviation of the accuracy associated with each machine learning technique.

Referring to the column of $k$-NNs in Table I, it is clear that Hand Shape 10 achieves the overall lowest accuracy of 41%. Hand Shape 2 achieves the highest accuracy of 91%. The average accuracy across all of the SASL hand shapes using the $k$-Nearest Neighbours machine learning technique is $64.4 \pm 13.4\%$.

For the Random Forests implementation, as shown in Table I in the column for the RFs, it can be seen that Hand Shape 7 achieves the highest accuracy of 89%. Hand Shape 2 achieves the lowest recognition accuracy of 50%. The average accuracy across all the SASL hand shapes using Random Forest learning is $71.6 \pm 13.5\%$.

It is interesting to note that the technique achieves a *higher* accuracy than $k$-NNs at almost the same standard deviation.

TABLE I
Recognition accuracy results for $k$-Nearest Neighbours ($k$-NNs), Random Forests (RFs) and Support Vector Machines (SVMs).

| Hand Shape | Recognition Accuracy (%) | | |
|---|---|---|---|
| | $k$-NNs | RFs | SVMs |
| 1 | 76 | 87 | **93** |
| 2 | **91** | **50** | 72 |
| 3 | 61 | 72 | 81 |
| 4 | 61 | 74 | 65 |
| 5 | 72 | 70 | 80 |
| 6 | 55 | 78 | 87 |
| 7 | 67 | **89** | 76 |
| 8 | 61 | 63 | 81 |
| 9 | 59 | 57 | 70 |
| 10 | **41** | 65 | **63** |
| Mean | 64.4 | 71.6 | 76.2 |
| Std. Dev. | 13.4 | 13.5 | 10.0 |

Furthermore, it can be observed that the highest and lowest accuracies achieved by the two methods are by completely different hand shapes. This indicates that no specific hand shape was inherently problematic.

For SVMs, Hand Shape 1 achieves the highest recognition accuracy of 93% and Hand Shape 10 achieves the lowest recognition accuracy of 63%. The hand shape that achieves the highest accuracy is different from both previous techniques, but the hand shape that achieves the lowest accuracy in this case is the same as one of the lowest performing hand shapes in $k$-NNs. Hand Shapes 5 and 7 appear to perform at a relatively high accuracy for all three methods, achieving above 60% accuracy. Hand shape 10 performs relatively poorly for $k$-NNs and RFs, but exceeds 60% accuracy for SVMs.

Overall, there does not appear to be any trend in accuracy across hand shapes for the three methods. The accuracies obtained per hand shape clearly appear to be specific to the machine learning technique, and not dependent on the actual hand shape.

Overall, SVMs achieve the highest average recognition accuracy of $76.2 \pm 10.0\%$ across all ten SASL hand shapes. This is a much higher accuracy than both previous methods, at a lower standard deviation. It appears that SVMs are, in fact, an optimum choice for high-accuracy SASL hand shape recognition.

## VII. Conclusion

This paper investigated and compared the classification accuracy of three machine learning techniques—Support Vector Machines, $k$-Nearest Neighbours and Random Forests—in the context of South African Sign Language hand shape recognition, for ten SASL hand shapes.

The results showed that the machine learning technique that achieves the overall highest average recognition rate is the SVM, with an average accuracy of $76.2 \pm 10.0\%$. Therefore, it can currently be concluded that Support Vector Machines are more suited to the task of SASL hand shape recognition than $k$-Nearest Neighbours and Random Forests are.

Future work for this research will look to optimize the parameters of each machine learning technique to improve

the recognition accuracy. SVMs can use several kernels, and each kernel has a set of parameters that can be optimized. $k$-NNs can be optimized in terms of the value of $k$. While trial and error showed that a value of $k = 1$ was optimal, further investigation is required. The number of trees in the RF can also affect the classification accuracy of the RF.

Furthermore, the use of other popular machine learning techniques such as Artificial Neural Networks and Naive Bayesian Classifiers will also be investigated.

Once the optimization of these machine learning techniques is achieved, they can be used in other recognition systems of the SASL group to recognize hand location, hand motion, hand orientation and facial expressions.

## ACKNOWLEDGEMENTS

The authors would like to thank Telkom, Cisco, THRIP and Aria Technologies for generous funding towards the Centre of Excellence at the Department of Computer Science at the University of the Western Cape and making this research possible.

## REFERENCES

[1] J. Connan, "Integration of signed and verbal communication: South African Sign Language recognition and animation," 2011, http://www.coe.uwc.ac.za/index.php/SASL.html.

[2] P. Li, M. Ghaziasgar, and J. Connan, "Hand shape estimation by 2D appearance and 3D animation for sign-language," in *Proceedings South African Telecommunication Networks and Applications Conference (SATNAC 2011)*, East London, Eastern Cape, South Africa, 2011, pp. 409–414.

[3] I. Achmed and J. Connan, "Upper body pose recognition and estimation towards the translation of South African Sign Language," in *Proceedings South African Telecommunication Networks and Applications Conference (SATNAC 2010)*, Spier, Stellenbosch, South Africa, 2010.

[4] I. Achmed, I. Venter, and P. Eisert, "A framework for independent hand tracking in unconstrained environments," in *Proceedings of the Southern Africa Telecommunication Networks and Applications Conference (SATNAC 2012)*, Fancourt, South Africa, 2012.

[5] J. Whitehill, "Automatic real-time facial expression recognition for signed language translation," Master's thesis, Department of Computer Science, University of the Western Cape, 2006.

[6] M. Kadous, "Machine recognition of Auslan signs using powergloves: towards large-lexicon recognition of sign language," Master's thesis, University of South Wales, Australia, 2010.

[7] I. Nitze, U. Schulthess, and H. Asche, "Comparison of machine learning algorithms random forest, artificial neural network and support vector machine to maximum likelihood for supervised crop type classification," in *Proceedings of the 4th Geobia*, Rio de Janeiro, Brazil, 2012.

[8] P. Viola and M. J. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of IEEE Computer Society's Computer Vision and Pattern Recognition (CVPR 2001)*, vol. 1, 2001.

[9] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *Proceedings of the 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, 1999.

[10] G. Borgefors, "An improved version of the chamfer matching algorithm," in *Proceedings of the 7th International Conference on Pattern Recognition*, vol. 2, Montreal, Canada, 1984, pp. 1175–1177.

[11] G. Bradski, "Real time face and object tracking as a component of a perceptual user interface," in *Fourth IEEE Workshop on the Applications of Computer Vision*, 1998, pp. 214–219.

[12] H. Freeman, "Computer processing of line-drawing images," *ACM Computing Surveys*, vol. 6, no. 1, pp. 57–97, 1974.

[13] N. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[14] G. Bradski and A. Kaehler, *Learning OpenCV: Computer Vision with the OpenCVlibrary*. O'Reilly Media, 2008.

**Roland Foster** received his B.Sc. Honours in 2012 from the University of the Western Cape and is presently studying towards his Master of Science degree at the same institution. His research interests include image processing, sign language recognition and hand shape recognition.

**Mehrdad Ghaziasgar** is the project manager of the South African Sign Language (SASL) research group. He has a wide range of interests that include internet programming, mobile computing, computer vision and machine learning.

**James Connan** heads up the South African Sign Language (SASL) research group. His interests are computer vision and machine learning.

**Reg Dodds** is a lecturer at the University of the Western Cape. His interests include symbolic symbol systems, machine learning, algorithms and their complexity and the theory of computation.

# Spam Email Classification with Generalized Additive Neural Networks using Ensemble Methods

Pieter Labuschagne†, Jan V. du Toit‡
School of Computer, Statistical and Mathematical Sciences
North-West University, Potchefstroom Campus
Private Bag X6001, Potchefstroom, 2520, South Africa
Tel: +27 18 2992548, Fax: +27 18 2992570
email: {21269777†, 10789901‡}@nwu.ac.za
†Corresponding author

**Abstract - Solutions to minimize the amount of junk mail we receive are not yet optimal and thus research towards an improved spam email classifier is ongoing. Traditional content and list-based filters are less effective than machine learning approaches, since continuous maintenance and human input are necessary. In this paper an automated construction algorithm for a Generalized Additive Neural Network (GANN) is applied to the Enron and PU1 email corpora and compared to the popular Naïve Bayesian filter. In an attempt to further enhance the GANN performance, both Bagging and Boosting ensemble methods were applied. It was found that both ensemble methods improved results making the GANN the better choice.**

**Index Terms – Classification, Ensemble, GANN, Spam**

## I. Introduction

Email is an application that was developed by Ray Tomlinson in 1971 (Tomlinson, n.d.) which changed the way people communicated. It evolved from a primitive service used by few into a prevalent communication medium used by many people on a daily basis. Currently, the majority of emails sent promote some sort of propaganda that varies greatly in content and structure. When recipients receive unsolicited bulk email messages that are irrelevant or purposely trying to mislead them it is called spam (Cranor & LaMacchia, 1998). Email messages with the opposite properties of spam are sometimes referred to as ham. Spammers, the people who send the above mentioned unwanted messages, cause annoyance to other email users by exploiting the service's immediacy, affordability and convenience.

This study will try to improve results obtained by Metsis, Androutsopoulos & Paliouras (2006) and Androutsopoulos, Koutsias, Chandrinos & Spyropoulos (2000) with the Naïve Bayesian technique applied to the Enron and PU1 corpora respectively. Since the GANN has shown promise on other classification tasks, this method is employed on the two corpora. Additionally, two ensemble techniques are applied to the results obtained by the GANN to possibly improve them further. This will provide some insight into whether ensemble techniques can improve results produced by a GANN.

The rest of the paper is organized as follows. Section II highlights some of the issues caused by spam that affect recipients and briefly discusses different types of filters. The need for new and improved spam filters will also be emphasized. In Section III the Multilayer perceptron neural network that forms part of the GANN architecture is discussed. Section IV will focus on the GANN that represents the spam filter under observation. The ensemble methods applied for potential classification improvements are discussed in Section V. Sections VI and VII will explain the experimental design for each experiment performed and provide an analysis of the results obtained through the use of evaluation measures. Finally, some conclusions are presented in the last section.

## II. Spam Email

According to Gudkova (2012); Gudkova (2013) the percentage of spam in global email traffic has decreased over the past four years. Table 1 illustrates that the overall increased level of anti-spam protection applied throughout the last four years are indeed paying off. The last column represents the decrease of spam in global email traffic compared to the previous year.

| Year | Spam in Global Email Traffic | Decrease of Spam in Global Email Traffic |
|---|---|---|
| 2010 | 82.2% | 3.0% |
| 2011 | 80.3% | 1.9% |
| 2012 | 72.1% | 8.2% |
| 2013 | 69.6% | 2.5% |

Table 1: Yearly decrease of spam in global email traffic

The major decrease of 8.2% in 2012 is due to spam filters that are in place on most email systems. Although spam filters are getting better, spammers are adapting

their methods by making use of different techniques such as content obscuring to disguise their spam messages as legitimate and safe (Guzella & Caminhas, 2009). With continuous attempts by spammers to outsmart spam filters it is possible for these unwanted messages to reach the recipients' inboxes despite the efforts made by the spam filter. When spam emails pass through the spam filter and are misclassified as legitimate messages it could affect the recipients (individuals as well as companies) in several ways.

With spammers focusing more on fraudulent and malicious messages, the security risks caused by spam increase if the recipient's system is susceptible to malware like viruses. Emails are known to spread malicious code that is usually embedded within attached files or injected into hyperlinks, that once clicked, redirects the user to an untrusted page that contains malware. In addition, cybercrime involving phishing scams are becoming more common and users with insignificant knowledge regarding these types of scams could become victims of stolen account numbers, passwords and login details. If spammers get hold of this information they could potentially gain access to the victim's social media, instant messaging and cloud services as most of these services integrate with email for user identification and verification purposes. Companies could suffer huge financial losses without even knowing it as the majority of spam cost stem from employees who spend working time identifying and deleting spam. Consequently, human resources are wasted that could have been used for more productive tasks. Finally, many spam messages in the inbox prevents easy access to legitimate messages causing unnecessary frustration.

Currently, spam filters have proven to be the most effective measure against spam messages, but none of the filtering techniques provide a total solution (Clark, 2008). According to Goodman, Cormack & Heckerman (2007) machine learning components are present in most spam filtering systems. Depending on the filtering technique, spam filters are categorized into one of the following groups:

1. **Content-based filters** evaluate the authenticity of messages by making use of words or phrases found in the different messages. Examples include Word-based-, Heuristic- (rule-based) and Bayesian filters.

2. **List-based filters** use different types of lists in an attempt to stop spam. Email senders are grouped as trusted users or spammers which determine if their messages are blocked or allowed past the filter. Examples include Blacklist-, Whitelist-, Greylist- and Real-time blackhole list filters.

3. **Other filters** include the use of statistical and mathematical models to help distinguish between spam and non-spam messages. Examples include Support vector machines, Neural networks and Bayesian methods.

The reduction in spam email shown in Table 1 indicates progress towards an optimal solution, but the smaller decline in spam percentage indicated for 2013 of just 2.5% suggests that spammers are responding to this change. To better address the ever changing spamming techniques used by spammers, a dynamic automated spam filter is required. The automated construction algorithm for the GANN discussed in Section IV complies with these requirements, but before this technique is considered, Section III will examine the Multilayer perceptron that forms the basis of the GANN.

## III. Multilayer Perceptron

The Multilayer perceptron (MLP) is the most common artificial neural network used (Potts, 1999), capable of pattern recognition. With a three tier architecture each represented by a layer (Bishop, 1995), data is provided to the neural network through the first layer called the input layer. Weights are applied to these inputs when passed to the second (hidden) layer that is represented by one or more hidden nodes. From the hidden layer the data is progressed to the output layer where a final result is obtained. An MLP has the form

$$g_0^{-1}\big(E(y)\big) = w_0 + w_1 \tanh\left(w_{01} + \sum_{j=1}^{k} w_{j1}x_j\right) + \cdots + w_h \tanh\left(w_{0h} + \sum_{j=1}^{k} w_{jh}x_j\right), \quad (1)$$

where $g_0^{-1}$ is the inverse activation function, $y$ the target value, $w_h$, $w_{0h}$ and $w_{jh}$ the weights and $x_j$ the inputs. However, there is criticism towards the use of MLPs, because it is considered a black box. The GANN architecture consists of separate MLPs and can produce partial residual plots capable of providing meaningful insight into the results obtained. The spam researcher can use this information to help overcome the black box effect posed by MLPs to some degree.

## IV. Generalized Additive Neural Network

A Generalized additive model (GAM) (Hastie & Tibshirani, 1990) has the following expression:

$$g_0^{-1}\big(E(y)\big) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_k(x_k), \quad (2)$$

where $g_0^{-1}$ is the inverse of the output activation function, $y$ is the target value, $f_k$ are the individual univariate functions and $x_k$ are the inputs. MLPs can be used as the univariate functions of the GAM, since it are universal approximators, capable of modelling any continuous function (Du Toit & De Waal, 2010). When implementing a GAM as a neural network, it is known as a GANN and is defined as

$$f_j(x_j) = w_{1j}\tanh(w_{01j} + w_{11j}x_j) + \cdots + w_{hj}\tanh(w_{0hj} + w_{1hj}x_j), \quad (3)$$

where $f_j$ are the individual univariate functions, $w_{hj}$, $w_{0hj}$ and $w_{1hj}$ are the weights and $x_j$ are the inputs.

Previous research on the GANN automated construc-

tion algorithm (Du Toit, 2006) included mortality prediction (Bras-Geraldes, Papoila, Xufre & Diamantino, 2013) and spam email detection (Du Toit & De Waal, 2010); (Goosen & Du Toit, 2009). A GANN is less susceptible to being perceived as a black box, because partial residual plots can be generated as a by-product. These plots provide a graphical representation of the relationship between the inputs and the target making analysis and interpretation easier (Du Toit & De Waal, 2010). The automated construction algorithm for the GANN is capable of performing tasks such as variable selection exonerating users from providing input to the algorithm during optimal model searches. Pattern recognition is another benefit of a GANN as it recognizes the transformations spam messages go through as spammers change their tactics. This allows the models to dynamically adapt to new spamming techniques (Du Toit & De Waal, 2010). The next section presents the ensemble learning algorithms used that may improve the GANN classification model.

## V. Ensemble Methods

Ensemble algorithms usually obtain better predictive performance than a single technique by combining classifiers or by changing the underlying distribution of training data (Aggarwal & Zhai, 2012). Although combining classifiers may improve classification performance, ensembles are time consuming and models become hard to explain. Bootstrap aggregating (Bagging) (Breiman, 1996) and Boosting (Freund & Schapire, 1996) are both powerful ensemble algorithms used in machine learning to improve the prediction accuracy of classifier learning systems.

Bagging and Boosting are data-centered methods, as they use different segments of the training data to train a classifier and create distinctive models rather than combining different types of classifiers. The result reported is a combination or average of these models' output. Bagging performs the following steps:

1. Create $B$ separate bootstrap sets randomly of size $n$ from the training data set $X$ with displacement to obtain $X_1, X_2, ..., X_B$.

2. Fit a predictor or classifier to the individual bootstrap sets $X_i$ where each model is denoted as $h_i$ and $i = 1, 2, ..., B$.

3. Average $h_i$ to derive the final prediction or classification result denoted as $\frac{1}{B}\sum_{i=1}^{B} h_i$.

Boosting follows similar steps as Bagging but differs by building the ensemble incrementally rather than constructing the training models independently. The construction of sequential training models help in adjusting the samples to accommodate previously computed inaccuracies by applying heavier weights on misclassified observations.

In the next section the experiments performed will be discussed.

## VI. Experimental Design

The automated construction algorithm for the GANN was applied to the Enron (Metsis et al., 2006) and PU1 (Androutsopoulos et al., 2000) publicly available spam corpora. Both corpora have been divided into subsets. The Enron corpus has six subsets each consisting of legitimate and spam messages accumulated over a fixed time period from different users. These time periods are presented in a month/year format. The six subsets are known as Enron 1 to Enron 6 and are summarized in Table 2. The PU1 corpus has four subsets each with different pre-processing steps performed. Lemmatiser (Lemm) is the process of grouping together the different inflected forms of a word so they can be analysed as a single item. Stop-list (Stop) refers to a list of common words which are removed prior to modelling of the data. The PU1 subsets are described in Table 3.

| | # Messages | | Accumulation Period | |
|---|---|---|---|---|
| Subset | Legit | Spam | Legit | Spam |
| Enron 1 | 3672 | 1500 | 12/99 - 01/02 | 12/03 - 09/05 |
| Enron 2 | 4361 | 1496 | 12/99 - 05/01 | 05/01 - 07/05 |
| Enron 3 | 4012 | 1500 | 02/01 - 02/02 | 08/04 - 07/05 |
| Enron 4 | 1500 | 4500 | 04/01 - 02/02 | 12/03 - 09/05 |
| Enron 5 | 1500 | 3675 | 01/00 - 05/01 | 05/01 - 07/05 |
| Enron 6 | 1500 | 4500 | 06/00 - 03/02 | 08/04 - 07/05 |

Table 2: Enron subset description

| | # Messages | | |
|---|---|---|---|
| Subset | Legit | Spam | Pre-processing |
| PU1 Bare | 618 | 481 | N/A |
| PU1 Stop | 618 | 481 | Stop |
| PU1 Lemm | 618 | 481 | Lemm |
| PU1 Lemm Stop | 618 | 481 | Lemm and Stop |

Table 3: PU1 subset description

Publicly available spam corpora are widely used in spam filtering research. Unfortunately, the data contained in most corpora doesn't replicate real world email data, making it difficult to predict how filters would perform once implemented in such environments. This is due to numerous privacy and legal restrictions applied to most emails containing sensitive information not intended for others to see. The Enron corpus is one of the few mass email collections available to the research community which replicates real world email data. The PU1 corpus is a good example where email messages are encrypted to protect the privacy of each contributor by substituting words and email addresses with numerical values making it difficult to decipher.

For the above mentioned spam corpora pre-processing steps were carried out where every message was stripped from email attachments and HTML tags. The Enron cor-

pus also applied a stop-list to remove common words from the messages. The PU1 corpus pre-processing is listed in Table 3 and used a stop-list too where indicated. As part of the pre-processing steps, the messages were converted into a Bag-of-words representation (Mitchell, 1997), and feature selection by mutual information (MI) (Sahami, Dumais, Heckerman & Horvitz, 1998) was performed to identify the top 100 word attributes contributing the most information for use by the classifiers. When applying feature selection each word has the same probability of being chosen. MI(X;C) is defined as

$$\sum P(X=x, C=c) \log \left( \frac{P(X=x, C=c)}{P(X=x)P(C=c)} \right), \qquad (4)$$

where $X$ is the candidate attribute, $C$ the category, $x \in \{0,1\}, c \in \{Spam, Legitimate\}$, $P(X,C)$ is the joint probability distribution function of $X$ and $C$, and $P(X)$ and $P(C)$ are the marginal probability distribution functions of $X$ and $C$ respectively. Next, cross-validation was performed which makes the results less prone to random variation. The Bagging and Boosting ensemble methods were also applied to the results obtained by the GANN. To evaluate the classification performance of a spam filter which is based on text classification tasks, performance metrics are used. These metrics are subsequently described.

Let $N_L$ and $N_S$ denote the entire set of legitimate and spam messages respectively, that the filter must classify, and $n_{X \to C}$ the number of messages associated with category X which is classified by the filter as belonging to category $C$. Table 4 summarizes these four classes followed by the definitions of the evaluation metrics.

| Class | Classification | Notation |
|---|---|---|
| False positives | Legit → Spam | $n_{L \to S}$ |
| False negatives | Spam → Legit | $n_{S \to L}$ |
| True positives | Spam → Spam | $n_{S \to S}$ |
| True negatives | Legit → Legit | $n_{L \to L}$ |

Table 4: Different email classification classes

Spam recall (SR) measures the percentage of spam messages that the filter manages to block and is an indication of the filter's effectiveness. SR is defined as

$$\text{SR} = \frac{n_{S \to S}}{n_{S \to L} + n_{S \to S}}. \qquad (5)$$

Ham recall (HR) measures the percentage of legitimate messages that successfully pass through the filter and is also an indication of the filter's effectiveness. HR is defined as

$$\text{HR} = \frac{n_{L \to L}}{n_{L \to S} + n_{L \to L}}. \qquad (6)$$

Spam precision (SP) measures the degree to which the blocked messages are indeed spam, indicating the filter's safety. SP is defined as

$$\text{SP} = \frac{n_{S \to S}}{n_{L \to S} + n_{S \to S}}. \qquad (7)$$

Weighted accuracy (WACC) (Androutsopoulos et al., 2000) compares the accuracy rate to a baseline, where no filter is present - legitimate messages are never blocked and all spam messages pass through the filter. For this case, each legitimate message is treated as $\lambda$ messages. Every false positive counts as $\lambda$ errors and as $\lambda$ successes when classified correctly. If $\lambda$ is set to 1, all legitimate and spam messages are weighed the same. More realistic results are obtained when using the WACC cost-sensitive measure, because $\lambda$ assigns a higher cost to false positives (Clark, 2008). Misclassifying legitimate messages as spam can be more severe than letting a spam message pass through the filter. In this study $\lambda$ is set to 1 based on the scenario where spam messages are only flagged and remains in the recipient's inbox. WACC is defined as

$$\text{WACC} = \frac{\lambda \cdot n_{L \to L} + n_{S \to S}}{\lambda \cdot (n_{L \to L} + n_{L \to S}) + n_{S \to L} + n_{S \to S}}. \qquad (8)$$

Finally, Total cost ratio (TCR) (Androutsopoulos et al., 2000) is an indicator of spam filter performance. Thus, higher TCR values are desired as it measures the average performance of the filter. If TCR values are smaller or equal to 1 (baseline), it is better not to use a filter. TCR is defined as

$$\text{TCR} = \frac{n_{S \to L} + n_{S \to S}}{\lambda \cdot n_{L \to S} + n_{S \to L}}. \qquad (9)$$

All experiments were performed by SAS® Enterprise Miner™ 5.3 software on a HP Compaq Elite 8300 CMT system running Windows XP SP3 (32-bit) with 4GB DDR2 memory and an Intel® Core™ i5-3470 CPU at 3.20GHz.

## VII. Results

Table 5 summarises the 10-fold cross-validation results obtained by Androutsopoulos et al. (2000) using the Naïve Bayes (NB) filter as well as the results obtained by the GANN filter on the PU1 corpus for $\lambda = 1$.

| | 10-Fold Cross-Validation | | | | |
|---|---|---|---|---|---|
| Filter | $\lambda$ | SR | SP | WACC | TCR |
| NB Bare | 1 | 83.98% | 95.11% | 91.08% | 4.90 |
| NB Stop | 1 | 84.19% | 96.76% | 91.17% | 4.95 |
| NB Lemm | 1 | 78.14% | 98.25% | 89.80% | 4.29 |
| NB Lemm Stop | 1 | 79.60% | 97.96% | 90.34% | 4.53 |
| Baseline (No Filter) | 1 | 0 | - | 56.23% | 1 |
| GANN Bare | 1 | 94.98% | 94.81% | 95.50% | 9.79 |
| GANN Stop | 1 | 94.83% | 94.50% | 95.35% | 9.36 |
| GANN Lemm | 1 | 92.05% | 93.92% | 93.98% | 7.19 |
| GANN Lemm Stop | 1 | 94.47% | 94.31% | 95.08% | 8.91 |
| Baseline (No Filter) | 1 | 0 | - | 56.23% | 1 |

Table 5: Results for the PU1 corpus

Both filters performed very well with TCR values higher than 1. In all the cases, based on TCR values, the GANN outperformed the NB filter. This indicates that the filters are effective regarding spam filtering and should be useful when implemented in real world applications. Although the GANN achieved a lower SP

rate than the NB filter on all subsets, the higher SR rates suggest that it is more aggressive in blocking spam emails as shown by the results which is more than 10% higher than each PU1 subset. The GANN also outperformed the NB filter based on WACC values on all subsets.

Table 6 shows the TCR values obtained by the GANN when Bagging and Boosting were applied. It can be clearly seen that the ensemble methods improved on the results obtained by 10-fold cross-validation when using a GANN. Bagging provided the best results by achieving the highest TCR values on three of the four PU1 subsets. Boosting did second best with lower TCR values except for the PU1 Stop subset where it performed the best.

| Subset | 10-Fold Cross-Validation | Bagging | Boosting |
|---|---|---|---|
| PU1 Bare | 9.79 | 16.08 | 12.87 |
| PU1 Stop | 9.36 | 10.16 | 10.72 |
| PU1 Lemm | 7.19 | 9.60 | 8.73 |
| PU1 Lemm Stop | 8.91 | 10.67 | 9.60 |

Table 6: TCR scores on the PU1 corpus for the comparative techniques when $\lambda = 1$

Tables 7 and 8 summarize the SR rates and HR rates obtained by the GANN with 6-fold cross-validation applied to the Enron corpus. The results obtained by the GANN are compared to five different versions of the NB filter as presented by Metsis et al. (2006).

| | 6-Fold Cross-Validation | | | | | |
|---|---|---|---|---|---|---|
| Subset | FB | MV Gauss | MN TF | MV Bern | MN Bool | GANN |
| Enron 1 | 90.50 | 93.08 | 95.66 | 97.08 | 96.00 | 87.27 |
| Enron 2 | 93.63 | 95.80 | 96.81 | 91.05 | 96.68 | 97.46 |
| Enron 3 | 96.94 | 97.55 | 95.04 | 97.42 | 96.94 | 89.07 |
| Enron 4 | 95.78 | 80.14 | 97.79 | 97.70 | 97.79 | 72.38 |
| Enron 5 | 99.56 | 95.42 | 99.42 | 97.95 | 99.69 | 91.70 |
| Enron 6 | 99.55 | 91.95 | 98.08 | 97.92 | 98.10 | 96.33 |
| Average | 95.99 | 92.32 | 97.13 | 96.52 | 97.53 | 87.74 |

Table 7: Spam recall (%) results for the Enron corpus

According to Table 7 the GANN managed to block an average of 87.74% of spam messages of the complete Enron corpus. The best result was achieved by MN Bool obtaining a SR rate of 97.53%. MN TF achieved the second best results by blocking 97.13% of the spam messages. The GANN did very well on five of the six Enron subsets obtaining the highest score on the Enron 2 subset with a SR rate of 97.46%. The GANN performed the worst on the Enron 4 subset with a SR rate of 72.38%. The only other NB filter that seemed to perform weak on the Enron 4 subset was MV Gauss with a SR rate of 80.14%. Overall the five NB classifiers outperformed the GANN when using 6-fold cross-validation with SR rates of more than 92%.

The HR rates presented in Table 8 provide an overview of how well the filters managed to classify legitimate messages correctly. Low HR rates are an indication that legitimate messages were misclassified as spam. All the classifiers achieved average HR rates of more than 80%. Once again all the NB filters outperformed the GANN with the NB filters achieving HR rates of more than 89%. The GANN as well as FB and MV Bern performed relatively weak on the Enron 6 subset indicating that it was a more difficult subset to classify compared to the other Enron subsets. The NB filters performed the best on the Enron 4 subset with HR rates of more than 95%. The GANN is the only classifier that didn't perform as well with a HR rate of 71.93%.

| | 6-Fold Cross-Validation | | | | | |
|---|---|---|---|---|---|---|
| Subset | FB | MV Gauss | MN TF | MV Bern | MN Bool | GANN |
| Enron 1 | 97.64 | 94.83 | 94.00 | 93.19 | 95.25 | 88.62 |
| Enron 2 | 98.83 | 96.97 | 96.78 | 97.22 | 97.83 | 82.71 |
| Enron 3 | 95.36 | 88.81 | 98.83 | 75.41 | 99.88 | 88.36 |
| Enron 4 | 96.61 | 99.39 | 98.30 | 95.86 | 99.05 | 71.93 |
| Enron 5 | 90.76 | 97.28 | 95.65 | 90.08 | 95.65 | 82.33 |
| Enron 6 | 89.97 | 95.87 | 95.12 | 82.52 | 96.88 | 44.00 |
| Average | 94.86 | 95.53 | 96.45 | 89.05 | 97.26 | 80.87 |

Table 8: Ham recall (%) results for the Enron corpus

Table 9 represents the Enron corpus averages for SR- and HR rates obtained when applying the Bagging and Boosting ensemble methods to the GANN.

| Measure | 6-Fold Cross-Validation | Bagging | Boosting |
|---|---|---|---|
| Spam recall | 87.74% | 97.66% | 97.92% |
| Ham recall | 80.87% | 94.52% | 94.79% |

Table 9: Enron corpus averages for Spam- and Ham recall (%) obtained by the GANN and enhanced by the ensemble techniques

Boosting performed the best obtaining a SR rate of 97.92% and a HR rate of 94.79%. Bagging performed weaker with slightly lower values. Both ensemble methods greatly improved the results of the GANN's 6-fold cross-validation and is now comparable to the NB filters' higher SR and HR rates. Boosting applied to the GANN resulted in the best results for the GANN with the highest SR rate of 97.92%. MN Bool and MN TF with SR rates of 97.53% and 97.13% performed second and third respectively. Regarding HR rates, MN Bool is still the best filter with 97.26% compared to the GANN's best HR rate of 94.79%.

## VIII. Conclusions

In this study, the GANN was discussed and evaluated on its spam classification capabilities. The GANN and NB classifiers presented in this paper delivered satisfactorily performance with TCR values greater than 1. The two publicly available corpora used are good representations of real world email data emulating the random inconsistencies of ham-spam ratios over time (Metsis et al., 2006).

Overall the GANN and NB techniques produced comparable results. By applying the Bagging and Boosting ensemble methods to the GANN's results, gains in spam classification performance were achieved. The GANN outperformed the five NB filters obtaining the highest average SR rate (97.92%) on the Enron corpus. For the PU1 corpus the GANN also performed better than the NB filter. The reported results are an indication that ensemble methods may enhance the results obtained by the GANN making the GANN a more preferable choice for spam classification over the NB classifiers.


## Acknowledgements

Gratitude is expressed toward SAS Institute Inc. for providing Base SAS® and SAS® Enterprise Miner™ software used in computing the results presented in this article. This work forms part of the research done at the North-West University within the TELKOM CoE research programme, funded by TELKOM, and THRIP.


## IX. Bibliography


Aggarwal, C. C. & Zhai, C. X. (2012), *Mining Text Data*, Springer Publishing Company, Incorporated.

Androutsopoulos, I., Koutsias, J., Chandrinos, K. & Spyropoulos, C. D. (2000), An experimental comparison of naive bayesian and keyword-based anti-spam filtering with personal e-mail messages, *in* 'Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', Athens, Greece, pp. 160–167.

Bishop, C. M. (1995), *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford.

Bras-Geraldes, C., Papoila, A., Xufre, P. & Diamantino, F. (2013), Generalized additive neural networks for mortality prediction using automated and genetic algorithms, *in* 'Proceedings of the 2nd International IEEE Conference on Serious Games and Applications for Health (SeGAH)', Vilamoura, Portugal, pp. 1–8.

Breiman, L. (1996), 'Bagging predictors', *Machine Learning* **24**(2), 123–140.

Clark, K. P. (2008), 'A survey of content-based spam classifiers'. Date of access: 15 May 2014. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.173.2685

Cranor, L. F. & LaMacchia, B. A. (1998), 'Spam!', *Communications of the ACM* **41**(8), 74–83.

Du Toit, J. V. (2006), Automated Construction of Generalized Additive Neural Networks for Predictive Data Mining, PhD thesis, School for Computer, Statistical and Mathematical Sciences, North-West University, South Africa.

Du Toit, J. V. & De Waal, D. A. (2010), Spam detection using generalized additive neural networks, *in* 'Proceedings of the Southern Africa Telecommunication Networks and Applications (SATNAC) Conference'.

Freund, Y. & Schapire, R. E. (1996), Experiments with a new boosting algorithm, *in* 'Proceedings of the 13th International Conference', Machine Learning, pp. 148–156.

Goodman, J., Cormack, G. V. & Heckerman, D. (2007), 'Spam and the ongoing battle for the inbox', *Communications of the ACM* **50**(2), 25–33.

Goosen, J. C. & Du Toit, J. V. (2009), Spam detection with generalized additive neural networks, *in* 'Proceedings of the Southern Africa Telecommunication Networks and Applications (SATNAC) Conference'.

Gudkova, D. (2012), 'Kaspersky security bulletin: Spam evolution 2012'. Date of access: 15 May 2014. http://www.securelist.com/en/analysis/204792276/Kaspersky_Security_Bulletin_Spam_Evolution_2012

Gudkova, D. (2013), 'Kaspersky security bulletin: Spam evolution 2013'. Date of access: 15 May 2014. http://www.securelist.com/en/analysis/204792322/Kaspersky_Security_Bulletin_Spam_evolution_2013

Guzella, T. S. & Caminhas, W. M. (2009), 'A review of machine learning approaches to spam filtering', *Expert Systems with Applications* **36**, 10206–10222.

Hastie, T. J. & Tibshirani, R. J. (1990), *Generalized Additive Models*, Vol. 43 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

Metsis, V., Androutsopoulos, I. & Paliouras, G. (2006), Spam filtering with naive bayes - which naive bayes?, *in* 'Proceedings of the 3rd Conference on Email and Anti-Spam (CEAS)'.

Mitchell, T. M. (1997), *Machine Learning*, McGraw-Hill Series in Computer Science, WCB/McGraw-Hill, Boston, Massachusetts.

Potts, W. J. E. (1999), Generalized additive neural networks, *in* 'KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, New York, NY, USA, pp. 194–200.

Sahami, M., Dumais, S., Heckerman, D. & Horvitz, E. (1998), A bayesian approach to filtering junk e-mail, Technical Report WS-98-05, Learning for Text Categorization – Papers from the AAAI Workshop, Madison Wisconsin.

Tomlinson, R. (n.d.), 'The first network email', *Raytheon BBN Technologies* . Date of access: 15 May 2014. http://www.raytheon.com/newsroom/rtnwcm/groups/public/documents/content/rtn12_tomlinson_email.pdf



**Pieter Labuschagne** received his B.Sc. in 2010 and his Hons. B.Sc. in 2011 from the North-West University, Potchefstroom Campus in Computer Science. Presently he is studying towards his Master of Science degree at the same institution in the field of Artificial Intelligence.

# Development of an online reputation monitor by using existing components

G.J.C. Venter, W.C. Venter, A.J. Hoffman
School of Computer and Electronic Engineering,
North-West University, Potchefstroom Campus,
Private Bag x4001, x442, Potchefstroom 2520
Tel: +27 76 358 8241, Fax: +27 18 299 1977
Email: mgmGertV@gmail.com, willie.venter@nwu.ac.za, alwyn.hoffman@nwu.ac.za

**Abstract- Customer opinion about companies are very important and companies often get customer feedback via surveys or other official methods. Some customers prefer to voice their opinion on the internet where they take comfort in anonymity. Currently this form of customer feedback is not closely monitored. This project aims to address this shortcoming by developing a system capable of monitoring various web and social networking sites for customer feedback.**

**Index Terms— Online Reputation Monitor, Web crawler, Facebook, Twitter, dtSearch.**

## I. INTRODUCTION

Advertising is a tool that is used to establish a basic awareness of a product or service in a potential customer by providing selected information [1]. Research shows that advertising has a major influence on customer preferences and can help consumers to make decisions [1][2]. However, according to Reichheld [3] the tremendous cost of marketing makes it hard for a company to grow profitable. Reichheld [3] believes that the only path to a profitable growth rate lies in the company's ability to get loyal customers to promote a company by sharing information about the company's products and the customer's experience [4].

Getting customers to talk about a company is not enough as customers may either promote or advise against the company. As such the company needs to monitor its public reputation. Most companies know this and often employ techniques such as focus groups and surveys to generate various statistics but these methods are not always effective. People often feel under pressure when their opinions are personally asked and therefore adjust their answers to avoid any potential confrontation. Instead many customers prefer to voice their opinion on the internet where they take comfort in anonymity. Therefore companies need to determine their online reputation.

Determining online reputation is not a new field and has been done for years by organizations such as BrandsEye and Brand.Com. However, the services these companies offer have certain limitations. The services normally require a monthly fee and companies are not allowed to purchase the online reputation monitoring (ORM) software. The user also has very little control over the software itself: such services search for user specified keywords over a range of websites, yet the user cannot specify which websites to monitor or change the keywords without restarting the service. Additionally the services rarely make use of historical data, meaning a company's reputation can only be analyzed from the present day onwards.

To solve this problem a new ORM system is proposed. The system will allow the user to scan a number of web pages and social networking sites such as Twitter and Facebook at a sufficient rate to ensure that all results are continually kept up to date. Once results from the web and social networking sites have been gathered the user will be presented with various methods capable of analyzing the results. The user will be capable of retrieving information regarding the result's source such as the website, page and paragraph, the date at which the result was generated as well as the sentiment and the category of the result. Once all the results have been analyzed the user will have the option to generate various reports that will indicate how his or her company is seen by the online community.

## II. BACKGROUND

The basic functionality of ORM involves scanning a variety of web pages and social networking sites, analyzing the results to determine their relevancy, using the analyzed results to generate various statistics and visualizing the results. Clearly an ORM system isn't a single process, but a collaboration of several processes that includes a web and social networking crawler, a method of analyzing the results and a sentiment analysis tool.

### *A. Web Crawler*

Web crawlers are programs that explore the World Wide Web, retrieve information and store the results for future use [5]. This process is known as *web crawling*. The type of data that is extracted from web pages depend on the implementation of the web crawler. Some web crawlers are configured to extract only specified phrases [6], while others extract and index each word in a web page for future use [7].

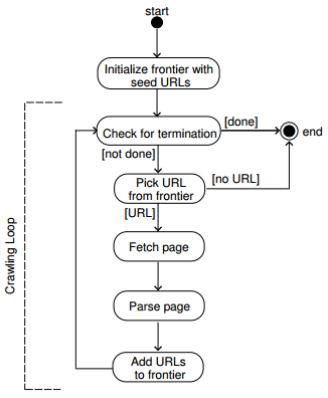



**Figure 1: Basic Crawler Architecture**

*Figure 1* shows the architecture of a basic web crawler [5]. Before a web crawler is initialized a user must specify a list of seed URLs which is stored in the *frontier*, a list of unvisited URLs. When the web crawler starts, it will load the first URL in the frontier, download and scan the contents and store any results in a database or on a local storage device. Once the crawler has finished scanning the page it will load the next URL from the frontier and repeat the process. This is known as the *crawling loop.*

Many web sites contain multiple pages, which in turn contains additional subpages. This is illustrated in *Figure 2*.

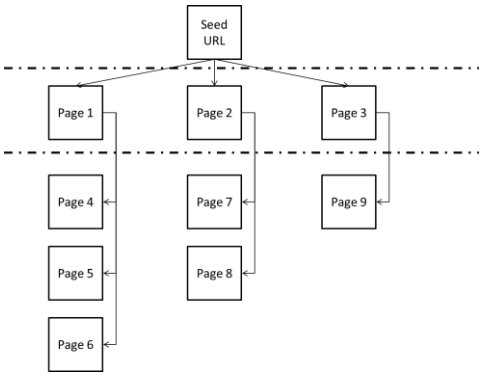



**Figure 2: Crawl Depth Illustration**

Web crawlers are often configured to scan only a certain *depth*. This is the extent to which a web crawler scans the page and is specified when the user adds the seed URL. If the user wish to scan only the original page the crawl depth will be set to 0. If the user sets the crawl depth to 1, the web crawler will scan the seed URL and add the URLs for *Page 1*, *Page 2* and *Page3* to the top of the frontier. Once the web crawler has finished scanning the seed URL, it will proceed to download and scan *Page 1*, followed by *Page 2* and finally *Page 3*. The same process will be followed for *Page 4-9* if the user sets the crawl depth to 2.

Web crawler execution ends when all the links in the frontier have been scanned. At this stage timers are usually initialized which will redeploy the crawler at a specific time or after a specified interval.

### B. *Social Networking Crawlers*

Unlike regular web sites, web crawlers cannot scan social networking sites. Social networking sites only present data once the user has registered and even if logged in the site will only show data regarding the people the current user is connected to. Normal web crawlers do not possess authentication capacities and therefore cannot access social networking sites and those that do will only have access to limited information.

Therefore various social networking sites have designed an interface that allows registered applications to access the site's public data. This interface is known as an Application Programming Interface (API).

All social networking sites have different API's and therefore different methods of extracting information. As such it will not be possible to design a universal method of interfacing with all social networking sites. The project will only be able to interface with some social networking sites due to time constraints. It was decided to interface with Facebook and Twitter as they had the highest Alexa rank at the time, which is an indicator of network traffic.

### C. *Sentiment analysis tools*

Once the web or social networking crawler returns a result it must first be analyzed to determine whether the result contains any relevant data before it can be presented to the user. However, the user often wishes to know whether the sentiment of the result is positive or negative.

There are two main methods used to calculate the sentiment of some text, namely the lexical approach and the machine learning approach.

A system that is based on the lexical approach uses a dictionary of pre-tagged words. Each word in the text that has to be analyzed is compared against the dictionary to find its polarity. This indicates whether the word has a positive or negative sentiment as well as the strength of the sentiment. After the polarity of each word has been determined, the polarity of the given text is calculated by summing the polarity for each word. According to Annett [8], the accuracy of such a system varies between 64% and 82%, depending on orientation of statistic metrics and the dictionary that was used.

A system that uses the machine learning approach uses two components; a series of feature vectors and a collection of tagged corpora which is a collection of documents that the system uses to train itself [9]. Feature vectors are usually a variety of *uni-grams*, single words from within a document, or *n-grams*, two or more words from a document that are in sequential order. Other features that are often proposed include the number of positive words, number of negative words, and the length of a document. Both the feature vectors and collection of tagged corpora are used to train a classifier, which can be applied to an untagged document to determine its sentiment. According to Annett [8], the accuracy of such a system varies between 63% and 82%, but the results are dependent of the features that were selected.

## III. DESIGN

As stated in the introduction, existing ORM system have various technical limitations that limit their usefulness and therefore it was decided to implement a new ORM system. To include all the functionality of ORM as discussed in Section II it was decided to develop a system consisting of three parts, namely a Back-End, a Front-End and a Website.

### A. *The Back-End*

To begin the ORM process information must be retrieved from the internet. This will be done by the Back-End, which will initialize and maintain all the web and social networking crawlers.

Before the ORM process can be started the user must specify a list of seed URLs, their respective crawl depths as well as a series of keywords. The URLs and the crawl depths will be used by the web crawler whereas the social networking

crawler will use the keywords to find any results from Facebook and Twitter.

Information gathered by the web and social networking crawlers will be saved either to a database or a local storage device. The storage location will depend on the type of web and social networking crawler that will be used. Once the Back-End has finished crawling the web and social networking sites it will proceed to wait for a predetermined amount of time before restarting the process.

The Back-End will be designed to function with as little human interaction as possible. The only interaction that may be required would be to verify that the Back-End has not encountered an error or to turn off the system in times of maintenance.

### B. The Front-End

After the data has been acquired by the Back-End the results need to be processed. This is the goal of the Front-End. The Front-End will proceed to load the stored results for a specified date range and process the data by identifying the paragraphs in which the keywords feature before passing the paragraphs through a similarity filter and a sentiment analysis tool. Afterwards the user will be able to verify the results before saving them to a database.

The similarity filter will filter out results that are too close to the keywords, such as tags within a web page. This will lessen the amount of results that will have to be processed by the sentiment analysis filter. The sentiment analysis tool will use the results that have passed through the similarity filter and determine whether the results are positive or negative. These results will be used to determine the overall opinion of the company or brand that is being investigated. Once the sentiment of the results have been calculated they will be saved back in the database.

The Front-End will be designed to allow multiple instances of the software to run in parallel. This will allow multiple system administrators to access the system at the same time which will increase the amount of results that can be processed at once. While it will be possible to fully automate the Front-End, it would be wise to allow some degree of human interaction to verify if the results that pass through the filters are correct. The main reason for this is the fact that the meaning of words can differ depending on the context which it is used in. Using the keyword “Kalahari” to monitor the public opinion of the online marketplace “Kalahari.com” may yield results for both the online marketplace and the Kalahari Desert, which will influence the results and statistics.

### C. The Website

Once the results have been processed the system administrator will require a method to display the processed results to the user or any other interested party. This can easily be accomplished via a web site as it will allow the data to be accessed anywhere.

## IV. Component Selection

The implementation of the ORM system will make use of existing components - there is no use in redesigning the wheel. Available components will be evaluated and component selection will be based on the results of the evaluation.

### A. Web crawler

Ideally the ORM system would make use of a powerful web crawler such as *GoogleBot* or *BingBot*, respectively developed by Google and Microsoft. Unfortunately these crawlers are not available for public use. An investigation discovered many free and commercially available web crawlers available on the internet. A commercial text retrieval engine as well as several open source web crawlers were acquired and evaluated.

The dtSearch Engine is a commercial text retrieval engine with many features that provides the user with a web crawler that is actively maintained. The user has no control over the internal functionality of the dtSearch Engine as it is not open-source but the dtSearch Engine can be manipulated by using the settings that are presented. The dtSearch Engine creates index files when scanning the internet which can be accessed by a result generator included with the web crawler.

The HTML Agility Pack (HAP) is a free library that allows the user to parse web pages directly from the web. The HAP does not have any internal web crawling capabilities but provides the user with components to build a custom web crawler and result generator. The HAP is open source; as such the HAP is completely cost-free but support for the library may stop if the community stops working on the project. A second open-source web crawler named *Abot* was evaluated. The *Abot* crawler contains pre-programmed functionality that is based on the HAP while still presenting the user with as much functionality as possible.

Various aspects of the crawlers were compared using a weighted average to determine the web crawler best suited for this project. The aspects considered were the speed with which the web crawler is able to scan the web, the time it will take to incorporate the web crawler into the ORM system, additional features and the customizability of the web crawler as well as its cost.

The results are shown in *Table 1*. For a full discussion of the weight assigned to the different aspects as well as the scores of the different web crawlers for each aspect please refer to the dissertation “Development of an Online Reputation Monitor” by G.J.C. Venter at the North-West University.

**Table 1: Web crawler comparison**

| Category | dtSearch Engine | HAP | Abot |
|---|---|---|---|
| **Speed (30%)** | 26 | 27 | 19 |
| **Implementation time (30%)** | 30 | 6 | 12 |
| **Existing Features (15%)** | 15 | 2 | 7 |
| **Customizability (15%)** | 7 | 15 | 12 |
| **Cost (10%)** | 0 | 10 | 10 |
| **Total (100%)** | **78** | **60** | **60** |

From *Table 1* it can be seen that despite its cost the dtSearch Engine will be best suited for the ORM service. While the HAP and the Abot web crawlers may be almost as fast and free, they lack in the “implementation time” and “existing features” categories.

### B. Social networking crawlers

As already explained, to extract sufficient information from social networking sites the ORM system will have to make use of the social network's API. The API provided by Twitter enables the user to extract data in one of two ways: by performing a search query via the REST API or getting a continuous feed of data from the Streaming API. This project will make use of the Streaming API because querying the RESP API is slow and creates significant internet overhead.

Two free Twitter components are available to allow the user to access the Twitter Streaming API, namely TweetInvi and Linq2Twitter. The differences between the components are marginal; both return only results which match a specific number of conditions.

Various aspects of the components are compared using a weighted average to determine the component best suited for the project. The aspects considered were the amount of results, the time it will take to implement the component within the ORM system, additional features and customizability.

The results of the components are shown in *Table 2*. For a full discussion of the weight assigned to the different aspects as well as the scores of the different components for each aspect please refer to the dissertation "Development of an Online Reputation Monitor" by G.J.C. Venter at the North-West University.

**Table 2: Twitter API component comparison**

| Category | TweetInvi | Linq2Twitter |
|---|---|---|
| **Amount of results (30%)** | 27 | 30 |
| **Implementation time (30%)** | 30 | 10 |
| **Existing Features (20%)** | 7 | 20 |
| **Customizability (20%)** | 10 | 20 |
| **Total (100%)** | **74** | **80** |

From *Table 2* it can be seen that it will be faster to implement the TweetInvi component but the limited information of the results makes it less viable than Linq2Twitter. TweetiInvi would also require additional filters in order to filter out non-English results which would take additional time. In accordance with the results Linq2Twitter was chosen to be used in the ORM system.

Facebook has a different API than Twitter and requires a different interfacing method and a different component. Searches have shown the Facebook SDK for .Net is the component that is used the most to interface with Facebook.

### C. Result Analysis Method - String Similarity Formula

The results of the web and social networking crawlers that contain the specified keywords will be saved, but a certain portion of the results, such as tags within a web page, will have insignificant meaning. Manually filtering out these results will cost the user valuable time, therefore a similarity filter based on the Levenshtein and Kuhn-Munkres algorithms will be used to filter out unnecessary results. There are other string similarity formulas available, as proposed by Mihalcea [10] and Li [11], but are deemed too complex for this implementation.

The Levenshtein algorithm is used to measure the similarity between two strings by calculating the least number of edit operations that are necessary to modify one string into becoming another [12]. The Kuhn-Munkres algorithm is an algorithm capable of solving linear assignment problem (LAP) instances [13].

The similarity between a result and its keywords can be calculated by breaking up both the result and the keywords into separate lists of their individual terms called tokens, followed by comparing the similarity of the tokens within each list against each other before comparing the tokens in the list of keywords against the tokens in the list of results.

The Levenshtein algorithm is used to compare the similarity between the tokens by providing a set of rules that calculates the cost of changing a token to another by using a series of one step operations. Each one-step operation has an associated cost; substitution costs 2 units whereas cost of insertions and deletions is 1 unit.

After the tokens in each list have been compared the algorithm will compare the tokens between the list of results and the list of keywords. This will return a matrix, which will be solved by the Kuhn-Munkres algorithm.

### D. Sentiment Analysis Tool

Once a record has been generated and processed its sentiment must be calculated to determine whether the message is positive or negative. Various sentiment analysis tools are available on the internet such as the *AlchemyAPI* [14] and the *Salience Engine* provided by *Lexalytics* [15].

To use the AlchemyAPI the user must first register on the AlchemyAPI website to acquire a *software development kit (SDK)* and two registration keys. In order to determine the sentiment of a sentence the user must use the SDK with the 2 keys to verify the application and pass the sentence that must be analyzed as a parameter. AlchemyAPI will proceed to analyze the sentence and present the user with the result.

The Salience Engine from Lexalytics provides the user with a multi-lingual text analysis engine that can be integrated into systems for business intelligence and social media monitoring. The Salience Engine includes many features, such as sentiment analysis, named entity extraction and summarization.

From features alone the Salience Engine would be better suited for this project, but in order to acquire the Salience Engine the user would have to contact Lexalytics and request access to the system. There are also no pricing opinion available on the website, whereas the AlchemyAPI SDK could easily be acquired once registered. As such the AlchemyAPI was used for this project but the Salience Engine can be investigated in the future.

## V. IMPLEMENTATION

The advantages and disadvantages of different implementations of the chosen tools are discussed in this section.

### A. dtSearch Engine – Web crawling

The scanning of thousands of web pages using the basic dtSearch Engine implementation will take quite long because the dtSearch Engine scans a single page at a time while the other pages are kept in a queue. To overcome this problem multiple instances of the dtSearch Engine are initiated in the ORM and the web pages are divided amongst them.

Multiple instances requires the use of *threading* at the cost of more system resources. To investigate the threading option the number of threads were increased while the CPU and memory usage as well as the internet bandwidth usage were monitored. All tests were done using a 1Mb/s (128 KB/s) internet connection. The results are shown in *Table 3*.

**Table 3: Threading comparison**

| Threads | 1 | 2 | 4 | 8 | 16 |
|---|---|---|---|---|---|
| **Time (s)** | 2381 | 1259 | 722.37 | 573.41 | 430.03 |
| **CPU Usage (%)** | 1.5 | 2.5 | 4.1 | 5.0 | 5.1 |
| **Memory Usage (%)** | 1.4 | 2.5 | 4.1 | 7.08 | 14.5 |
| **Bandwidth (kB/s)** | 23.67 | 57.02 | 98.16 | 127.18 | 133.28 |
| **Websites loaded** | 788 | 788 | 788 | 788 | 788 |
| **Stable** | Yes | Yes | Yes | Yes | Yes |

From *Table 3* it can be seen that as the number of threads increased the amount of time required to scan all the sites decreased, but the usage of the system resources increased. More threads could be used but at 32 threads there were too many simultaneous web connections which caused some of the web crawlers to time-out as they could not receive information fast enough. This problem can be solved with a faster internet connection.

### B. Social networking sites - Twitter

The Linq2Twitter component provides results when received from the Twitter API. The number of results are not dependent on the computer speed and multiple instances are therefore not needed.

During testing it was noticed that some results from the Linq2Twitter component cannot be used by the ORM service, e.g. non-English results and results that have been previously detected. Saving these results to the database will increase the execution time of the Front-End as the results have to be filtered out each time the results for a specific keyword are requested. To prevent the saving of these results an additional filter was developed that only saves English results. The record has not been detected before and the message of the keyword does contain the keyword. Results failing any of the checks were discarded.

**Table 4: Twitter Filter Results**

| | Test1 1 | Test 2 | Test 3 |
|---|---|---|---|
| **Correct** | 1377 | 1657 | 1942 |
| **Incorrect Language** | 3426 | 3055 | 3184 |
| **Duplicate record** | 295 | 830 | 658 |
| **Total** | **5098** | **5542** | **5784** |
| **Percentage Correct** | **27.0%** | **29.8%** | **33.5%** |

The effect of the filter can be seen in *Table 4*. Each result was passed through the filter as it was received. It can be seen that only about 30% of the results are indeed viable for the ORM service.

### C. Social networking sites – Facebook

Unlike Twitter, the Facebook streaming components do not provide the user with continuous posts but with limited information regarding user and page status *updates*. To retrieve information regarding the API the user will have to perform a search request using another component of the Facebook SDK called the Graph API.

The Graph API will provide the user with various information regarding the latest Facebook public posts such as user ID, the message and the date at which it was created and will always return the latest 25 results. During integration it was discovered that if the Graph API is queried once per minute the ORM system will not lose any results as results are generated very slowly on the Facebook system, but will receive various duplicates. In order to prevent the saving of these results an additional filter was developed to filter out any duplicate results. To test the filter 5 queries were executed each a minute apart. For the 1st query the filter reported 70.4% unique results, for the 2nd 16%, for the 3rd 16.8%, for the 4th 14.4% and for the fifth 24.8%. The initial query contains the most unique results and subsequent queries contain fewer unique results as more results are duplicated and filtered out.

## VI. RESULTS

### A. Web Crawling

The Web crawler was loaded with 250 websites, crawl depth of 1. The results for the crawler test are shown in *Table 5*.

**Table 5: Web crawler results**

| Crawler ID | Time | #Seed URLs | #Links detected | #Links scanned | Crawl Rate |
|---|---|---|---|---|---|
| **1** | 3 hours 31 minutes | 16 | 6332 | 2877 | 0,23 |
| **2** | 3 hours 43 minutes | 16 | 4221 | 1585 | 0,12 |
| **3** | 3 hours 17 minutes | 16 | 4604 | 2075 | 0,18 |
| **4** | 2 hours 29 minutes | 16 | 3182 | 1175 | 0,13 |
| **5** | 2 hours 32 minutes | 16 | 3352 | 1239 | 0,14 |
| **6** | 2 hours 55 minutes | 16 | 4056 | 1773 | 0,17 |
| **7** | 2 hours 40 minutes | 16 | 4085 | 1555 | 0,16 |
| **8** | 3 hours 19 minutes | 16 | 4091 | 1458 | 0,12 |
| **9** | 2 hours 53 minutes | 16 | 3897 | 1690 | 0,16 |
| **10** | 2 hours 50 minutes | 16 | 3611 | 1317 | 0,13 |
| **11** | 3 hours 29 minutes | 15 | 4302 | 1617 | 0,13 |
| **12** | 3 hours 11 minutes | 15 | 4805 | 1805 | 0,16 |
| **13** | 2 hours 5 minutes | 15 | 2741 | 1184 | 0,16 |
| **14** | 3 hours 6 minutes | 15 | 6631 | 2194 | 0,2 |
| **15** | 3 hours 34 minutes | 15 | 3683 | 1326 | 0,1 |
| **16** | 2 hours 4 minutes | 15 | 3014 | 1028 | 0,14 |
| **All** | 2 hours 59 minutes | 250 | 66607 | 25898 | 2.41 |

From *Table 5* it can be seen that the web crawler scanned the web pages at a rate of 2.41 web pages per second. This will result in scanning 1000 web pages within 40 minutes. If a higher scan rate is required it can be realized by using a more powerful computer with a faster internet connection. To test whether the web was scanned correctly a single keyword was chosen and the number of instances found were verified against the original web page. All instances were found. The test was executed on a computer with a 3GHz CPU, 16GB RAM and a 1Mb/s internet connection.

It should be noted that less links were scanned than detected. This occurred because of duplicate links that were detected within web pages such as style sheets, as well as links that are not allowed to be crawled, as dictated by the website.

### B. Twitter Scanning

The Twitter Scanner was loaded with 26 keywords and executed for 1 hour 35 minutes. During the execution 74 201 results were detected at an average rate of 13.01 results per second. Of the 74 201 results 51 556 results were in a non-English language and 2 882 were duplicates of previously detected results. When passed through the similarity filter 6 542 results were removed, which provides the ORM service with 13 221 viable results. Due to the nature of social networking sites it will be impossible to verify the results against the Twitter database and therefore only the results as presented by Linq2Twitter can be verified.

### C. Facebook Scanning

The Facebook Scanner was loaded with 26 keywords and executed for 1 hour 54 minutes. During the execution 13 250 results were detected at an average rate of 2.12 results per second. Of the 13 250 results 11 469 were duplicates of previous detected results which left 1 781 results. When passed through the similarity filter 766 results were found viable for the ORM service. Due to the nature of social networking sites it will be impossible to verify the results against the Facebook database and therefore only the results as presented by Facebook SDK can be verified.

### D. Sentiment Analysis

To test the sentiment analysis tool the sentiment of 125 random results from each of the three web crawlers were calculated and manually verified. For the web crawler 91.8% of the sentiments were calculated correctly, for the Twitter Scanner 91.2% and for the Facebook Scanner 95.2%.

## VII. Conclusion

This paper demonstrated an online reputation monitoring system capable of monitoring both the web and various social networking sites. The system is capable of extracting information from the internet in real time from various user-defined web sites as well as Twitter and Facebook and can determine which results are most likely to contain any relevant information via various filters. For an additional discussion on how the various components act together to calculate a company's online reputation, please refer to the dissertation "Development of an Online Reputation Monitor" by G.J.C. Venter at the North-West University.

While this system is capable of performing ORM services there are various areas available for future research. Development of an offline sentiment analysis tool as well as a custom web crawler will greatly improve the cost-efficiency of the system. Some of the included filters could also be further improved upon; while the Twitter Scanner will automatically filter out the non-English results, this feature is not available for the Facebook Scanner.

## VIII. Bibliography


[1] A. B. Ayanwale, T. Alimi and M. A. Ayanbimipe, "The influence of advertising on consumer brand preference," *Journal of Social Science*, vol. 10, no. 1, pp. 9-16, 2005.

[2] S. J. Hoch and Y.-W. Ha, "Consumer learning: advertising and the ambiguity of product experience," *Journal of consumer research*, pp. 221-223, 1986.

[3] F. F. Reichheld, "The one number you need to grow," *Harvard business review*, vol. 81, no. 12, pp. 46-55, 2003.

[4] N. Hu, L. Liu and J. J. Zhang, "Do online reviews affect product sales? The role of reviewer characteristics and temporal effects," *Information Technology and Management*, vol. 9, no. 3, pp. 201-214, 2008.

[5] G. Pant, P. Srinivasan and F. Menczer, in *Crawling the web*, Springer, 2004, pp. 153-177.

[6] Web2Mine, "Easy Web Extract Software," Web2Mine, 2013. [Online]. Available: http://webextract.net/. [Accessed July 2013].

[7] dtSearch, "How dtSearch Works," dtSearch, [Online]. Available: http://www.dtsearch.com/PLF_howdtworks.html. [Accessed March 2013].

[8] M. Annett and G. Kondrak, "A comparison of sentiment analysis techniques: Polarizing movie blogs," in *Advances in artificial intelligence*, Springer, 2008, pp. 25-35.

[9] R. Feldman, "Techniques and applications for sentiment analysis," *Communications of the ACM*, vol. 56, no. 4, pp. 82-89, 2013.

[10] R. Mihalcea and P. Tarau, "A language independent algorithm for single and multiple document summarization," *Proceedings of IJCNLP*, vol. 5, 2005.

[11] Y. Li, D. McLean, Z. A. Bandar, J. D. O'shea and K. Crockett, "Sentence similarity based on semantic nets and corpus statistics," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 18, no. 8, pp. 1138-1150, 2006.

[12] RosettaCode.org, "Levenshtein Distance," 25 January 2014. [Online]. Available: http://rosettacode.org/wiki/Levenshtein_distance. [Accessed 3 March 2014].

[13] D. Dasgupta, G. Hernandez, D. Garrett, P. K. Vejandla, A. Kaushal, R. Yerneni and J. Simien, "A comparison of multiobjective evolutionary algorithms with informed initialization and Kuhn-Munkres algorithm for the sailor assignment problem," in *Proceedings of the 2008 GECCO conference companion on Genetic and evolutionary computation*, ACM, 2008, pp. 2129-2134.

[14] AlchemyAPI, "About Us," Alchemy API, 2014. [Online]. Available: http://www.alchemyapi.com/company/about-us/. [Accessed 17 April 2014].

[15] Lexalytics, "Salience Engine," Lexalytics, 2014. [Online]. Available: http://www.lexalytics.com/technical-info/salience-engine-for-text-analysis. [Accessed 28 04 2014].



**G.J.C. Venter** received his undergraduate degree from the North-West University (NWU) in Potchefstroom in 2012. He is currently studying towards his Masters in Engineering at the same institution. His research includes web data extraction and online reputation monitoring.

# Facebook Crawler Architecture for Opinion Monitoring and Trend Analysis Purposes

Sinesihle I. Mfenyana, Nyalleng Moorosi, Mamello Thinyane
Telkom Centre of Excellence of ICT for Development
Department of Computer Science
University of Fort Hare, Private Bag X1314, Alice 5700
Tel: +27 406022464, Fax: +27 406022464
Email: {smfenyana, nmoroosi, mthinyane}@ufh.ac.za

**Abstract-In the contemporary era of web 2.0, the Internet is being used to build and reflect social relationships among people who share similar interests and activities. Currently, there is a very high usage of Social Networking Sites (SNSs) and blogs where people share their views, opinions, and thoughts. This leads to the production of a lot of data by people who post such content on SNSs. This paper presents a system prototype for performing opinion monitoring and trend analysis on Facebook. The proposed system prototype crawls Facebook, indexes the data and provides a User Interface (UI) where end users can search and see the trending of topics of their choice. The system prototype could also be used to check the trending topics without having to search. The main objective of this research project was to develop a framework that will contribute in improving the way government officials, companies or any service providers and normal citizens communicate regarding services they provide. This research project is premised on the conceptualization that if the government officials, companies or any service providers can keep track of the citizen's opinions, views and thoughts with regards to services they provide it can help improve the delivery of such services.**

**Index Terms—Facebook, Web Crawling, Information Retrieval, Siyakhula Living Lab (SLL), Service Delivery**

## I. INTRODUCTION

SNSs are used by people to connect with each other for business or personal purposes. They build and reflect social relationships among people who share similar interests and activities[1][2][3]. These SNSs are also used by many businesses and government to inform, engage and serve citizens. In government, this is done through government 2.0, which is the usage of social media tools to facilitate the conversations between governments institutions and citizens to shape polices, support local government democracy and improve services [4].

These free and easy to use cloud based applications provide individuals, organizations and societies an easy and cheaper way of communication [5]. SNSs' services provide content sharing through the publication of documents and links and also through messages exchanged using communication tools. These functionalities allow for the use of SNSs as a collaborative opinion mining platform. There are many factors that motivate the use of SNSs for opinion trend analysis. Firstly, SNSs are easy to use and provide a cheap way of communication as evidenced by a growth in their usage [6]. Additionally, most SNSs data can easily be accessed by the use of supported Application Programming Interface (APIs).

The applicability of the proposed system is wide-ranging; however in this research the focus is on improving service delivery in marginalized rural communities, which form the field site of this research. In this research we mine opinions from constituents with a special interest in monitoring trends about service delivery in marginalized rural communities of South Africa. The proposed system seeks to improve or optimize consultation principle, one of the eight Batho Pele Principles initiatives. Batho Pele (Sesotho word, meaning "People First" in English) was launched in 1997 in South Africa because the preceding public service system was believed not to be people-friendly and lacked the skills and attitudes to meet the developmental challenges facing the country. The idea of this initiative was the need to transform public service at all levels [7]. The consultation principle refers to the process of interacting with, listening to and learning from the people whom you serve by staying in touch with them so that you can find out what services they need, how they would like their services to be delivered and with which they are not satisfied [7].

The proposed system prototype will provide a cost effective solution for communication between companies, government officials, or organizations which seek to find out about the public or clients' views, opinions, ideas etc. about the services they provide. In this research project, the proposed system will be used to enable the local government in Dwesa (and other communities in South Africa) to determine the trending topics and discussions that are related to service delivery within their jurisdiction. This is achieved through the implementation of a focused Facebook crawler and opinion trend analysis tool. The crawler extracts statuses and comments feeds which are then used for opinion monitoring and trend analysis purposes.

The remainder of the paper is organized as follows. The Related work and the Research methodology are given in section 2 and 3 respectively. Section 4 presents System architecture and Implementation. Section 5 presents System Testing and Experimental Results. Section 6, the last section is the conclusion and future work.

## II. RELATED WORK

In 2013, Kim et al [8] implemented a prototype system for detecting trend and bursty keywords from Twitter stream data. To detect trend and bursty keywords, it first selects candidate keywords from tweets by performing simple syntactic feature based filtering. Then, merge various

keyword variants using several heuristics and select bursty keywords based on the term frequency. By tracing the popularity transition of such trend keywords, it then determined bursty keywords. Their system first collects user tweets via Tweeter Streaming API and extracts candidate keywords from them (tweets) by calculating their term frequency (TF). It then identifies various word variants which are considered some full keywords and merge them semantically with the same keywords also adjusting their term frequencies accordingly. It then determines trend keywords based on their rank and bursty keywords are selected from them based on the temporal pattern of their popularity. Their work is different to ours in that their work is not domain specific as ours focuses on government service delivery. The other difference is that their work was based tweets, which are limited to only 140 characters unlike Facebook statuses which are not limited. In 2013, Fotis et al[9] presented the architecture of a distributed crawler which harnesses information from multiple OSNs. They argued that the extraction of facts and aggregated information from individual Online Social Networks (OSNs) has been extensively studied in the last few years while cross–social media–content examination has received limited attention. The motivation of their work was that content examination involving multiple OSNs gains significance as a way to either help to verify unconfirmed-thus-far evidence or expand the understanding about occurring events. On their work they also demonstrated that contemporary OSNs feature similar, if not identical, baseline structures by proposing an extensible model termed SocWeb that articulates the essential structural elements of OSNs in wide use today. To accurately capture features required for cross-social media analyses, SocWeb exploits intra-connections and forms an "amalgamated" OSN. They introduced a flexible API that enables applications to effectively communicate with designated OSN providers and discuss key design choices for their distributed crawler. Their approach helped to attain diverse qualitative and quantitative performance criteria including freshness of facts, scalability, quality of fetched data and robustness. They work is different to our work in the sense that were gathering data from to OSNs (Facebook and YouTube) while our developed crawler only crawl Facebook.

## III. Research Methodology

The research method that was followed in this research consists of a combination of well-established research methods which are requirements gathering, system development and system testing.

**Requirements Gathering**-During the course of this research, the research site was constantly visited for one week per month by the SLL research team. During the visits, the SLL research team conducted computer literacy trainings and also got to know more about the research site at large.

Questionnaires ware used to find out how people of Dwesa are currently reporting the issues related to government service delivery that they are facing. The questionnaire focused on finding out about; SNSs currently being known and used, methods currently being used to communicate their problems with their current local government, and the possibility of the improvements the developed system could bring.

Through a literature review of similar previous works on our research domain, we were able to design and foresee the appropriate tools which could be used to develop our system. A review of programming languages and most of the tools which were previously used to develop similar systems was also carried out. A brief review of the tools that were chosen and used to develop the system and its usage is given on the System Architecture and Implementation section below.

**System Development**–The Iterative Incremental Development approach was used in developing the system. The Iterative Incremental Development approach is a combination of the Iterative approach, which refers to an approach that allows cycling through the development phases from requirements gather to the final developed system deliverable[10] and Incremental approach which is an approach that allows the development of various parts in different stages and schedules which are then integrated when they are completed[10]. This development approach allowed us to develop the proposed system in different modules which were then incremented and were also revised later whenever research objectives were not met.

**System Testing** - During its development phase, the system was constantly modularly tested. The detailed system testing and evaluation with the experimental results are given in section 6.

## IV. System Architecture and Implementation

The system is developed with open source technologies, because the SLL from where the system will be deployed uses open source software which is free. The system is developed in different modules which were constantly tested and later integrated together to form one system package. The developed framework serves two high level functionalities which are outlined as follows:

1. Data extraction from Facebook, text extraction, text preprocessing and text indexing.

2. Index searching which consists of four sub-modules which are outlined as follows:

   - Keyword searching, content matching and frequency analysis;
   - Keyword searching, content matching and correlation analysis;
   - Content matching and frequency analysis; and
   - Keyword searching.

These sub-modules are integrated together with one UI. The UI is developed in such a manner that it eases the usage of the system. These high level system functionalities and the undertaken literature review informed the design of the system. The following diagram illustrates the high level system architecture of the developed system in this research.

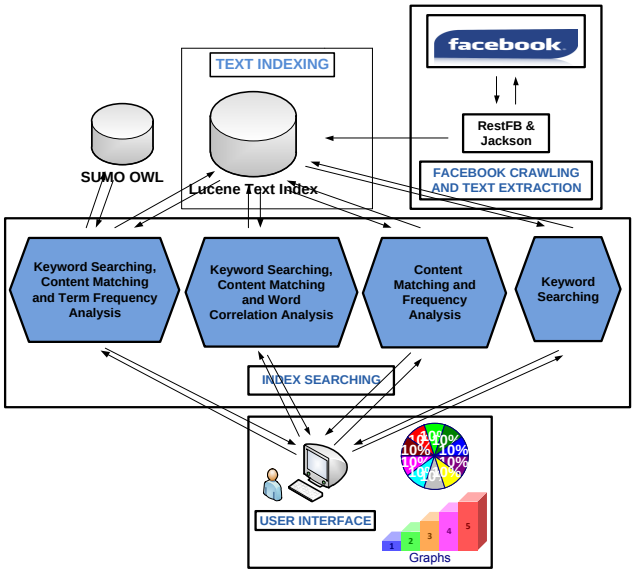



**Figure 1: System Architecture**

The system architecture above in Figure 1 illustrates the internal and external structure of system modules integrated together in one package to form one system. The following subsections provide a brief background overview of the tools used and the design details of the different modules of the developed system starting from the back-end to the front-end. The whole system is developed on Linux Ubuntu operating system and Netbeans IDE platforms because they are open source technologies.

### *A. Facebook Crawling and Text Extraction*

This system module was developed to crawl Facebook. This is the system back-end module developed on top RestFB and Jackson java libraries. RestFB is a simple client java based library alternative for the Facebook Graph API written in Java [11]. This library is used by developers for communicating with Facebook database servers. FacebookClient interface was used to communicate with Facebook Graph API. This interface specifies how the Facebook graph API client is supposed to operate [12]. The sub-classes FetchConnection and FetchObject of FacebookClient were used for extracting friends' user account ID's and data feeds from Facebook respectively. All data extracted from Facebook was returned in JsonObjects, an unordered collection of name/value pairs. JavaScript Object Notation (JSON) is a lightweight data-interchange format between machines and humans [13]. The data was also filtered by specifying the parameters to retrieve from Facebook and that eased the process of data mapping through java beans when parsing text. The extracted Facebook data was then parsed to plain text using Jackson library. The parsing of Facebook data was done for two reasons:

1. The user account IDs were parsed to get actual value of user account ID. The user account ID was then used as reference to retrieve the latest status update together with associated comments of the specific user.

2. The data feeds which is the actual data indexed, were parsed because Lucene only indexes plain text.

Lucene open source java library developed by the Apache organization for high text indexing and efficient search algorithms performance was used for indexing and adding searching functionality on the system. Collecting data from Facebook Breath-first-search (BFS) SNS sampling algorithm was used, which includes an agent that collects seed's friends user-ID's and an agent which is responsible for crawling friends' data feeds. In BFS sampling algorithm web pages are crawled according to the way they are discovered [14].

The Facebook Crawling and Text *Extraction module* operates as follows: it contacts the Facebook server, providing Facebook access token required for the authentication and permissions for accessing users' data. Once logged in, the agent starts crawling seed's (logged in Facebook user) friends list extracting friends' user-IDs and also crawling data feeds of each and every user in First-In-First-Out (FIFO) queue manner. Facebook data feeds are arranged in a chronological order with the most recent data feed appearing at the top Only the top data feed that is the last status update in every friend's Facebook account is crawled and indexed. This allows the tracking of the trending topic, on the network of the seed node.

### *B. Text Indexing*

This system module was built on top Lucene information retrieval java library. The Facebook Crawling and Text Extraction module as aforementioned was developed for Facebook data extraction and text parsing; the text was then indexed using Lucene. Lucene only indexes data available in textual format. The input text data in Lucene is stored in an inverted index data structure, which is stored on file system or in memory as a set of index files. On the development of this module we used different Lucene library classes to accomplish the objective of developing a working index. The following are the core indexing classes which were used during the development of the index: Directory, Analyzer, and IndexWriter.

Directory is the abstract class used to represent the location where the index files are to be or stored. We also used its sub-class FSDirectory store index files in the actual file system. This sub-class was used because we were getting large sizes of data and such data was also intended to be used for trend analysis purposes. The Analyzer class was used for converting the text data into a fundamental unit of searching, which is called a term. During analysis, the text data goes through multiple operations such as extracting the words, removing common words, ignoring punctuation, words stemming, also converted into tokens, and these tokens are added as terms in the Lucene index. There are already built-in Lucene analyzers which differ in the way they tokenize the text and apply filters, but in our development we used the StandardAnalyzer.

The IndexWriter class was used for creating and maintaining an index and its constructor accepts a Boolean that determines whether a new index must be created or an existing index is opened [15]. The IndexWriter class also provides methods for adding, deleting, or updating documents in the index. To avoid duplications and ensuring that most recent statuses were used, we used the deleteAll method every time new data was about to indexed. Lastly, Document class was used to represent a collection of fields,

and the data was stored under two fields "title" and "content". Title being status or comments and the content being the actual data we obtained from Facebook.

### C. Index Searching

This module defines how end user's queries, content matching, correlation analysis and trend analysis are done. This module from our system consists of four modules whose designs will be discussed in the following sub-sections.

#### 1) Content Matching and Frequency Analysis

This system module adopted a fully-automatic method for trend analysis from text and term-document matrix. Term-document matrix is the frequency of terms or words in a collection of documents, with documents being columns and rows representing terms. We provided our system with the collection of textual data in an array, which are regarded as stop-words train dataset. This module then, starts by checking if the words from the array are also available from the index. The frequencies of words which are found from the array and the index are not calculated. To access data from Lucene we used MatchAll query to retrieve all the documents from the Lucene index, then tokenized the retrieved data to enable us to match terms from the array and tokenize terms from index and to also be able to calculate word frequencies which are found not matching.

#### 2) Keyword Searching, Content Matching and Frequency Analysis

This system module adopted the semi-automatic method for trend analysis for text in the index. It used SUMO data and data from the Lucene index for content matching and frequency analysis. SUMO is the largest formal public ontology in existence today with different domain ontologies on it [16]. The use of SUMO was to ensure that our system would not be restricted to the government service delivery related topics only. The usage of the SUMO aimed at giving the developed system in this research the ability to retrieve data from different domains in cases where end-users wanted to check trending topics which are not government services delivery related.

Through this system module, an end-user supplied a keyword to be searched. The module then regards the search keyword as the class which could be found from the SUMO. Thereafter, the module then searches for the SUMO with the keyword. If the class name matching the searched keyword is found the supper-class, sub-class and the instances of such specific found are then retrieved and preprocessed by removing unnecessary delimiters and separating the joined words. The words are then stored in an array to be used as the train dataset.

This module also uses MatchAll Lucence query to retrieve all the data found from the index. The data found from the index is then tokenized to single tokens. The tokens are then also stored in an array. The two arrays, the one which acts as the train dataset created from data found from the ontology and the actual data found from the Lucene index are then matched. The words which are found matching from the two arrays are stored in a different array. The algorithm for selecting these words is: every token from the index is checked if it is there from the train dataset array. If the token is found from the train dataset array is then added to a new array regardless of how many times the token is found from the index data. Its instances are stored in the new array according to its number occurs. The newly created array is then used to calculate the frequency of the words that are found matching from the SUMO and Lucene index. The results are then visualized to the end-user the trending words on a bar graph. This module is used to check the trends from the entire index in the system.

#### 3) Keyword Searching, Content Matching and Correlation Analysis

This system module also adopted a semi-automatic method for detecting the trends from the textual data found from our index. This module also requires the keyword to be searched but it differs from the previously described module in that it searches the index only and it does not use SUMO. This module searches the index using the normal keyword searching method and retrieves all the data found matching the searched keyword. The data is then tokenized and stored in an array. This module has a stop-words array provided to it as the train dataset for preprocessing data retrieved from the index by removing them, as the stop-words occur almost in every textual data. The two arrays are then matched by looking for the words that occur in both arrays. The words whose frequencies match are not calculated. The correlation in this module is calculated and determined based on the word frequencies, with the words with higher frequencies being regarded as the words with higher correlations. The results are then visualized on a scatter plot for the end-users to analyze and draw conclusions.

#### 4) Keyword Searching

This sub-model enables the keyword searches. Based on results from the above described modules an end-user can search the most correlating and/or trending words through this system feature, to get the content where they can see what has been said on Facebook in relation to the words which were found correlating or trending. Keyword searching is not limited to searching for words found in the above described modules but can also accept any query and retrieves the documents that match the query. To limit the number of documents retrieved we used Lucene TopScoreCollector class and limited the results to top 10 documents with the highest hit scores. The reason for limiting the results which are retrieved through this system module is to make it easy for the end-user to easily get the sentiments of the retrieved content form the index.

### D. User Interface

The UI was created using JavaServer Pages (JSP) and linked to each of the four index searching modules with different java servlets. The servlets are responsible for accepting queries from JSP and passing them to index searching modules and accepts the feedback from these modules and passes them back to be displayed and/or visualized on JSP page. Jfreechart bar graphs and scatter plot are used to visualize results from content matching and frequency analysis, keyword searching, content matching and frequency analysis, and keyword searching, content matching and correlation analysis modules. Results from the keyword are displayed in the plain text. The UIs for each and every module was developed and tested separately and they were all called to function in one UI to make one system package. The interfaces were developed according to the functionality they are supposed to offer and the expected results.

## V. System Testing and Experimental Results

In this section we discuss the visual results for our work, the results that the end-users will come in contact with. However, there are also a lot of technical results we achieved during this research project to make the system work as it should.

### A. Keyword Searching, Content Matching and Frequency Analysis

Testing this module had some limitations such as the data format from SUMO, was totally different to the data that is likely to be found from Facebook. This was because of the writing styles and wrong spellings which are normally found from Facebook and SUMO uses correct spellings of the words.

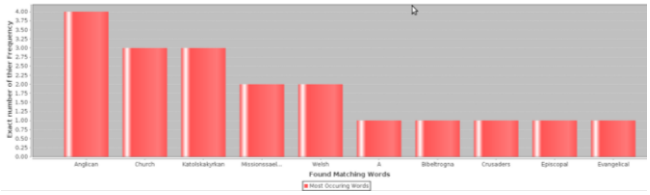


**Figure 2: Keyword Searching, Content Matching and Frequency Analysis**

Trying to find out about all text delimiters which are used in defining the words on SUMO was also one of the drawbacks in using SUMO. The last drawback which was noticed in using SUMO is that it was taking time to search the ontology which jeopardized the response time of the system. In order to test the functionality of this module, we needed three people with Facebook user accounts to update their statuses and write some comments on their time lines with some few words we were getting from the ontology. We then initiated our system to crawl Facebook, preprocess the data we were getting from Facebook and then index. We then performed the searching through this module to see the functionality of the system. We searched the word "Church" and the results from Figure 2 were found. The results suggest that it possible to use ontology for e trend analysis from the SNSs data.

### B. Keyword Searching, Content Matching and Correlation Analysis

After the death of one of the greatest world icon, the former South Africa first black President Nelson Rolihlahla Mandela was announced on the 5th December 2013, it filtered all over the SNSs and the Internet. That in turn prompted us to know what people on the network from Facebook which we were using to test our system, were saying about the death of Mr. Mandela.

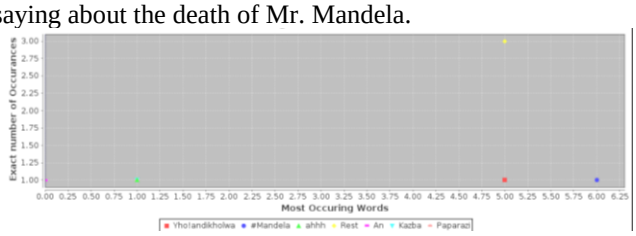


**Figure 3: Keyword Searching, Content Matching and Correlation Analysis Results**

On the 6th of December 2013, the day after the tragic of passing the South African president we crawled Facebook. We collected statuses and comments of the Facebook users which are in the network with the Facebook user account we were using in this research project for testing purposes. The reason gathering data from Facebook on such day, we wanted to see the trends about the passing of our former first black president. We then searched his surname "Mandela" to see the words which were mostly used in the statuses and comments that matched the searched keyword. The results of the searched keyword are visualized in Figure 3. The word "Rest" was the word with the highest frequency as can be observed in Figure 3. This word was most occurring on the statuses and comments of Facebook users which were mourning about the passing of our former president. The results suggest if there could be any government services delivery related data from Facebook it could be possible to collect and analyze. At the moment, when the system was tested people which were on the Facebook network which was used to test the system, hardly ever post posts which were related to service delivery.

### C. Keyword Searching

The section discusses the keyword searching module. This module retrieves statuses and/or comments that match the searched keyword. It also limits the maximum number of the retrieved statuses and/or comments to ten based on their hits.

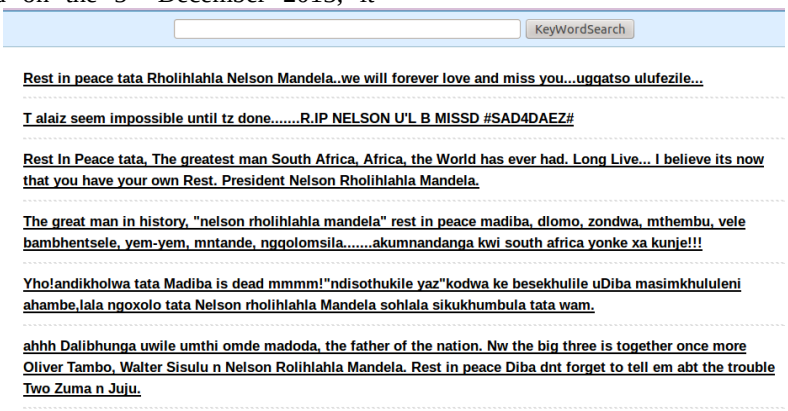



**Figure 4: Keyword Searching Results**

Figure 4 above presents the results which are the statuses and/or comments which were retrieved after we searched the keyword "Nelson" from our system. The retrieval of the actual statuses and comments feature in our system was

developed so that end-users could get the actual sentiments of the content which is found matching the searched keyword. With the results displayed of the keyword searched, an end-user who knows IsiXhosa and English could simply understand that the people who wrote the statuses and comments in Figure 4 were actually expressing their condolences to the Mandela family. The data as aforementioned was collected on the 6th December 2013. These results suggest had there have been data related to survive delivery during the time the data was collected. The system end users were to be able to learn what people were saying about service delivery.

### A. *Content Matching and Frequency Analysis*

The section presents the results of the last module in index searching modules. This module does not require the keyword search input for it to start searching the index. This module starts retrieving the data from the index and starts calculating the frequencies of the words found from the index. That implies that the trending words are not calculated based on the keyword searched but on the whole index.

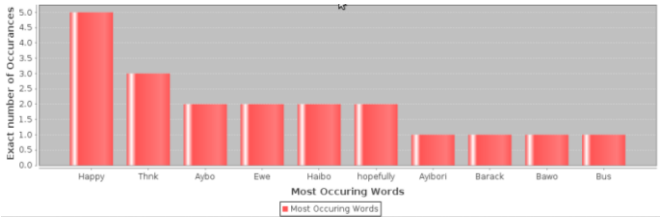



**Figure 5: Content Matching and Frequency Analysis Results**

The results in Figure 5 were the results of the content matching and frequency analysis found after we collected and indexed on 12 December 2013. These results were the words with higher frequencies, even though there were probably more words with frequencies of one but we limited our results to ten.

## VI. Conclusion and Future Work

In this paper we have presented the research done and outlined the development of the proposed system that will probably improve the way the government and local people communicate. The implementation of the Facebook crawler for opinion monitoring and trend is a success story, although its deployment at Dwesa where the system is intended to be used still remains a future proposal and/or development. The success of the system when performing its actual main duty, which is to provide a cheap way of communication between government officials and the local people, will be determined by the reviews of the targeted end-users, after the system has been tested and deployed at Dwesa.

## VII. Acknowledgement

This work is based on the research undertaken within the Telkom CoE in ICTD supported in part by Telkom SA, Tellabs, Saab Grintek Technologies, Easttel, Khula Holdings, THRIP and National Research Foundation of South Africa (UID : 84006). The opinions, findings and conclusions or recommendations expressed here are those of the authors and none of the above sponsors accepts any liability whatsoever in this regard.

## VIII. References

[1] P. Chiu, C. M. K. Cheung, and M. K. O. Lee, "Online Social Networks: Why Do 'We' Use Facebook?," *J. Mark. Res.*, vol. 19, pp. 67–74, 2008.

[2] C. M. K. Cheung, P. Chiu, and M. K. O. Lee, "Computers in Human Behavior Online social networks : Why do students use facebook ?," *Comput. Human Behav.*, vol. 27, no. 4, pp. 1337–1343, 2011.

[3] A. Nadkarni and S. G. Hofmann, "Why do people use Facebook?," *Pers. Individ. Dif.*, vol. 52, no. 3, pp. 243–249, 2012.

[4] M. Srivastava, "Social Media and Its Use by the Government," *J. Public Adm. Gov.*, vol. 3, no. 2, pp. 161–172, Jul. 2013.

[5] R. Bassett, T. Chamberlain, S. Cunningham, and G. Vidmar, "Data Mining and Social Networking Sites: Protecting Business Infrastructure and Beyond," *Data Min. Soc. Netw. Sites*, vol. Volume XI,, no. 1, pp. 352–357, 2010.

[6] E. Costa, R. Ferreira, P. Brito, I. I. Bittencourt, O. Holanda, and A. Machado, "Expert Systems with Applications A framework for building web mining applications in the world of blogs : A case study in product sentiment analysis," *Expert Syst. Appl.*, vol. 39, no. 5, pp. 4813–4834, 2012.

[7] "Batho Pele: Improving government service." [Online]. Available: http://www.etu.org.za/toolbox/docs/govern/bathopele.html. [Accessed: 18-Sep-2013].

[8] D. Kim, S. Rho, and E. Hwang, "Detecting Trend and Bursty Keywords Using Characteristics of Twitter Stream Data," *Int. J. Smart Home*, vol. 7, no. 1, pp. 209–220, 2013.

[9] F. Psallidas, A. Ntoulas, and A. Delis, "SocWeb : Efficient Monitoring of Social Network Activities," in in *Web Information Systems Engineering--WISE 2013*, Springer, 2013, pp. 118–136.

[10] A. Cockburn, "Using Both Incremental and Iterative Development," *Technol. Softw. Eng.*, no. May, pp. 27–30, 2008.

[11] "RestFB - A Lightweight Java Facebook Graph API and Old REST API Client." [Online]. Available: http://www.restfb.com/. [Accessed: 20-Sep-2013].

[12] "FacebookClient (RestFB)." [Online]. Available: http://restfb.com/javadoc/com/restfb/FacebookClient.html. [Accessed: 20-Sep-2013].

[13] "JSON." [Online]. Available: http://www.json.org/. [Accessed: 20-Sep-2013].

[14] M. Najork and J. L. Wiener, "Breadth-First Search Crawling Yields High-Quality Pages," pp. 114–118, 2001.

[15] M. Mccandless, E. Hatcher, and O. Gospodnetic, *Lucene in Action*, Second Edi. Manning Publications Co. 180 Broad St. Suite 1323 Stamford, CT 06901.

[16] "The Suggested Upper Merged Ontology (SUMO) - Ontology Portal." [Online]. Available: http://www.ontologyportal.org/.

**Mfenyana Sinesihle I.** received his Honours in Computer Science in 2012 from University Fort Hare and completed his Master of Science degree at the same institution in 2013. His research interests include social networks, intelligent systems, web crawling and data retrieval systems.

# Using online assignment submissions at an open distance learning institute to predict future Internet traffic for a higher educational institution!

Arthur J Swart
Department of Electrical, Electronics and Computer Engineering
Central University of Technology, Private BagX20539, Bloemfontein, 9300
Department of Electrical and Mining Engineering
University of South Africa, PO Box 392, Science Campus, UNISA, South Africa, 0003
Tel: +27 51 507 3907, Fax: +27 86 607 1786
Email: drjamesswart@gmail.com

**Abstract-Internet traffic has increased dramatically over the past few years. However, Internet penetration remains low in Africa, while high internet costs, limited finances for new computer technology and poor e-skills of many citizens hamper the effective use of the Internet. The purpose of this paper is to present current Internet traffic for two modules offered at an open distance learning institute within the Electrical Engineering environment, with three noteworthy predictions. These future predictions are primarily based on the file size of online assignment submissions which are routed between students and academics at an open distance learning institute. These predictions point to significant Internet traffic increases for online assignment submissions during the next five years and that not ALL registered students have the necessary e-skills and computer technology to submit their written assignments online.**

**Index Terms—online submissions, Internet traffic, written assignments,**

## I. INTRODUCTION

In South Africa (SA), as in most of Africa, using or having an Internet connection remains a luxury and is not as widely used as in developed counties [1, 2]. Consider just the cost of Internet access in SA, which is still relatively expensive when compared to Europe and the United States [3, 4]. The total cost for residential uncapped Internet access in SA is currently around R 364 per month according to Telkom's official webpage [5]. In 2011, one out of every two black African households spent less than R 840 per month on each of its members [6]. This means that roughly 43% of their expenditure would have to go to Internet access, something which is just not possible in SA's current economic climate.

Large numbers of previously disadvantaged people live in many rural areas of SA with few employment opportunities [7], and subsequently limited household income. Open and Distance Learning (ODL) students from these rural areas do not have the financial means to afford Internet access or the latest computer technology to connect effectively to the Internet. They are therefore disadvantaged anew, in that access to an online Learning Management System (LMS) or a University's assignment router is just not possible.

Internet penetration in SA is approaching 20% (Africa at 18% - see Figure 1), with the Internet user base having grown from 6.8 million in 2010 to 8.5 million at the end of 2011 [8]. Despite this growth, nearly 8 million people in SA still access the Internet on their mobile phones. In fact, personal Internet access is as low as 3% in some of the nine local provinces [9]. Internet access is therefore limited in many rural communities in SA, with access often gained through cell phone usage which has a limited bandwidth. In fact, Ko [10] declared that with the current low levels of ICT access, SA would find it difficult to provide most citizens with access to public services, such as e-government, e-entrepreneurship, and e-learning services.

SA has citizens that are lacking in e-skills [11]. ODL students falling into this category would have difficulty in accessing and effectively using any online LMS or assignment router. This could well be true of freshman or first-year students who have not yet been exposed to advanced ICT programmes. However, senior university students would have been required to complete one or other ICT course during their studies and would therefore possess a set of specific e-skills, advantaging them more than their first-year counterparts.

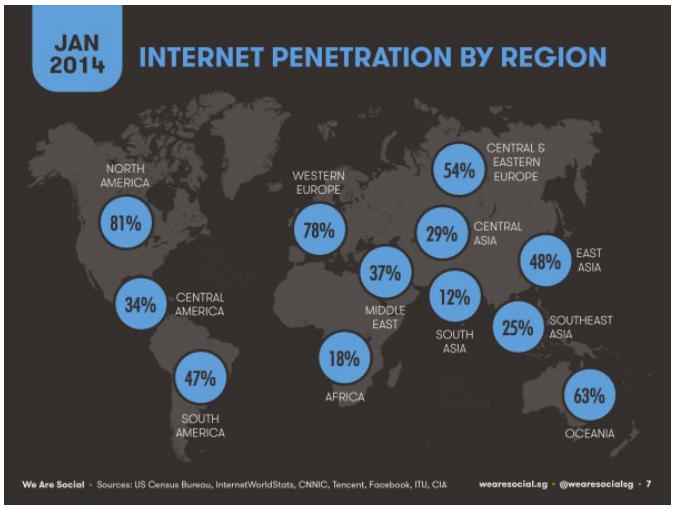


Figure 1: Global Internet penetration [12]

These challenges (high internet costs, limited financial means for new computer technology, low rural Internet penetration and poor e-skills) hamper the effective use of the Internet in the ODL environment, proving problematic to overcome in an ever dwindling economy. This is especially

so when many ODL institutions are moving completely online. For example, the University of South Africa (UNISA) has mandated an institutional move to online delivery of all teaching materials [13] as well as onscreen marking (OSM) of all written assignments which should be submitted online [14]. Written assignment submissions play a major role in ODL pedagogy, as they are designed to help prepare students for a final venue-based examination [14]. Students are assisted to understand the structure of an examination paper, and become used to the examiner's style of questioning. Preparing students for the examination paper includes timely feedback, which assists students to rectify any misconceptions which they have regarding the study material.

The following research questions thus arise: (1) "What Internet traffic will be generated by students registered with UNISA in terms of online assignment submissions?" (2) Do ODL students have the necessary e-skills to effectively make use of the Internet in terms of online assignment submissions? (3) Will ALL registered students be able to access the system thereby proving that they have both the financial and technological ability?

The purpose of this paper is to present current Internet traffic for two modules offered at UNISA within a telecommunications course, using this data to predict future Internet traffic for this higher education institute. The context of this study is firstly outlined and then followed by current infrastructure challenges. The research methodology is given along with pertinent results and conclusions follow.

## II. THE ODL CONTEXT AT UNISA FOR ASSIGNMENTS

UNISA, the largest ODL institute in SA, has experienced tremendous student growth over the past few years (see Figure 2). In 2013 it provided distance education to almost 400 000 non-residential students [15]. UNISA currently has seven colleges and one Graduate School of Business Leadership, with the College of Economic and Management Sciences attracting more than 40% of all registered students.

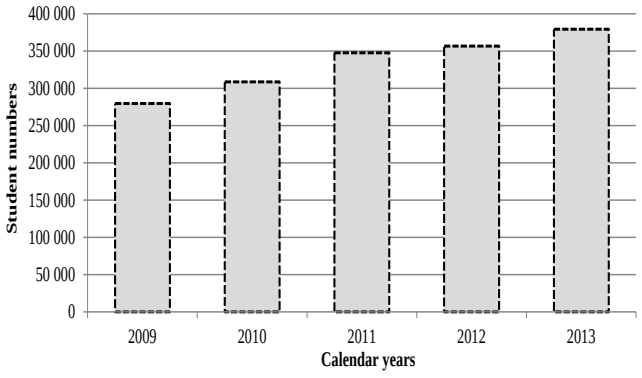



Figure 2: Student growth at UNISA

Normally, two written assignments are required per module to obtain a year mark, which currently contributes 20% to the final mark of the student within the College of Science, Engineering and Technology. UNISA offers both semester (14 week period) and calendar year (7 month period) modules. Semester modules may require a minimum of one written assignment while the calendar year modules may require a minimum of two written assignments (due dates shown in Figure 3). Written assignments are submitted online via UNISA's assignment router, termed the J-Router (located in Pretoria). Academics are required to log into the J-Router once they receive an email notification of student assignments which need to be marked. Academics may browse through various J-Router windows until they identify their module, whereupon they select the module and relevant assignments for download to their personal workspace. Once downloaded, they can log out of the J-Router and mark the assignments offline. Once completed, they log back into the J-Router and place the marked assignments in their respective outboxes, from where it is routed back to the relevant students for feedback purposes. Two uploads and two downloads of the assignment therefore occur, as shown in Figure 3. The procedure to effectively work with the J-Router as well as assignment marking was discussed by Swart [14].

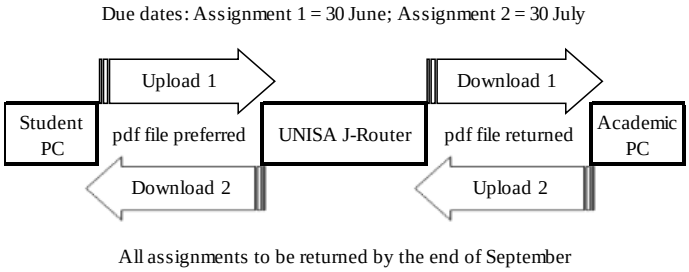



Figure 3: Data transfer between students, UNISA and the academic over a four month period

## III. CURRENT ONLINE SUBMISSION CHALLENGES

The installation of the J-Router and OSM tools may take up to 5 hours to finalize [14]. This is because different settings have to be selected and verified. The latest computer technology is also required, with Windows 7 being preferred as the minimum operating software. An ADSL is preferred (minimum of 1 Mbps) to be able to route the assignments between the J-Router and the PC of the academic. Adding to the challenge are external markers, who live more than 100 km away from a UNISA campus and who do not have the latest computer technology available at home. How does UNISA accommodate these markers and the student assignments which are meant for them?

The user friendliness of the J-Router is problematic when selecting modules in different J-Router windows [14]. Routing the assignment from the inbox to the workspace may take a number of minutes, depending on the number and file size of the assignments. Moving between the different J-Router windows does not prove problematic. However, moving between different modules available in different windows requires multiple logouts and logins in order to select the desired module. Is it possible for the software to be adapted to address this concern?

Internet access by students is problematic at times [14]. UNISA caters to students all over Africa, including those living in rural communities. The only form of contact with the outside world is still the post office, as the

telecommunications network must still reach many of these outlying and remote areas. Expecting these students to have a fast Internet connection or the latest computer technology to upload their assignments is unfair and unreasonable. At most they will have a standard modem connected to their telephone line for Internet communications. What is currently being done to enable more students to access the Internet at acceptable speeds?

Bandwidth limitations impact negatively on the system's response time, especially when many assignments are routed around by markers at the same time [14]. This is especially of concern when markers work from home, where they may have limited bandwidth due to limited financial means. Why could a service level agreement not be negotiated between UNISA and telecommunications network operators for lower connection costs for markers and students?

SA has citizens that are lacking in e-skills [16]. This includes students who may struggle to format their assignments correctly, leading to unusually large file size submissions [17]. Large files lead to long download times [18] which may cause increased frustration and dissatisfaction with the online submission process [19]. Large files necessitate increased storage requirements by all users [20] and can result in slow processing, which needs more memory and increased transmission costs [21]. Would it not be advisable for ODL institutes to include assignment file formatting, submission and handling in their orientation courses for freshman students?

Five rhetoric questions were raised in the aforementioned discussion to emphasize the unique challenges faced by ODL institutions, academics and students. Suitable answers to these questions may contribute to students successfully uploading and downloading their written assignments within a shorter period of time as compared to the traditional postage system. Students will benefit from having more time to reflect on their answers [22], identifying their mistakes and making appropriate corrections before their final examination.

## IV. Research Methodology

This study's sample involves using two specific modules, termed Satellite Communication IV (SCM4) and Radio Engineering 4 (RAE4), which forms part of a telecommunications course offered during 2013. Both modules are offered over a year period (approximately 7 months) at UNISA and require two written assignments.

A software program called NetWorx V5.2.10 is used to monitor the Internet traffic over a WIFI connection of a DELL notebook (E6530) which is connected to a 1 Mbps ADSL from Telkom (home subscriber in Vanderbijlpark using a Tenda W548D V2.0 modem). NetWorx is a free and simple, yet powerful tool that helps you objectively evaluate your bandwidth situation [23]. It can be used to collect bandwidth usage data and measure the speed of your Internet or any other network connection. NetWorx can help to identify possible sources of network problems, ensure that one does not exceed the data limit specified by an ISP which could incur additional fees, or track down suspicious network activity characteristic of Trojan horses and hacker attacks. Incoming and outgoing traffic is represented on a line chart and logged to a file, so that one can view statistics about daily, weekly and monthly bandwidth usage and dial-up duration. A polyline was selected with minute intervals within the graph settings. Data usage summary is displayed on the graph in bytes which shows Internet traffic from right to left.

Internet traffic download from the J-Router is considered when unmarked assignments are put into the academics' workspace; i.e. assignments are downloaded from UNISA's J-Router to a specific folder on the academic's PC to enabling OSM. Internet traffic upload to the J-Router is considered when marked assignments are put into the academics' outbox; i.e. assignments are uploaded from the academic's PC to UNISA's J-Router for administration purposes and for routing to students.

Assignments downloaded to the workspace of the academic are further analyzed in terms of file size and correlation to the final mark obtained for the specific assignment. This is done in order to determine if large file sizes result in better assignment marks in order to establish if students really need to submit large pdf files. Criticism has been leveled against systems that grade students for the quantity of work done rather than its quality [24].

Linear regression may be used to project future growth rates [25] and was used to predict future internet traffic for online assignment submissions (uploading and transferring the assignments between the relevant parties). This was done using the current mean file size for assignment submissions and the projected linear growth rate of students.

## V. Results and Discussion

Figure 4 shows a logon (roughly 20 seconds to complete between Vanderbijlpark (academic's home) and Pretoria (UNISA's J-Router location) and page selection. Logon and page selection (moving between the different J-Router windows – eight visible in this sketch) averaged about 185 KB (average between Figure 4 and 5). Requests for data (Internet traffic upload) averaged about 50 KB.

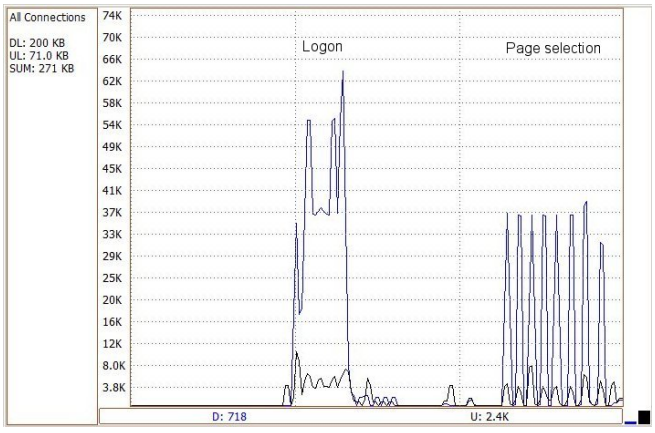



Figure 4: J-Router logon (roughly 20 seconds) and page selection (1 minute intervals between vertical dotted lines)

Figure 5 highlights that a logon profile may take more than 1 minute to complete, being very jagged and inconsistent. Although the supply of bandwidth is improving year by year, there still exists a great demand for bandwidth management due to the expense incurred in obtaining a

connection [26], which may vary between 20 seconds and more than 1 minute in this case.

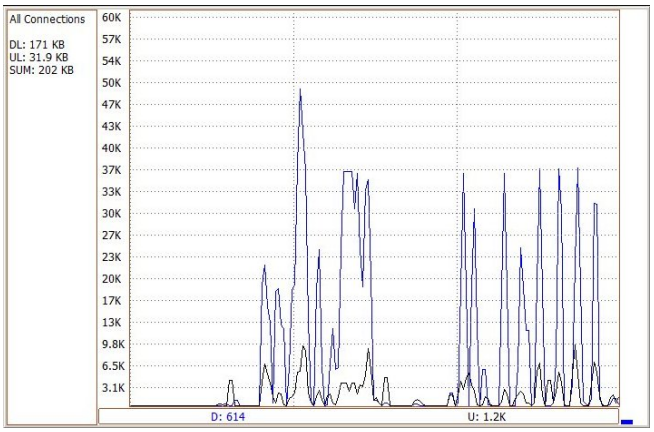



Figure 5: J-Router logon (roughly 1 minute) and page selection via a notebook's WIFI connection to a 1 Mbps ADSL from Telkom (19 September 2013 using NetWorx)

Figure 6 presents the time required to upload 13.844 MB of marked assignments (average speed of 450 kbps) while Figure 7 shows 34.105 MB of Internet traffic download for unmarked assignments (average speed of 900 kpbs). Download speed is thus double the upload speed to the J-Router.

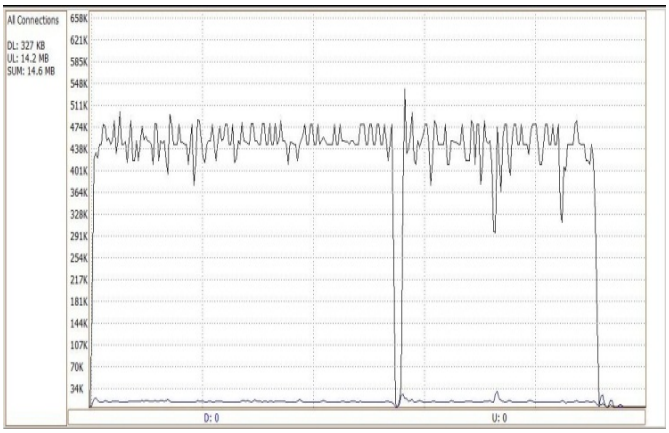



Figure 6: 13.844 MB uploaded (roughly 4.5 minutes)

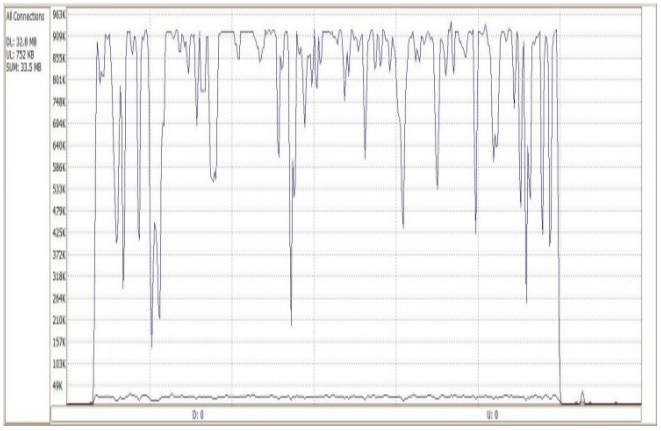



Figure 7: 34.105 MB downloaded (roughly 5.6 minutes)

Table 1 presents statistical data of the downloaded unmarked assignments. The mean file size for the four assignment submissions in the two BTech modules was 1,873,866 bytes. The high skewness values for Assignment 3 reveals (more than 3) that more than 50% of the assignments were smaller than 1.8 MB with extreme values to the right (much larger file sizes exist). The high kurtosis values (more than 11) indicates that the distribution has a sharp peak with long and fat tails. However, no statistical significant relationship exists between the file size and the final marks obtained for the assignment (p-value > 0.05). Larger file sizes do not mean that a better assignment has been submitted which will result in higher marks. Table 1 furthermore shows that only about 45% of registered students (20 of 47 for RAE4 and 22 of 46 for SCM4) submitted Assignment 3 online.

Table 1: Descriptive statistics of one traffic upload to UNISA's assignment router (J-Router)

| | RAE4 Assign 2 (n = 26) | RAE4 Assign 3 (n = 20) | SCM4 Assign 2 (n = 22) | SCM4 Assign 3 (n = 22) | Calculations |
|---|---|---|---|---|---|
| Mean (bytes) | 2,059,166 | 2,514,644 | 1,479,580 | 1,442,075 | Average 1,873,866 |
| Kurtosis | -0.232 | 11.342 | 3.576 | 12.064 | |
| Skewness | 1.044 | 3.111 | 2.040 | 3.183 | UNISA students |
| Registered students | 47 | 47 | 46 | 46 | 400,000 |
| Correlation to final mark | 0.191 | 0.362 | -0.046 | 0.306 | |
| Significance | 0.954 | 1.648 | 0.204 | 1.438 | One data upload |
| P-value | 0.175 | 0.058 | 0.420 | 0.083 | 749,546,426,084 |

Figure 8 portrays a prediction of future Internet traffic increase for online assignment submissions at UNISA. Interpolating student numbers, based on the past four years, and multiplying it by the mean file size (1.8 MB from Table 1) could result in almost 9 TB of student's assignments being uploaded and downloaded in 2020 over a 4 month period (from Figure 3). This increase is very plausible, as Cisco [27] has stated that global IP traffic has multiplied eightfold over the past five years with annual global IP traffic forecasted to exceed 1.3 ZB by the end of 2016.

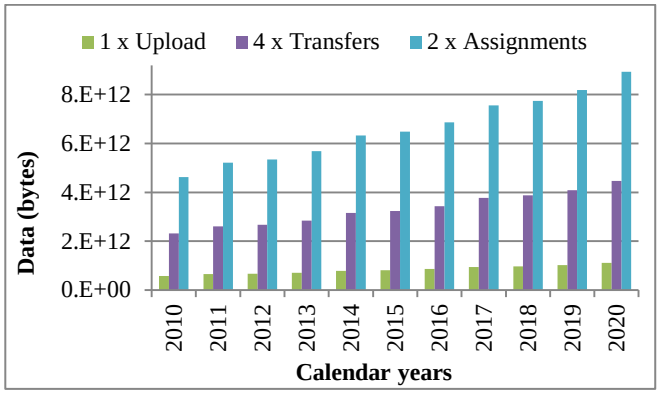



Figure 8: Future Internet traffic predictions for online written assignment submissions at UNISA

## VI. CONCLUSIONS

The purpose of this paper was to present current Internet traffic for two modules offered at UNISA within a telecommunications course, highlighting a number of future predictions. The average file size of online assignment submissions was calculated to be around 1.8 MB, with no significant correlations to the final mark given for the assignment. Furthermore, establishing a connection to

UNISA's assignment router may take between 20 seconds and more than 1 minute to complete.

The following pertinent results were found: (1) predicted average Internet traffic will grow substantially over the next few years in accordance with the number of registered students at UNISA, with June through September of 2020 expected to reach almost 9 TB of assignment data; (2) some ODL students do NOT have the necessary e-skills to effectively make use of the Internet, as they upload pdf files in excess of 16 MB to UNISA's assignment router; (3) not all ODL registered students accessed UNISA's assignment router during 2013, highlighting that many may not have the technology required or the financial means to gain access.

## VII. References


[1] South African Foundation, "Telecommunication prices in South Africa: An international peer group comparison," Occasional Paper No 1, April 2005.

[2] Business Leadership South Africa, "South African telecommunications prices: An updated international price comparison, with regulatory recommendations," Occasional Paper No 3, November 2007.

[3] B. Barry, V. Chukwuma, M. Petitdidier, L. Cottrell, and C. Barton, "Digital divide in sub-saharan africa universities: recommendations and monitoring.," presented at the atIST-Africa 2008 Conference & Exhibitio, Windhoek, Namibia., 2008.

[4] N. M. Theron. *Economic report: The Internet service provider market*. Available: http://mydrive.co.za/uploads/economic.report.ispa.pdf. Accessed: 16 January 2005.

[5] Telkom SA. *do Uncapped Basic*. Available: http://www.telkom.co.za/sites/athome/productsandservices/internetandbroadband/internetanddatabundles/&productname=do%20Uncapped%20Basic#.U80ha_mSx1Y. Accessed: 21 July 2014.

[6] Statistics SA. *Income and expenditure of households, 2010/2011*. Available: http://www.statssa.gov.za/Publications2/P0100/P01002011.pdf. Accessed: 20 March 2012.

[7] L. McLaren and E. Heath, "The public sector as a key enabler in sustainable rural tourism," African Journal of Public Affairs, vol. 5, pp. 93-104, 2012.

[8] World Wide Worx. *Internet matters: The quiet engine of the South Africa economy: Internet access in South Africa.* Available: http://www.internetmatters.co.za/. Accessed: 14 January 2012.

[9] A. Goldstuck, "Internet access in South Africa," World Wide Worx, Ed., ed, 2008.

[10] Y. S. Ko, "New technologies in implementing e-government: E-Government workshop," ed. Pretoria: University of South Africa, 2009.

[11] N. M. Ochara and T. Mawela, "Enabling Social Sustainability of E-Participation through Mobile Technology," Information Technology for Development, vol. 0, pp. 1-24, 2013.

[12] We are social. *Homepage*. Available: http://wearesocial.net/blog/2014/01/social-digital-mobile-worldwide-2014/. Accessed: 20 March 2014.

[13] J. Murray, D. Byrne, and L. Koenig-Visagie, "Teaching gender studies via open and distance learning in South Africa," Distance Education, vol. 34, pp. 339-352, 2013.

[14] A. J. Swart, "Onscreen marking: An effective assessment tool for engineering education in the information age," in ICEE/ICIT2013, Cape Town, South Africa, 2013.

[15] University of South Africa. *Homepage*. Available: http://www.unisa.ac.za/default.html. Accessed: 11 March 2013.

[16] N. M. Ochara and T. Mawela, "Enabling social sustainability of e-participation through mobile technology," Information Technology for Development, pp. 1-24, 2013.

[17] P. Bridge and R. Appleyard, "A comparison of electronic and paper-based assignment submission and feedback," British Journal of Educational Technology, vol. 39, pp. 644-650, 2008.

[18] J. W. LeLoup and R. Ponterio, "On the Net," Language Learning & Technology, vol. 5, pp. 4-7, 2001.

[19] P. R. Selvidge, B. S. Chaparro, and G. T. Bender, "The world wide wait: effects of delays on user performance," International Journal of Industrial Ergonomics, vol. 29, pp. 15-20, 2002.

[20] B. G. Beckerman and M. D. Schnall, "Digital information management: a progress report on the National Digital Mammography Archive," presented at the International Symposium on Biomedical Optics, San Jose, California, 2002.

[21] P. A. Haleem and M. Sebastisan, "Manuscript Info Abstract," International Journal, vol. 2, pp. 251-294, 2014.

[22] P. Dillenbourg and M. Betrancourt, "Collaboration load," Handling complexity in learning environments: theory and research, pp. 142-163, 2006.

[23] NetWorx. *Homepage*. Available: http://www.softperfect.com/products/networx/manual/index.htm#intro. Accessed: 1 May 2014.

[24] R. Ashita, "Beyond Testing and Grading, Using assessment to Improve Teaching–Learning," Research Journal of Educational Sciences, vol. 1, pp. 2-7, 2013.

[25] A. Jadallah and S. Loach, "System and method for network capacity planning," ed: Google Patents, 2013.

[26] B. Irwin, I. Siebörger, and D. Wells, "Bandwidth Management and Monitoring for Community Networks," in The Southern Africa Telecommunication Networks and Applications Conference (SATNAC), Spier Estate, Stellenbosch, South Africa, 2010.

[27] Cisco White Paper, "Cicso Visual Networking Index: Forecast and Methodology 2011-2016," May 2012.



**James Swart** received his DTech: Electrical: Engineering degree in 2011 from the Vaal University of Technology. His research interests include engineering education and alternative energy.

# Securing Mobile Payments on Unsecure Mobile Devices

Rossouw de Bruin[1], Prof S.H. von Solms[2]
Academy of Computer Science and Software Engineering
University of Johannesburg, P. O. Box 524, Auckland Park 2006
Tel: +27 11 559 2967
email: debruin.rossouw@gmail.com[1], basievs@uj.ac.za[2]

**Abstract-Marking the 10 year anniversary of mobile malware, what has been done in order to ensure the security, confidentiality, integrity and non-repudiation of mobile payments? How secure are mobile payments and mobile payment applications? From early on, malicious users and cyber criminals sought ways of compromising the security of not only mobile devices, but also of mobile payments and mobile payment applications – particularly since the mobile payments industry is a multi-billion dollar industry (South Africa alone contributed R7 billion at the end of 2012). Furthermore, as a result of poor or improper technology implementations, SETA (Security Education, Training and Awareness) has a tendency to lack in its ability to enforce a strong sense of information security. This plays a crucial role in mobile payment environments, where users can unknowingly, unwittingly and/or accidentally make their credentials and/or sensitive account details available for others to snoop. In this paper, we take a look at the history of mobile malware and how it evolved into what we have today, and how it is being used in order to compromise the security and confidentiality of mobile payments and mobile payment applications. Furthermore, we will propose a solution to securing mobile payments and mobile payment applications, regardless of how unsecure the user's mobile device is.**

**Index Terms—mobile payments, mobile security, mobile malware, NFC, mobile wallets**

## I. INTRODUCTION

*"App stores and mobile apps are the greatest hostile code and malware delivery mechanism ever created."* – Winn Schwartau, chairman of MobileActiveDefense.com

From the very first wireless phone call made on April 3, 1973, we have seen one great human need being satisfied: connectivity [1]. This need has led us to improve mobile technology, where we are even able to conduct work as if we are using a desktop or laptop computer.

This, however, have not stopped our curious nature to understand the technology and determining its weaknesses, loopholes and entry points for devious activities.

Marking the 10 year anniversary of mobile malware, we have become concerned for not only our virtual safety, but also our safety in the physical world, where privacy and the security thereof is just as valuable to criminals as what our virtual value is.

As mentioned by the SANS (derived from **S**ysAdmin, **A**udit, **N**etworking and **S**ecurity) Institute, mobile device security is still not yet mature enough to provide a secure and confidential mobile payment platform [2]. Currently, SETA (derived from **S**ecurity **E**ducation, **T**raining and **A**wareness) programs are rigorously deployed, yet they are still not stopping users from downloading malicious mobile applications and/or visiting malicious web sites.

Today, we see numerous variants of mobile malware, such as adware, spyware, chargeware, ransomware, Trojans, worms and viruses. Yet, with all these variants, we are still prone to social grooming and other social engineering implementations in order to gain knowledge of our virtual selves.

In this paper, we discuss a brief history on mobile attacks and mobile malware, and how it evolved into what we have today.

In the second part of this paper, we will categorise mobile attacks into three basic categories. In this categorisation, each type of attack either focuses on compromising the security of the mobile application, compromising the security of the mobile device or compromising the security of the wireless infrastructure.

The third part of this paper proposes a solution on how we can provide extra measures in order to improve the security of mobile payments and mobile payment applications.

Lastly, to illustrate the effects of the proposed solution, we discuss four scenarios of compromisation, where each scenario attempts to compromise the mobile payment application.

## II. BRIEF HISTORY ON MOBILE ATTACKS

The history of mobile (cellular) devices started way back on April 3, 1973 where Motorola was the first company to make the first wireless phone call via a mobile phone [1]. From 1990, these mobile devices started to change into something that we use today – sleek, powerful, personal, portable and, to a great extent, an extension of our bodies.

However, when it comes to technology, we have an uncanny desire to know more, to play more and to experiment more. For mobile phone technology, this desire was born in June 2004, when Ojam engineered their mobile game *Mosquito* with an anti-piracy Trojan. It's not considered a mobile virus *per sē*, it can be seen as the instigator of mobile malware [3]. In July 2004, the first, real, mobile malware was created: SymbOS.Cabir [4]. SymbOS.Cabir was written by 29A, an international group of virus writers, that made it possible for the malware to replicate itself by making use of the mobile device's

Bluetooth technology [5]. Fortunately for the victims, SymbOS.Cabir was not harmful; it merely displayed "Cabir" on the mobile phone's display each time the mobile phone was switched on.

In August 2004, cyber criminals (although a term not yet coined at that time) saw that it was possible to use mobile malware for monetary gain. This was made possible when a cracked version of *Mosquito* made its appearance, which contained a Trojan called Trojan.Mos. Trojan.Mos would send out premium rate text messages to very expensive vendors [5]. This resulted in the victims building up very high phone bills – all without their knowledge and consent.

Cyber criminals noticed that they could easily exploit weaknesses in mobile phones. The first attack making use of such exploits, was called SymbOS.Skulls. SymbOS.Skulls would either replace icons of applications with a skull, or it would corrupt system and application files, essentially making the mobile device or application useless [5].

In 2005, SymbOS.CommWarrior entered the scene, which sent SMS messages to all the victim's contacts located in the victim's phone book [5]. Spyware.FlyxiSpy, with the aid of social engineering attacks, entered the scene in early 2006, which made it possible to monitor compromised mobile devices [5].

From 2006 until today, malware evolved into becoming more powerful, more aggressive and much more specialised.

As technology improved and started to provide more features and more possibilities, online banking began to make use of mobile devices for authentication as early as 2010 [5]. This movement to mobile devices saw cybercriminals creating malware that specifically targets mobile devices [5]. One such malware variant is SymbOS.ZeusMitmo which is used to compromise the authentication messages, enabling the cybercriminal to log into the victim's online banking account [5].

The first Android botnet to be created, was created in 2011 and is called Android.Geinimi [5]. Android.Geinimi explicitly targets the Android family where it made use of click-fraud and premium rated text messages [5]. The very first mobile botnet was created in 2009 [6].

Late in 2013, Symantec detected a variant of mobile malware – known as mobile ransomware – called Android.Fakedefender. Android.Fakedefender masqueraded itself as an anti-virus application (also called a fake anti-virus) called Android Defender.

However, due to compatibility issues with various Android mobile devices, the success of this mobile ransomware application is questionable.

The types of attacks these malware deploys can essentially be divided into three categories (discussed in the next section), where they aim to either compromise the security of the mobile device, to compromise the security of the mobile application or to compromise the security of the wireless networks.

## III. Categorising Mobile Attacks

When it comes to mobile payment applications, we can essentially group the mobile payment applications into three categories:

a. Those that make use of mobile wallets, such as PayPal, Google Wallet and Isis Wallet,
b. Those that are used as money transfer and micro-financing services, such as M-Pesa, and
c. Those that are connected to existing mobile banking applications which allows transfers to be made within is specific proximity, such as First National Bank (FNB) South Africa's mobile banking and mobile payment application.

These applications are useful and successful in their own right, but none of them are without security risks and threats.

Figure 1 illustrates a common model for mobile payment systems making use of mobile wallets. It further indicates that there are essentially three aspects of mobile payments and mobile payment applications that can be compromised:

a. The mobile device and the applications it contains, as referred to by point C in Figure 1,
b. The communication that takes place between a payer and his mobile wallet and between a payee and his mobile wallet, as referred to point A in Figure 1, and, lastly
c. The communication that takes place between the payer and payee's bank (provided that they use two different bank) and between the mobile wallets and the banks, as referred to by point B in Figure 1.

These three aspects, and how they can become compromised, are discussed further in this section.

### A. Compromising Mobile Applications

Referring to point C in Figure 1, mobile devices and mobile applications are one of the three aspects of mobile payments and mobile payment applications that can become

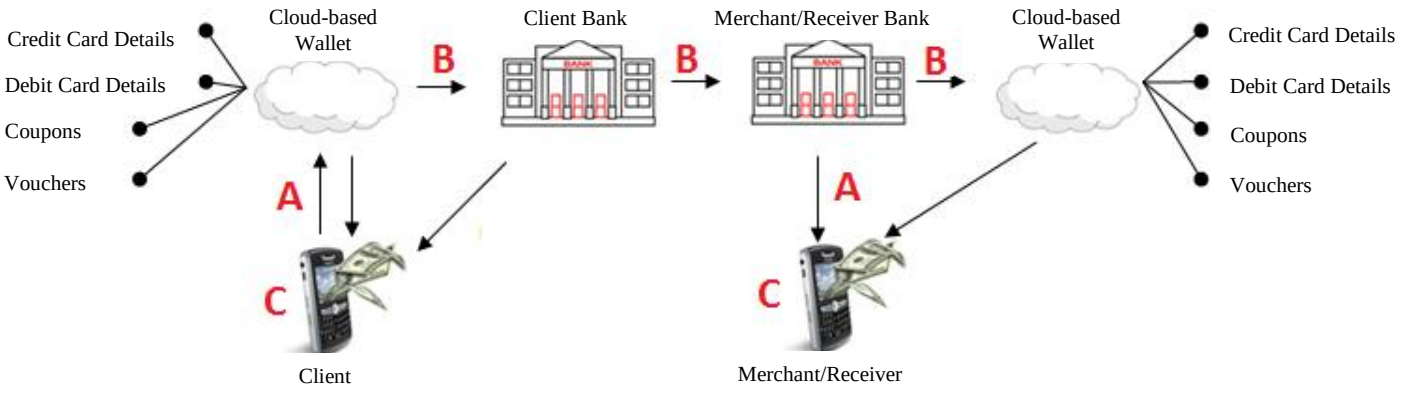



*Figure 1 Mobile Payment Systems [7]*

compromised. These attacks can be relatively simple or complex – each with its own purpose and motivation.

Compromising mobile applications can be done in numerous ways, but each of these essentially falls into two sub-categories, namely, malware and repackaging.

By making use of malware (and we will includes social engineering in this term), we can attempt to compromise the log in credentials of the victim. These compromised log in credentials could then be used to compromise the financial profile of the victim and also to transfer funds into our own accounts. By making use of Trojans, probably the most dangerous and effective strategy is to gain root access, or elevated privileges, of the mobile device (as discussed more in section B).

By making repackaged applications – as was the case with August 2004 Trojan contained in *Mosquito* – cyber criminals can fool us into download legitimate looking and even enticing applications. These repackaged applications can be used to compromise mobile application and mobile device security. Repackaged attacks are possible due to the way in which mobile applications are stored and how we can access them in order to download and install them.

With numerous third party websites – and even legitimate application stores – making mobile applications available, malware can easily be embedded within mobile applications. The victim may not even be aware that they have downloaded and installed a compromised mobile application.

### *B. Compromising Mobile Devices*

Compromising the actual mobile device and not the mobile applications contained within the device, can again be broken up into two sub-categories: hard attacks and soft attacks.

In hard attacks, theft – including tampering – and misplacement of the mobile device are the only real threats. Although these threats can be prevented in some circumstances, it still does happen. When a mobile device is stolen or lost, it is more than just the mobile device that can be compromised. Each application or file contained on the mobile device can be compromised.

We can consider soft attacks as the virtual equivalent of hard attacks. These attacks can be likened to jailbroken and rooted devices (rooting a mobile device is similar to jailbreaking a mobile device, the difference between the two is the operating system being targeted. For sake of simplicity, when we refer to jailbreaking, we also refer to rooting). Although the owner of the mobile device can decide to jailbreak the mobile device himself, there are instances where a malicious user can jailbreak a device. This is effectively done by making use of malware. When a mobile device is jailbroken by the user, numerous potential security holes are opened up for exploitation. When a mobile device is jailbroken by malware, the malicious user typically wants to gain super user access as well as remote control access [8]. When a malicious user jailbreaks a victim's mobile device, the malicious user can remotely view everything the victim does – the malicious user can even obtain keystroke information on password fields in order to compromise an application remotely.

### *C. Compromising Communication Mediums*

Referring to point A and B from Figure 1, the communication media used to transfer information from the payee to his mobile wallet, from the payer to his mobile wallet, from the mobile wallet to the bank and from the bank to the mobile wallet, are the lifeblood of modern mobile payment applications.

With its new and successful product forming part of their mobile banking suite, FNB's mobile banking application diverts greatly from this original concept of using mobile wallets. FNB's mobile banking application offers clients to send payments to other clients, by making use of GeoPayments [9].

GeoPayments uses the user's GPS, GPRS and/or Wi-Fi information in order to obtain a list of potential users within a range of 500 m [9].

This is a large range with a big scope of potential security risks and security threats.

Compromising communication media essentially targets two components of wireless communication: the actual infrastructure and the wireless medium.

By compromising the infrastructure, a malicious user can use a rogue access point in order to obtain confidential information or even infect the user's mobile device with malware [10].

With the way and tempo in which mobile devices evolve and improve, they can be used as wireless access points (WAPs). If a compromised mobile device is used as a WAP, a malicious user can just as easily obtain confidential information as if he had compromised a wireless sensor, node or antenna [7].

By compromising the wireless medium, a malicious user can easily snoop the data packets that are sent from the origin to the destination. Also, a malicious user can easily divert, relay and replay data packets from the origin to two destinations; the malicious user's destination and the original intended destination.

The malicious user can then more easily analyse the data packets and determine how they are encrypted. Once this information has been obtained, the malicious user can compromise additional data packets in order to obtain more information. Besides obtaining additional information, the malicious user can change the contents of the data packets and send the modified data packets to the recipient.

These types of attacks all pose a serious threat to the security of mobile applications and mobile payments.

The most severe concern for mobile payments and mobile payment applications, is the way in which users can customise their mobile devices. If the user improperly customise their mobile device, the mobile device can become ridden with malware, which might not even be known to the user.

The way in which we can attempt to provide security, is not by stopping information from being compromised. Rather, we can secure the information in such a way that, if it does become compromised, the malicious user would have great difficulty in obtaining the original contents of the information

In the next section, we will propose a method of securing

mobile payments and mobile payment applications that is based on this principle.

## IV. Proposed Prototype of Securing Mobile Payments

As discussed previously, mobile malware can be grouped into three categories, where each category aims at compromising certain elements of mobile communication.

We are fighting an evolving force, as can be seen in a 2013 analysis done by Lookout Security. In this analysis, mobile threats consisting of adware, chargeware and malware are the most prevalent in countries such as Asia, Russia and parts of Eastern Europe and Africa [11].

North American countries and Japan had by far the lowest rate of mobile threats. This is due to North American countries making use of mobile application stores to download mobile applications. Also, Japan employs a set of strict regulations governing mobile application stores [11].

This begs the following question: with so many individual variants of malware and high prevalence rate, how does one go about to secure mobile applications – and, more particularly, mobile payment applications?

Should we go for a complete third-party shutdown and only allow downloads to be made from mobile application stores such as Amazon Appstore, Apple's App Store, BlackBerry World, Google Play, Nokia Store, Windows Phone Store, etc.?

Or is strict regulation the way to ensure that no mobile malware will ever run wild?

Each of these are efficient in their own right however, when it comes to Internet and connectivity, we – as individuals and companies – want freedom and a fair playground.

Keeping this in mind, how does one go about securing mobile applications without hindering a user's Internet freedom?

This paper aims at providing a means of assuring the user that their mobile payment information is kept secure and confidential. Also, we aim to provide assurance to the user that their mobile usage and connectivity experience is not hindered in any which way.

In [7], "*A Wallet-less Mobile Payment System using Near Field Communication (NFC)*" a prototype mobile payment systems making use of Near Field Communication (NFC) is discussed.

Using this prototype, a payer (an individual) can pay a payee (either an individual or a merchant), by making use of NFC. The payee would transfer their payment information (consisting of account information and the amount payable) to the payer. The payer would send this information to his bank, where the payment is executed.

In this prototype, as indicated in Figure 2, the need to use a mobile wallet has been eliminated. The payment information is transferred from the payee to the payer in a 0 - 5 cm range, which is much less than FNB's 500 m range.

Looking at the security risks as illsutrated in Figure 1 and Figure 2, we can see that the risks between the user and his mobile wallet has been eliminated. However, there still are security threats posed to the mobile application, mobile device and communication between the payer, payee and their bank.


*Figure 2 Possible Security Risks and Threats Present in Mobile Payment Applications [7]*

These security risks will always be inherent in any mobile payment application.

Therefore, the methodology employed in this prototope is not to eliminate these security threats and risks. Rather, the methodology aims to make the transmitted information and access to the mobile payment application as secure as possible. The idea is to make it extremely difficult for the criminal to obtain the original contents of the transmitted information.

This methodology makes use of a strong and efficient security and cryptography platform, where a guarantee about the security, confidentiality, integrity and non-repudiation of an activity can be made.

To combat the threats that attempt to compromise the security of the mobile application and mobile device (particularly threats that attempt to obtain key stroke information and remote control attacks), as illustrated in point B of Figure 2, the prototype makes use of complex identification, authentication and authorisation mechanisms. These mechanisms consist of a username, password, authentication tag and information contained within the mobile device.

The authentication tag is linked to the user and his banking profile maintained by his bank. Also, the mobile device is linked to registration information used to register for the application as well as with the authentication tag. These links associated with the authentication tag and mobile device creates a unique relationship between the authentication tag and mobile device. This relationship ensures that if either of the two elements is missing, the user would not be able to log into his mobile payment application.

To combat the threats that attempt to compromise the wireless communications between the payee and the payer, between the payer and his bank and between the payee and his bank, the prototype makes use of the same authentication tag and mobile device to encrypt the information before it leaves the mobile device. Also, information sent from the payee to the payer cannot be decrypted by the payer. Only the bank can decrypt the payment information from both the

payer and the payee.

To further aid in the security of transferring payment information from the payee to the payer, NFC has been selected as the preferred means of transmission. This is due to the speed at which information can be transmitted, the type of information that can be sent – which aids in securing and obscuring information – and the small range within which information can be transmitted [12].

This prototype welcomes the fact that authentication tags and mobile devices can be lost or stolen and that wireless communications can be intercepted. What the prototype aims at accomplishing is that, if the authentication tag and/or mobile device does become compromised, a malicious user would not be able to log into the mobile payment application. Likewise, if the communication does become compromised, the malicious user would have great difficulty – and, even near impossibility – to obtain and use the original contents.

The next section will discuss four scenarios of how the mobile application can be compromised and how the proposed solution attempts to handle it.

## V. Outcome of Prototype

To illustrate how the proposed solution can aid in making mobile payments and mobile payment applications more secure, we will look at four scenarios aiming at compromising the security of the mobile application.

In the first scenario, a malicious user obtains the username and password his victim uses in order to log into his mobile payment application. Using this information on his mobile device, the malicious user would be unable to log into the mobile application. This is due to the malicious user not having the victim's authentication tag and he is not using the victim's mobile device in order to log into the mobile payment application.

In the second scenario, the same malicious user also obtains the victim's authentication tag. By using his mobile device and the victim's username, password and authentication tag in order to log in, the malicious user would still be denied entry. This is due to the malicious user not using the victim's mobile device in order to log into the mobile payment application.

In the third scenario, the malicious user obtains not only the victim's username and password, but also his mobile device. Using the victim's username and password on the stolen mobile device, the malicious user is once again denied entry. Without the victim's authentication tag, the malicious user cannot log into the mobile payment application.

In the fourth and last scenario, the malicious user obtains the victim's username, password, authentication tag and mobile device. With all four elements, the malicious user can successfully log into the victim's mobile payment application. The victim would need to inform his bank, whereby the bank would send a request to the mobile payment application to uninstall itself from the mobile device. Additionally, the victim's bank would disable all the victim's information used for logging in purposes.

## VI. Conclusion

Mobile devices, as with any piece of modern technology, have greatly improved how we connect with the real and virtual world. Joining meetings and holding conferences at two opposite ends of the world, were never possible before the technology kicked in.

Mobile technology improved even more upon this and we are even using mobile devices to conduct business and making payments, ensuring that the mobile payment industry remains lucrative.

It is even lucrative in the criminal space. Keeping their eyes on a multi-billion dollar industry leads to more aggressive and efficient attempts in getting – and keeping – their hands in the honey jar.

In this article, we have briefly discussed the history of mobile attacks, mobile malware and the evolution into what we have today.

We have also proposed a solution into making mobile payments and mobile payment applications more secure, which can aid in securing the user's financial, virtual and physical well-being

## VII. References


[1] Goodwin R. The History of Mobile Phones: 1973 to 2007, 2013 [Internet]. Accessed 16 May 2014. Available from: http://www.knowyourmobile.com/nokia/history-mobile-phones/19848/history-mobile-phones-1973-2007.

[2] SANS Institute InfoSec Reading Room. Security of Mobile Banking and Payments; 2012. Available from: http://www.sans.org/reading-room/whitepapers/ecommerce/security-mobile-banking-payments-34062

[3] Cochrane A. Mobile Telephony Attacks; 2012. Available from: http://www2.cs.uidaho.edu/~oman/CS336/Cochrane_MobileTelephonyHacks.pdf

[4] Wueest C. A brief history of mobile malware, 2014 [Internet]. Accessed 16 May 2014. Available from: http://www.symantec.com/connect/blogs/tenth-anniversary-mobile-malware.

[5] Clooke R. The Tenth Anniversary of Mobile Malware, 2013 [Internet]. Accessed 16 May 2014. Available from: http://www.mobilecommercedaily.com/a-brief-history-of-mobile-malware.

[6] Maslennikov D. Mobile Malware Evolution: Part 6, 2013 [Internet]. Accessed 16 May 2014. Available from: http://www.securelist.com/en/analysis/204792283/Mobile_Malware_Evolution_Part_6.

[7] de Bruin R. A wallet-less mobile payment system using near field communication (NFC). Master Dissertation, available from the University of Johannesburg,

[8] Siciliano R. How Does Jailbreaking Or Rooting Affect My Mobile Device Security?, 2012 [Internet]. Accessed 19 May 2014. Available from: http://blogs.mcafee.com/consumer/how-does-

jailbreaking-or-rooting-affect-my-mobile-device-security.

[9] Wilson G. FNB geo payments: all the details, 2012 [Internet]. Accessed 22 May 2014. Available from: http://www.techcentral.co.za/fnb-geo-payments-all-the-details/31691/.

[10] Choi M, *et* al. Wireless Network Security: Vulnerabilities, Threats and Countermeasures; 2008. Available from: http://www.sersc.org/journals/IJMUE/vol3_no3_2008/8.pdf

[11] Lookout. Mobile Threats, Made to Measure: The Specialization of Mobile Threats around the World; 2014. Available from: https://www.lookout.com/resources/reports/mobile-threat-report

[12] Nosowitz D. Everything You Need to Know About Near Field Communication, 2011 [Internet]. Accessed 19 May 2014. Available from: http://www.popsci.com/gadgets/article/2011-02/near-field-communication-helping-your-smartphone-replace-your-wallet-2010/.

**Rossouw de Bruin** received his BSc IT degree in 2010 from the University of Johannesburg, South Africa. In 2012, he received his BSc IT (Hons) degree *cum laude* at the same university. He received his MSc IT degree *cum laude* from the same university. He's MSc IT focused on mobile payments and how to secure mobile payments.

# Mobile Health Monitoring System for Community Health Workers

George Sibiya, Ishmael Makitla, Samuel Ogunleye, Thomas Fogwill and Ronell Alberts
Meraka Institute,
CSIR, 627 Mering Naudea Road
Brummeria, Pretoria, 0001
Email: {gsibiya,imakitla,oogunleye,tfogwill,ralberts}@telkom.co.za

**Abstract: The leading global cause of high mortality has been identified to be chronic non-communicable diseases such as stroke, cancer, chronic respiratory conditions, heart disease, diabetes etc. These diseases affect communities both from the rural and urban areas. Deaths as a result of these diseases are relatively higher in rural communities as most of them have limited access to health care services. The limited access to health care services by rural communities is due to the difficulty experienced by governments in bringing those services to rural communities. Contributing factors to the difficulty is the inaccessibility of those communities and shortage of resources such as health professionals. The authors in this paper share the sentiment that community health workers can play a major role in assisting governments to deliver health services to rural communities regardless of their location. This paper therefore presents an application that supports health workers in diagnosing and monitoring non-communicable diseases in rural communities.**

- **Keywords:** Non-communicable diseases, Android application, mobile health application

## I. INTRODUCTION AND BACKGROUND

The leading global cause of high mortality has been identified to be Chronic non-communicable diseases (NCDs) such as stroke, cancer, chronic respiratory conditions, heart disease, diabetes etc. [1]. Cardiovascular diseases (CVDs) were found to be the major cause of NCD deaths in 2008 and they account for more than 15 million deaths worldwide, a statistic which is more than twenty per cent of global mortality [2]. Furthermore, more than ten per cent of global deaths are caused by hypertension, which is the leading risk factor for mortality [2]. The dominance of hypertension in global adult population was estimated to be approximately 25 per cent in 2000 and has been predicted to increase by about 70 per cent by 2025 [3]. Research has also shown an increase in hypertension cases in rural areas [4]. Non-communicable diseases in general are affecting rural communities most, [5] and [6]. This raises concerns as governments in developing countries still have challenges in delivering health services to rural communities.

Majority of the current healthcare technologies and health diagnostic devices are used in clinics or hospital environments. Due to the difficulty to access these facilities in rural areas the underlying symptoms of NCDs cannot be adequately monitored. This could lead to difficult or even incorrect diagnoses [7] of NCDs.

There are factors that are responsible for the lack of access to health care facilities in rural and remote areas in developing countries. These factors include transportation, health insurance and income and shortage of health care providers.

In developing countries, such as South Africa, Nigeria, Kenya etc., the majority of the rural communities are situated in remote areas where access to them is limited by poor road infrastructure or geographical landscapes. These factors also hinder the roll out of health services in those areas. There limited health care facilities in such areas, results into people having to stand in long queues at clinics and hospitals, sometimes for the whole day. In many rural areas there is no access to clinics and hospitals at all and even where there is access, it usually requires citizens to travel far to reach the nearest facility [8].

Furthermore in the South African context, South Africa has diverse communities with a large part of the population that needs significant medical attention being situated in those rural areas with under resourced health facilities. The South African Census in 2011 [9] indicated that the dependant population (those aged 0 – 14 or above 65) is found mainly in rural areas. This is the population segment that generally does not have means to travel and hence has difficulty accessing health facilities. A shortage of human resources is also cited as one of the hindering issues in delivering quality health care to the citizens by the government [10]. The limited existing health resources and facilities are also burdened by the scourge of tuberculosis and HIV and AIDS. To address accessibility of health care services, the South African government has adopted and supports the deployment of community health workers in communities and schools [11]. These community health workers need low cost mobile equipment and resources for use during their visits to the community. There is therefore an urgent need for intervention in rural communities in a cost effective way. These include health care devices such as blood pressure meters and blood sugar meters, among others.

In this paper we present a mobile health application that can be used by community health workers to collect observations in remote rural areas. This application makes

use of mobile health devices to take health care services to the people regardless of their location. The solution presented in this paper takes into account the South African demographics.

The paper is organized as follows; section II presents an overview of the requirements of our proposed mobile health system. In section IV, we present a description of the mobile health application in detail. In section V we present evaluation of our proposed system. In section VI we conclude the paper and acknowledgements in section VII.

## II. RELATED WORK

Research works that are aimed at providing technological solutions for community health workers are limited. Research works in [12],[13] and [14] focus on surveys on solutions that are provided to health workers in different countries and the use of mobile technology by health workers. Research works that are closely related to our technology solution that can aid community health workers are the research found in [15],[16] and[17]. The research work by Ngabo et al. in [17] however solely focused on addressing pregnancy monitoring, maternal and child birth deaths. The work by Mena in [15] and by Grossman in [16] focuses on blood pressure monitoring using mobile technologies. There is a need for a technology solution that can combine all necessary NCD related monitoring in a single application such as glucose, cholesterol and others in addition to blood pressure. Such solutions need to also be able to transmit messages in a format compliant with the HL7 standard [18].

## III. MOBILE HEALTH APPLICATION

As delivering health services to rural communities remain a challenge for many governments, technological innovations that can increase prevention and control of NCDs are needed. Wearable health devices such as ambulatory blood pressure (ABP) monitors are a step in the right direction. ABP monitors are portable, fully functional automatic devices that record blood pressure from a subject of care (patient) for a self-determined period of time, while users conduct their daily activities [19]. This method can be effective in monitoring cases of hypertension as it provides real time information and eliminates the need to visit a healthcare facility to take blood pressure readings.

Our proposed mobile health monitoring system enables faster computerization of data that has been recorded. This improves the quality and efficiency of delivering healthcare services to rural communities far better when compared to paper based traditional data collection processes which need to be followed by transcription to computer systems [20]. It is therefore meant to provide support to community health workers in delivering health services to their communities. It allows continuous monitoring and subsequent transmission to a standalone server in order to allow both the professional healthcare provider to offer an extensive health feedback. It also allows the national health authority to have an extensive clinical database for data mining and analysis of potential risk factors that can aid in the speedy delivery of health service to the citizens.

Based on background presented in this paper and the intended environment for the proposed mobile system, the authors propose the requirements of a mobile health system to support community-based health workers are as follows:

### *A. Light weight*

As health workers travel long distances on their day to day activities, it is ideal for them to carry light weight devices. It is then required for an application meant for health workers to run on light weight devices such as tablets or smart phones.

### *B. Intermittent Internet connectivity*

Areas that are travelled by community health workers often do not have Internet or have poor Internet connectivity. An application developed to aid health workers therefore needs to have the capability to function on both online and offline modes.

### *C. Blood pressure, heart rate and glucose measurements*

This research paper contributes towards addressing delivery of NCD related health services. The application therefore needs to be able to be take and process blood pressure, heart rate and glucose readings. These reading closely related to most common NCDs.

### *D. Feedback to health worker and the subject of care*

Community health workers are often not professionally trained on health. As a result they are not expected to have an expert knowledge and ability to interpret observations that they take from subjects of care. The application therefore needs to have intelligence to interpret the observations and provide feedback to both the subject of care and the health worker. This is to cater for the cases such as in which a subject of care requires urgent medical attention. An example of such cases is a case in which a subject of care's blood pressure would be critically high or critically low.

### *E. Historical preview*

Conclusions cannot be drawn from single readings on blood pressure and/or glucose on a subject of care. A historical preview of the observations recorded from a subject of care needs to be supported for intelligence to be gathered. The mobile health application therefore needs to support ability to take subject of care observations more than once.

## IV. APPLICATION DESCRIPTION

In this section we present functional and technical descriptions of our proposed mobile health application.

### *A. Functional description*

The application provides technology for real time, dependable and intelligent health monitoring by health workers in the field. It integrates a set of wearable wireless sensors with a mobile computing device, such as a

smartphone or tablet. The sensors are attached to the subject's body, usually the finger. The reading is then performed and data are collected by the sensor. The data is transmitted to the mobile device, which analyses the data in real-time and provides immediate, personalised feedback to the health worker. The data is also sent to a remote server via the mobile device's internet connection (if and when it becomes available). From this server, healthcare professionals can access the current and historical data of a subject of care to provide expert medical feedback and to support in clinical decision making. This data could also be sent to the national health authority to enable better planning and more effective and efficient allocation of healthcare resources. Figure 1 shows a communication among components in our mobile health system.


**Figure 1: Mobile health system communication**

### *B. Technical Description*

In this section we describe the technical features of the mobile health monitoring application. This features show how the function requirements described in section II are addressed.

The application is implemented as an Android [21] application. It makes use of Java, SQLite and is designed to run on any Android v4.2+ powered mobile device. The application integrates wirelessly with an ambulatory blood pressure meter, a pulse oximeter, a stethoscope and a glucometer.

Data is received from these sensors using a specialised protocol over a Bluetooth connection. The application includes a specialised protocol stack to handle the interpretation of a health device's data transmission protocol and to handle the sending and receipt of messages to and from each device.

Figure 2 depicts a code snippet of a protocol stack in this case representing a Stethoscope protocol stack. The model of a Stethoscope used as a test for this application works as follows:

- A Bluetooth connection between the Stethoscope and the mobile Android device is established.
- A user selects and uploads a file from the Stethoscope screen and the file transmission begins immediately.
- While file transmission is in progress, the application waits for the command `0x0F` (`TRANSMISSION_COMPLETE_COMMAND` variable in the code snippet) which signals the end of transmission of the file.
- On receipt of the 0x0F command, the cached data in the buffer are processed.

Data received from sensors and into the mobile Android device are processed by the central server to decide on messages to send to either a health worker or the subject of care. If abnormal readings are detected, the health worker is notified through a text message. This intelligence is only implemented on the server side to minimize processing on the mobile device which would consume more battery power. Critical and life threatening observation readings such as abnormally high or low blood is processed on-board to cater for areas without Internet connections.

Data obtained from the sensors are also geo-coded using the mobile Android device's Global Positioning System (GPS), if available. The data is then uploaded to a central server and linked to the subject of care (patient). This provides a historical log of readings taken for a particular subject of care.

```
  while(availBytes != 0 || !uploadComplete){
          if(dinput.available() >0){
             state  =  (state  ==  -1?  0:
state);
             int          length          =
dinput.available();
             while     ((state+length     <=
maxLengthOfBuffer)      &&      (read      =
dinput.read(testData, state, length))>0){
                  if(     (testData[0]     ==
(byte)0xB0  ||  testData[1]  ==  (byte)0xB0)  &&
!sent){
         state = 0; sent = true;
mmOutStream.write(StethoscopDeviceCommand.STA
RT_UPLOAD_COMMAND());
          continue;  }
  . . .
  if(state          >          0          &&
Arrays.copyOfRange(testData,                2,
4).equals(TRANSMISSION COMPLETE COMMAND)){
```

**Figure 2: Stethoscope protocol stack**

The Android mobile application uses Internet connection, when available, to incrementally upload data in JavaScript Object Notation (JSON) [22] to the central server using the Hypertext Transfer Protocol (HTTP). An example of a message in JSon format is as depicted in Figure 3.

The data representation is compliant with the HL7 [23] messaging format. The application supports intermittent network connectivity and fault tolerant upload of this data – for instance, if Internet connectivity is lost during an upload, it allows for the incomplete upload to be resubmitted later when connectivity is re-established.

On the central server, the received data are stored in a database and linked to the subject of care's electronic

record. There is a secure web application which allows authorized health professionals to view this data, to analyse it and to provide expert medical feedback.

{"registration":{"other":{"geometry":{"type":"Point","coordinates":[28.287718199635652,-25.743805733583176]},"language":"Afrikaans"},"addresses":[{"postCode":"01678","addressLines":[{"addressLine":"45 Tristan Ave"}],"countryIdentifier":"ZA","stateTerritoryProvince":"GT","suburbTownLocality":"Groenewaard"}],"electronicCommunications":[{"medium":"Mobile (cellular) telephone","detail":"0721234567"}],"identifiers":[{"designation":"8756778945"}],"names":[{"familyNames":[{"surname":"Msiza"}],"givenNames":[{"sequenceNumber":"1","name":"Lindokuhle"},{"sequenceNumber":"2","name":""}]}],"additionalDemographicData":{"sex":"Male","birthDate":"1995\/01\/05"}},"created_at":"2014\/01\/05 02:07:07","chw_id":"1","visit_id":"77e7cec3-952c-

**Figure 3: Subject of care registration information**

## V. MOBILE HEALTH APPLICATION EVALUATION

The application was evaluated in a lab environment. The purpose of the evaluation was to test how far the application satisfies the functional requirements presented in section II. A scenario used while evaluating the application is:

Consider a community health worker designated to work with a health facility say a clinic. The health worker will by default have their details registered with an online database. The community which the health worker will be responsible for is the community to which the clinic is designated to. The health worker registers online before they can be able to sign in into the mobile application while making home visits.

While doing the home visits, a subject of care may already exists in an online database and linked to the clinic. If the subject of care already exists, the health worker selects the subject of care's name on the application and proceeds with taking observations. If the subject of care does not exist, the health worker will need to register and upload the new subject of care details. The health worker then selects the subject of care and takes the observations. Finally, the health worker uploads the observations linked with the subject of care. In absence of Internet connection, the application needs to store the observations on the local file system. On establishment of Internet connection, the application needs to upload any pending observations from the local file system to the central server.

The application was then evaluated based on this scenario. The application was found to satisfy all requirements in section II. It satisfies the light weight requirement as it runs on Android devices which normally come as smart phones and tablets. It satisfies the intermittent Internet connectivity requirement by having a background process that monitors the Internet connectivity status. On establishment of Internet connectivity, the process initiates the uploading process of any pending observation stored locally on the device. The feedback requirement is satisfied by utilising the MOBI4D service [24] to deliver notifications to the health worker and the subject of care. Historical preview is enabled by linking each observation upload to the server with an individual subject of care.

## VI. CONCLUSION AND FUTURE WORK

In this paper an application that is meant to support community health workers is presented. Requirements for the application to effectively support a health worker are presented. The functional and technical description on how the application satisfies the requirements are presented. The evaluation of the application in a laboratory environment is also presented.

As part of future work, the application will be rolled out and evaluated further in the field. The evaluation criteria of the application will be designed using the guidelines presented in [25]. Post-test questionnaires will be prepared for the field workers to give feedback after using the application. Results obtained here will be used to improve the application. The performances of the application in different devices with different technical specifications still need to be carried out as well. This will be done to ensure that even health workers in possession of low end mobile devices will still be able to use the application efficiently.

The mobile application presented in this paper is only part of a larger health project that seeks to integrate technological health solution in South Africa. More developments are envisaged in this direction.

## VII. ACKNOWLEDGEMENTS

We acknowledge contribution to this research by Telemedicine Africa [26]. Telemedicine has contributed by providing devices that were used to test this application which included glucometers, ambulatory blood pressure meters, stethoscopes and pulse oximeters. The development of this application and the research presented on this paper is made possible through contributions by other team members which are Mathew Chetty, Johan van Zyl, Michael Ofori-Appia, JP Tolmay, Gugu Khalala, Marna Botha, Ilse Viviers and Ofentse Mokhuane.

## VIII. REFERENCES

[1] A. Alwan, D. R. MacLean, L. M. Riley, E. T. d'Espaignet, C. D. Mathers, G. A. Stevens, and D. Bettcher, "Monitoring and surveillance of chronic non-communicable diseases: progress and capacity in high-burden countries," *The Lancet* , vol. 376, no. 9755. pp. 1861–1868, 27-Nov-2010.

[2] W. H. Organization, *Global health risks : mortality and burden of disease attributable to selected major risks*. Geneva, Switzerland: World Health Organization, 2009.

[3] P. M. Kearney, M. Whelton, K. Reynolds, P. Muntner, and P. K. Whelton, “The Global burden of hypertension: analysis of worldwide data,” *Lancet*, vol. 365, pp. 217–223, 2005.

[4] S. Pastakia, S. Ali, J. Kamano, C. Akwanalo, S. Ndege, V. Buckwalter, R. Vedanthan, and G. Bloomfield, “Screening for diabetes and hypertension in a rural low income setting in western Kenya utilizing home-based and community-based strategies,” *Global. Health*, vol. 9, no. 1, p. 21, 2013.

[5] A. Bhagyalaxmi, T. Atul, and J. Shikha, “Prevalence of risk factors of non-communicable diseases in a District of Gujarat, India.,” *J. Health. Popul. Nutr.*, vol. 31, no. 1, pp. 78–85, Mar. 2013.

[6] B. M. Mayosi, A. J. Flisher, U. G. Lalloo, F. Sitas, S. M. Tollman, and D. Bradshaw, “The burden of non-communicable diseases in South Africa,” 2009. [Online]. Available: http://www.sudafrica.cooperazione.esteri.it/utlSudafrica/EN/download/pdf/The burden of non-communicable diseases in South Africa.pdf. [Accessed: 04-Jan-2014].

[7] O. Aziz, B. Lo, J. Pansiot, L. Atallah, G.-Z. Yang, and A. Darzi, “From computers to ubiquitous computing by 2010: health care,” *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 366, no. 1881, pp. 3805–3811, Oct. 2008.

[8] T. E. Allison, “FACTORS AFFECTING ACCESS TO HEALTH CARE FOR RURAL ARIZONA MINORITIES,” 2005. [Online]. Available: http://www.nursing.arizona.edu/Library/Allison_TE.pdf. [Accessed: 28-Nov-2013].

[9] “Census 2011.” [Online]. Available: http://www.statssa.gov.za/publications/P03014/P030142011.pdf. [Accessed: 02-Aug-2014].

[10] R. Visser, R. Bhana, and F. Monticelli, “The National Health Care Facilities Baseline Audit: National Summary Report.” Health Systems Trust, 2013.

[11] N. Languza, T. Lushaba, N. Magingxa, M. Masuku, and T. Ngubo, “Community health workers a brief description of the HST experience,” 2011. [Online]. Available: http://www.healthlink.org.za/uploads/files/CHWs_HSTexp022011.pdf. [Accessed: 04-Jan-2014].

[12] B. BUEHLER, R. RUGGIERO, and K. MEHTA, “Empowering Community Health Workers with Technology Solutions,” *IEEE TECHNOLOGY AND SOCIETY MAGAZINE*, pp. 44–52, 2013.

[13] R. Braun, C. Catalani, J. Wimbush, and D. Israelski, “Community Health Workers and Mobile Technology: A Systematic Review of the Literature,” *PLoS One*, vol. 8, no. 6, 2013.

[14] “Taking Innovation to Scale: Community Health Workers, Promotores, and the Triple Aim.” [Online]. Available: https://www.phi.org/uploads/application/files/dwjet18q0tvqvzg9iwizi6ts5shmektcxn9tntu7rrp5tugfk5.pdf. [Accessed: 01-Aug-2014].

[15] L. J. Mena, V. G. Felix, R. Ostos, J. A. Gonzalez, A. Cervantes, A. Ochoa, C. Ruiz, R. Ramos, and G. E. Maestre, “Mobile Personal Health System for Ambulatory Blood Pressure Monitoring,” *Comput. Math. Methods Med.*, vol. 2013, 2013.

[16] E. Grossman, “Ambulatory blood pressure monitoring in the diagnosis and management of hypertension.,” *Diabetes Care*, vol. 36 Suppl 2, no. Supplement_2, pp. S307–11, Aug. 2013.

[17] F. Ngabo, J. Nguimfack, F. Nwaigwe, C. Mugeni, D. Muhoza, D. Wilson, J. Kalach, R. Gakuba, C. Karema, and A. Binagwaho, “Designing and Implementing an Innovative SMS-based alert system (RapidSMS-MCH) to monitor pregnancy and reduce maternal and child deaths in Rwanda,” *Pan Afr. Med. J.*, vol. 13, no. 31, 2012.

[18] “Health Level Seven International - Homepage.” [Online]. Available: http://www.hl7.org/. [Accessed: 02-Aug-2014].

[19] T. G. Pickering, D. Shimbo, and D. Haas, “Ambulatory Blood-Pressure Monitoring,” *N. Engl. J. Med.*, vol. 355, no. 8, pp. 850–851, Aug. 2006.

[20] R. Shahriyar, M. F. Bari, G. Kundu, S. Ahamed, and M. M. Akbar, “Intelligent Mobile Health Monitoring System (IMHMS),” in *Electronic Healthcare SE - 2*, vol. 27, P. Kostkova, Ed. Springer Berlin Heidelberg, 2010, pp. 5–12.

[21] Google, “Android.” [Online]. Available: http://www.android.com/. [Accessed: 05-Jan-2014].

[22] C. Spence, J. Devoys, and S. Chahal, “Architecrutre Software as a Service for the Enterprise,” 2009.

[23] “ISO/TS 22220:2011 - Health informatics -- Identification of subjects of health care.” [Online]. Available: http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=59755. [Accessed: 09-May-2014].

[24] I. Makitla and T. Fogwill, “Mobi4D: Mobile Value-Adding Service Delivery Platform,” in *IT Convergence and Services SE - 6*, vol. 107, J. J. Park, H. Arabnia, H.-B. Chang, and T. Shon, Eds. Springer Netherlands, 2011, pp. 55–68.

[25] mHimss, “Selecting a Mobile App: Evaluating the Usability of Medical Applications,” 2012.

[26] “Telemedicine Africa.” [Online]. Available: http://telemedafrica.co.za/. [Accessed: 06-Jan-2014].

# A Consumer Health Informatics Application for e-Health Interventions in Marginalised Rural Areas

Chikumbutso Gremu, Alfredo Terzoli and Mosiuoa Tsietsi
Department of Computer Science
Rhodes University, P. O. Box 94, Grahamstown 6140
Tel: +27 46 6038291, Fax: +27 46 603 7608
email: g12g1792@campus.ru.ac.za. {a.terzoli, m.tsietsi}@ru.ac.za

**Abstract-Providing disadvantaged communities with health information can help them make informed decisions, which can lead to improved health. Health individuals do not frequently seek medical help and this relieves stress on health resources. However, for the information to be useful, it should be relevant, timely, concise, and in a language the recipients understand. Research-based communication strategies should be used to select health information to ensure that it is relevant to the target audience. This paper presents work done to develop a health informatics application that uses research-based communication strategies to disseminate health information to the target audience. The target users are health service organisations such as the South African Department of Health (DoH) that will produce health content and disseminate it to the people living in the marginalised rural area of Dwesa in the Eastern Cape Province of South Africa. The application consists of two components called the Dashboard for uploading health information, and the HealthMessenger for dissemination of the information. The HealthMessenger is built to run on top of TeleWeaver, an application integrator, currently being developed by the Telkom centres of Excellence at Rhodes and Fort Hare Universities. Apart from being designed to host applications for use in marginalised areas, TeleWeaver is aimed at supporting a business model to monetise access channels to marginalised users, in order to support ICT infrastructure development in their areas. Unit, integration, and system tests were conducted on the application, which verified that it is working as expected.**

**Index Terms—Application Server, Consumer Health Informatics, Internet-Based Interventions, Targeted Health Communication, TeleWeaver**

## I. INTRODUCTION

The advancement of the Internet technology has led to the proliferation of online content on different subjects including health. Health information consumers are now accessing a wide range of health topics on the Internet independent of their physical location or time zone. In this paper, we define a health information consumer or in short a consumer as a person who seeks information on how to stay healthy, prevent diseases, treat specific health conditions, and manage various health conditions and chronic diseases [14]. It includes persons with specific health conditions, their friends and family and also the public concerned about health.

The Internet is facilitating unprecedented access to unlimited general and specific health information for people of all ages, socio-economic statuses, and professions [1]. This has transformed the traditional "patient-provider" roles from a "paternalistic archetype" to a "participatory model" [15]. In a paternalistic archetype, healthcare professionals or providers are custodians of specialised knowledge and information (e.g. diagnostic, prognostic and treatment options [16]) and they make most of the health decisions for their patients, while in a participatory model, decision-making is based on the consensus of both the healthcare professional and the patient.

The paradigm shift in the consumer-provider relationship has opened up new opportunities for providers and consumers to partner in combating health-related challenges. As discussed by McMullan [16], consumers with access to the Internet usually go online to search for health information after seeing a healthcare professional to confirm his/her suggestions, decisions or recommendations or to gather additional information. When they share and discuss their findings with the healthcare professional, they both benefit from the expansion of knowledge and further understanding of how the health issue can be handled. Healthcare professionals also direct their patients to reliable online information to supplement what they provide them.

Despite the ubiquity of online health information, it is not equitably accessible to everyone. Some people are facing challenges of lack the of means of access (e.g. the availability of computers or mobile phones with Internet connection), high cost of access, content being in a language they do not understand, and limited knowledge on how to use computers and the Internet, or they lack of knowledge of the existence of online health information. These challenges are more prevalent in the rural areas because most people are of low economic status and hence cannot afford the costs of accessing the Internet, possess low literacy levels, and lack formal training in the use of technologies like the Internet.

In South Africa, Internet usage is low among people living below the poverty line (19%), and with low income (23%) despite the large number of new users coming from this section of the population [4]. This contributes to the digital divide, which is "the growing gap between persons who can and cannot benefit from the proliferation of online health information" [11, p.473]. Some of the factors contributing to this are the challenges discussed in the previous paragraph. To counter the challenges, some efforts are being made to bring ICT to the rural populations with the aim of uplifting their socio-economic statuses. Among these

efforts are those being made by the Siyakhula Living Lab (SLL), which are discussed in Section II A.

This paper presents research work carried out with the aim of designing and developing a health informatics application to enable Health service organisations such as the Department of Health (DoH) to disseminate and gather health information from people living in Dwesa, a marginalised rural area in the Eastern Cape province of South Africa. Information dissemination uses research-based health communication strategies (discussed in Section II D) to ensure that the target recipients access or receive only the information that is relevant to them. In addition to dissemination of health information, the application is developed to help generate revenue to support the SLL activities in Dwesa. Revenue generation is based on a business model supported by TeleWeaver (TW), an application service integration platform designed to host applications developed to be used in marginalised areas. In the next section we will present background and related work. This is followed by a presentation of the design, implementation, and testing of the application, a discussion of the work, and a proposal for work to be done in future respectively and the last section concludes.

## II. Background and Related Work

### A. *The Siyakhula Living Lab*

The SLL is a project by the Telkom Centres of Excellence at Rhodes University and the University of Fort Hare, in collaboration with partners and sponsors from the government and the industry [19]. Its design adopts the emerging Research Development and Innovation living lab methodology which aims at co-creating solutions with empowered users. It was launched in 2005, with Dwesa in the Mbashe municipality in the vicinity of the Dwesa-Cwebe Nature Reserve, as the main, deep rural test site. As reported by Dugmore [5], the project uses schools in the vicinity as points-of-presence to enable communities in the area to access computers with Internet connection. Schools were selected because they provide a neutral access point for the entire community, suitable teaching and learning structures, and they were one of the first structures to be connected to the national electricity grid. As of the year 2012, 17 schools in the area were connected wirelessly to each other and reasonably high speed Internet was available at all the schools for use by the community members. Through community engagement activities facilitated by the project, approximately 200 community members and 4500 learners drawn from the 17 participating schools were trained to use computers and browse the Internet.

### B. *TeleWeaver*

As discussed in Section I, TW is an application service integration platform designed to host applications developed to provide services to people living in marginalised areas. It was proposed after observing that researchers in the SLL were developing and implementing applications as individual projects and were reinventing the wheel instead of building on already existing projects [8]. Its design enables applications its hosts to interact by exchanging data which enables developers to reuse functions and data of existing applications, eliminating redundancy or duplication of efforts in new projects. The other objective is to support a business model (see Figure 1) developed to generate revenue to sustain activities of the SLL in marginalised areas. The business model was developed to guide the development of applications that can be used to generate revenue to support the development of ICT infrastructure in marginalised areas. The business model, called TW business model, works as follows:

- An entity (e.g. a municipality) acquires a licence for TW;
- Reed House systems, a company developing TW, installs it at a designated place in a marginalised area as directed by the municipality;
- The municipality opens up TW to other organisations to host their applications or use those already running to provide services to the communities in the area; and
- The organisations are billed based on the use of the services in TW by the target users.

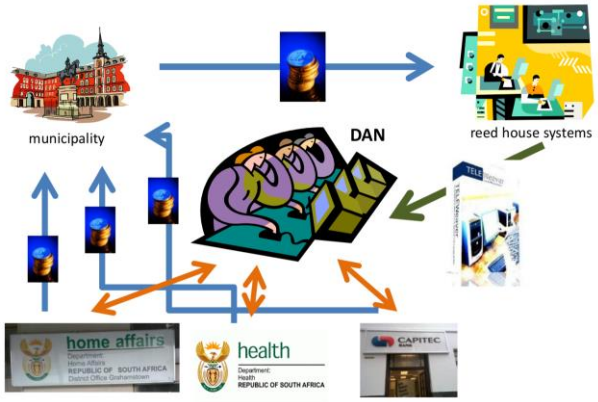



**Figure 1:** TeleWeaver Business Model. Source [20]

The SLL is currently upgrading TW to replace the current version which is based on Equinox, an Open Services Gateway Initiative (OSGi) technology with a new version, which is based on JBOSS, an enterprise web server designed to host medium to large applications[1]. The applications in the new version of TW will communicate through an Enterprise Service Bus (ESB). The ESB will handle routing of all the communications (requests and responses) between applications. The services provided by the applications will be accessible both online through the Internet and offline when the Internet is down through the local TW hosted on the school network.

### C. *Consumer Health Informatics*

Various authors have defined Consumer Health Informatics (CHI) differently. Common in the definitions is the development of computer-based interactive health communication applications to deliver health information and decision support to consumers [10, 15]. This paper adopts the definition by Eysenbach [6, p.1] who defines CHI as "the branch of medical informatics that analyses consumers' needs for information; studies and implements methods of making information accessible to consumers; and models and integrates consumers' preferences into medical information systems".

[1] http://www.jbossweb.jboss.org

As discussed by McDaniel *et al*. [15], CHI applications improve health outcomes by increasing personally relevant knowledge, enhancing self-efficacy for self-management, and enabling informed decision-making. Informed decision-making requires consumers to have access to information about the advantages and disadvantages of all possible courses of action [7]. This entails providing both patient-related information (diagnosis, pathology, personal risk factors, etc.) and general information (for example, effectiveness of different interventions for a certain disease). In addition to this information, consumers require significant lay knowledge and skills in areas ranging from health terminology to effective use of electronic media [11].

### *D. Health Communication and Strategies*

As defined by the National Cancer Institute and the Centres for Disease Control and Prevention [21], health communication is the study and use of communication strategies to inform and influence the target audience's decisions in order to enhance their health. Effective health communication can (1) increase the intended audience's knowledge and awareness of a health issue, problem, or solution; (2) influence perceptions, beliefs, and attitudes that may change social norms; and (3) increase the demand or support for health services [21]. For health communication to be successful, research-based strategies that select messages relevant to a group, or an individual, or that help determine appropriate channels to deliver health information should be used [9].

According to Hawkins *et al*. [9], health communication strategies are grouped into:

**Mass communication**: In this strategy, a relatively large, heterogeneous audience receives identical messages. Some of the common mass communication channels are newspapers, radio, television, and brochures. The advantage of this strategy is that it is less expensive compared to other strategies, because same materials are used to reach a large audience.

**Targeted communication**: In this strategy, the target audience is divided into smaller, more homogeneous subgroups in a process known as audience segmentation [12]. Some of the characteristics that are used to segment a population are demographics, behavioural, psychosocial, geographic, and risk-factors. As discussed by Kreuter *et al*. [13], in this strategy, intervention materials are different for each subgroup of the population. The intervention materials are developed based on characteristics that are common to the members of the subgroup. The assumption is that sufficient homogeneity exists among the members of the subgroup to justify using the same materials for all the members.

**Tailored communication**: In this strategy, health messages are designed to address the needs of an individual based on his/her unique characteristics [3]. To tailor the messages, information about an individual is gathered either through asking questions or from other sources like a database. The information is used to retrieve the most appropriate messages that meet the individual's unique health needs. Tailored messages are perceived to be more personally relevant, and to have the effect of stimulating cognitive activity as opposed to non-tailored or targeted health education messages [15]. According to Hawkins *et al*. [9], there are three message tailoring strategies which are:

- **Personalisation**: This strategy is used to increase consumers' attention, interest and motivation by implicitly or explicitly conveying that the message is meant for the recipient. An example is mentioning the name of the person in a communication.
- **Feedback**: In this strategy, individuals are presented with information about themselves in order to create an impression that they are understood.
- **Content Matching**: In this strategy, information provided by the user is used to retrieve matching messages/content to address a particular behaviour.

### *E. Internet-Based Health Intervention Studies*

Internet interventions are systematic treatment/prevention programs or activities, delivered through the Internet, to address one or more determinants of health of the target audience [1]. The growing popularity of the Internet among people of all ages and socio-economic statuses has led to its frequent use as a platform for the delivery of health interventions. Pull or Push tactics can be used to disseminate the information on the Internet [7]. In the push tactic, information is broadcasted to consumers, while in the pull tactic, consumers access it on demand. In each of the tactics, health communication strategies are used to ensure that users access only the information that is relevant to them.

The Internet has facilitated consumers' autonomy by allowing them to publish content (egalitarianism), to have multiple channels to provide input and feedback about content (adaptability), and to access health information from anywhere, and at any time (immediacy) [15]. However, egalitarianism brings about challenges of information quality and accuracy. If consumers access and use inaccurate health information, they stand a risk of diagnosing themselves wrongly, or taking medication not suitable for their condition. The other challenges of online health information include consumers being confused with the large amount of information leading to misinterpretation, or failing to access information that is relevant to their condition [7].

Some of the studies that have used the Internet to deliver health interventions are the interventions targeting rural men, who have sex with men (MSM) [2], and the intervention to prevent unplanned pregnancies, spread of sexually transmitted diseases and Human Immunodeficiency Virus (HIV) [18]. These studies concluded that the Internet is more superior as a platform for delivering health interventions than other platforms. Portnoy *et al*. [17] reviewed computer delivered interventions for health promotion, and behavioural risk reduction between the years 1988 and 2007 among which were Internet interventions. Similar to the other two studies, a conclusion was made that computer-delivered interventions are more efficacious than non-computer-based interventions.

## III. DESIGN AND IMPLEMENTATION

This section presents a discussion of the steps taken to develop the application. This includes a discussion of the

requirements gathering process, its architecture, functions, design, and Implementation.

### A. Requirements Gathering

A three stage process was used to gather requirements. The first step was a literature study in order to understand the process of health communication. The second step was to investigate existing CHI applications/websites to learn the functions they provide. Based on the literature study and the investigation of various CHI applications, a prototype application was designed and developed. The third and last step was a demonstration of the prototype to Raphael Centre, a non-governmental organisation based in Grahamstown that is involved in the dissemination of HIV/Aids messages. The aim was to get feedback on the functions, the usability of the application, and to understand the process of health communication from the practical point of view of a health service organisation. Because of other logistical challenges, the authors did not consult other organisations like the DoH.

### B. System Architecture

The application is web-based and follows a client/server architecture with the web browser as the client. The application consists of two components: a Dashboard and an end-user message dissemination component (HealthMessenger). Each of the components has its own database. They are exposed as web services through a RESTful interface to enable them to exchange data. According to the deployment plan, the HealthMessenger will be hosted on TW and will be used to disseminate and gather health information from consumers. The Dashboard will be hosted independent of TW and will be used by health service organisations to upload health messages and information to the HealthMessenger. The application was split into the two components to allow health service organisations to be able to use the Dashboard even when the TW is offline due to intermittent Internet services in the rural areas.

When deployed on TW, the HealthMessenger will interact with other applications such as the Profile application to obtain profile information to implement message targeting strategies, and the Billing application to post information to generate bills for health service organisations using the application to disseminate health messages (see Figure 2).

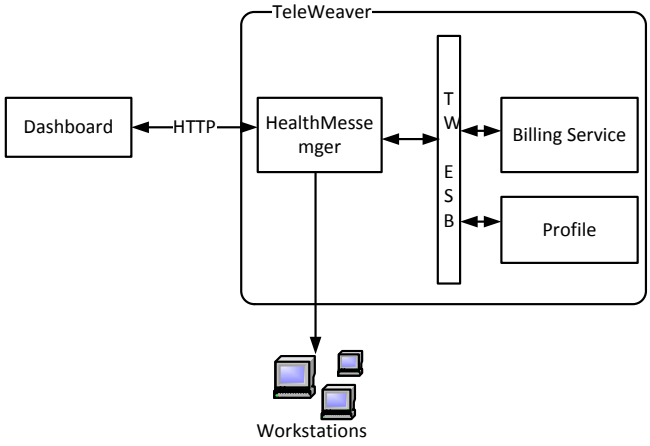



**Figure 2**: Overview of System Architecture

### C. Dashboard Functions

The functions of the Dashboard were designed to enable health organisations to create and manage campaigns, surveys, health topics, news, and events. Other functions are for production of reports on the use of the application, and for making application setting. The combination of the functions enables the dissemination and gathering of health information. The survey function is for data/information gathering, while the other functions such as campaigns, topic, and events are for uploading and disseminating health messages and information. The Dashboard functions provide for creating, editing/deleting, stopping/starting and publishing an activity (e.g. a campaign).

Campaigns, surveys, news, and events can be targeted using demographic information to indicate the target audience. A publish function is provided through a publish button on the Dashboard interface to retrieve health information from the Dashboard database and post in the HealthMessenger database. The data transmitted between the two applications is serialized as JavaScript Object Notation (JSON) format. JSON is preferred to other formats because it is lightweight, easy for humans to read and write, and easy for machines to parse and generate[2] making data transfer efficient. Communication between the two applications is achieved through Hypertext Transfer Protocol (HTTP).

### D. Message Dissemination

A blend of three health communication strategies discussed in Section II D is used to disseminate health messages and information. Where the information is targeted, targeted communication strategy is used. Where no targeting is done, mass communication is used. In both mass and targeted communication, personal information is used to personalise the health information dissemination in order to create an impression that the messages are designed specifically for the target recipient.

Only authenticated users with a completed user profile can access targeted health information. To determine appropriate information for a user, the user profile information is compared with the target information. When the target information is a subset or a complete match of the user profile information, the user can access the related health messages and information. Pull and push tactics (discussed in Section II E) are used to disseminate the messages. In the pull tactic, a user accesses health information either through searching or by clicking a menu item on the web interface. Popup message boxes are used in the push strategy in order to attract the attention of the users to act on the messages.

### E. Client-Side Implementation

The web browser is the client side of the applications. The function layout and presentation of the Dashboard follows a similar design in order to accelerate the learning process. Commonly used web icons, provided by a JavaScript library called "Font Awesome"[3], are used to direct users on the actions they can perform. A JavaScript library called

[2] http://www.json.org
[3] http://fortawesome.github.io/Font-Awesome/icons/

DataTable[4] is used to display a list of items (e.g. a list of surveys) to simplify sorting, searching, and other functions. To indicate the state of an activity (e.g. a campaign), different colours are used. For example, a campaign can be in "Active", "Running", "Finished", "Archived", or "Stopped" state (see Figure 3 for a screenshot) and each of these states is differentiated by a unique colour highlighting.

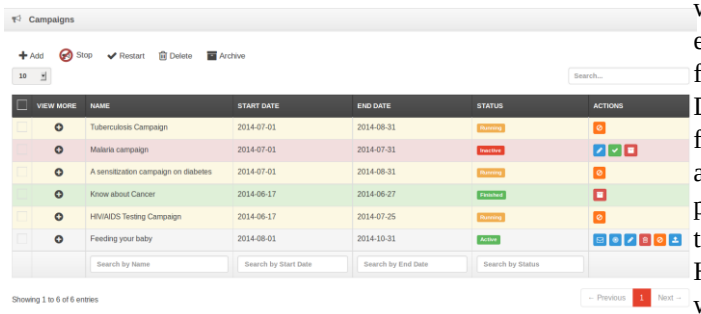



**Figure 3:** Screenshot of the dashboard interface

The HealthMessenger application is designed to support multiple languages. Two languages (English and isiXhosa) are currently supported. IsiXhosa was chosen because it is the dominant language among the people living in Dwesa. Language preference is set either when creating or updating a profile, or by selecting it on the web interface. A jQuery plugin called jquery-lang-js[5] was used to implement the multi-language support. To use the plugin, two language packs, one for English and the other for isiXhosa were created. Based on the set language, an appropriate language pack is loaded and the menus and labels are presented in the selected language.

### *F. Server-Side Implementation*

Java programming language was used to develop the server-side of the application. Spring Framework[6], an open source Java platform that provides support for developing Java applications, was used to accelerate the development work. Hibernate, an Object Relational Mapping technology was used to manage database access. Jersey, a JAX-RS API was used to develop a RESTful interface which enabled the component applications to exchange data. Jackson was used to process JSON data format.

## IV. Testing

Deployment of the HealthMessenger was not done on TW according to the deployment plan because TW is currently being upgraded and as such it is not ready for use. The current version of TW is based on OSGi which is not compatible with the programming style which was used to develop the application. To conduct a test of the application, two desktop computers were prepared by installing JBOSS AS 7.1 and extending it with the required libraries. One of the desktops acted as TW and hosted the HealthMessenger, while the other hosted the Dashboard. JBOSS was selected because no or minimal changes will be required for the HealthMessenger application to be deployed on the new version of TW. Unit, integration, and system tests were done to verify that the application is working as expected. Unit tests were done on each of the applications' components to confirm that the functions are working as expected.

A black box software testing method was used during the system testing phase to verify that the individual applications were working as per specifications and that they were able to exchange data. A test schedule of sample data was prepared for creating activities (e.g. surveys, campaigns) using the Dashboard. Different targets, and commencement and finishing dates to indicate the validity period were set on the activities. After creating the activities, the information was posted in the HealthMessenger using the publish button. A test schedule of user accounts was also created for the HealthMessenger. Ten accounts with different or sometimes with some profile information being similar were created. Each of the user account was used to access the application. It is a requirement that a user should complete a profile at first logon before accessing any of the health information as an authenticated user. A user accesses only the health information that is within the validity period and that he or she is allowed to access based on the target of the information. The target information each of the authenticated users was able to access was compared with his or her profile information to verify that the targeting algorithm is working. The tests confirmed the application to be working as expected. Further tests will be done when the new version of TW is completed which will include how the users interact with the application.

## V. Discussion and Future Work

The current design is suitable for environments where the Internet is stable to guarantee that both applications will be online at all times. If one of the applications is offline, the function to publish an activity fails. The solution is to use queuing technology to queue requests until the other component application comes online. Figure 4 show a proposed design with a broker managing the queuing of messages.

Hosting of the application in TW has many advantages. Some of the advantages are that the sources of health information can be controlled to prevent consumers from accessing information that is of questionable quality. The health messages and information are disseminated based on the needs and in a language the majority of the target audience understands.

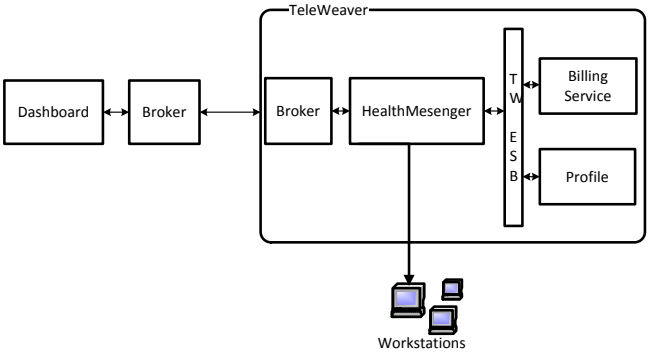



**Figure 4:** Proposed system architecture

[4] https://datatables.net/
[5] https://github.com/coolbloke1324/jquery-lang-js
[6] http://docs.spring.io/spring/docs/current/spring-framework-reference/html/overview.html

## VI. Conclusion

The Internet provides an alternative channel to disseminate health information for informed decision-making to people with Internet access living in marginalised areas. However, there are challenges of information overload and users not understanding the information if it is in a language they cannot easily comprehend. In this paper we discussed the design and development of a health informatics application to facilitate gathering and dissemination of health information by health service organisations to people living in marginalised rural area of Dwesa. The development of the application considered the importance of language localisation and the use of health communication strategies. Language localisation enables users to access the application in a language of their preference, while communication strategies are used to select appropriate information for the target users based on either the group or individual level characteristics. However, further considerations in the development of the application including the unreliability of the Internet in marginalised areas need to be considered. Future work should consider making the application usable at all times even when one of the applications is offline through the use of message queuing.

## VII. References


[1] G. G. Bennett and R. E. Glasgow. The delivery of public health interventions via the internet: Actualizing their potential. *Annual Review of Public Health*, 30:273–292, 2009.

[2] A. M. Bowen, K. Horvath, and M. L. Williams. A randomized control trial of internet-delivered hiv prevention targeting rural MSM. *Health Education Research*, 22:120–127, 2007.

[3] S. Davis. Internet-based tailored health communications: History and theoretical foundations. http://bcis.pacificu.edu/interface/?p=3363. Online [Last accessed: 01/04/2014], June 2007.

[4] I. de Lanerolle. The new wave. Technical report, University of Witwatersrand, 2012.

[5] H. Dugmore. The Siyakhula Living Lab: An important step forward for South Africa & Africa. Technical report, Rhodes University, November 2012.

[6] G. Eysenbach. Recent advances: Consumer health informatics. *British Medical Journal*, 320:1713–1716, 2000.

[7] G. Eysenbach and T. L. Diepgen. The role of e-health and consumer health informatics for evidence-based patient choice in the 21st century. *Clinics in Dermatology*, 19:11–17, 2001.

[8] S. Gumbo, H. Thinyane, M. Thinyane, A. Terzoli, and S. Hansen. Hansen. Living lab methodology as an approach to innovation in ICT4D: The siyakhula living lab experience. *IST-Africa 2012 Conference Proceedings*, pages 1–9, May 2012.

[9] R. P. Hawkins, M. Kreuter, K. Resnicow, M. Fishbein, and A. Dijkstra. Understanding tailoring in communicating about health. *Health Education Research*, 23(3):454–466, March 2008.

[10] T. K. Houston and H. E. Ehrenberger. The potential of consumer health informatics. *Seminars in Oncology Nursing*, 17(1):41–47, February 2001.

[11] A. Keselman, R. Logan, C. A. Smith, G. Leroy, and Q. Zeng-Treitler. Developing informatics tools and strategies for consumer-centered health communication. *Journal of the American Medical Informatics Association*, 15:473–483, 2008.

[12] M. W. Kreuter and S. M. McClure. The role of culture in health communication. *Annual Review of Public Health*, 25:439–455, 2004.

[13] M. W. Kreuter, V. J. Stretcher, and B. Glassman. One size does not fit all: The case for tailoring print materials. *Annals of Behavioral Medicine*, 21(4):276–285, 1999.

[14] D. Lewis, B. L. Chang, and C. P. Friedman. Consumer health informatics. In D. Lewis, G. Eysenbach, R. Kukafka, P. Z. Stavri, and H. Jimison, editors, *Consumer Health Informatics: Informing Consumers and Improving Health Care*, chapter 1, pages 1–7. Springer Science+Business Media, Inc., 2005.

[15] A. M. McDaniel, D. L. Schutte, and L. O. Keller. Consumer health informatics: From genomics to population health. *Nursing Outlook*, 56(5):216–223, 2008.

[16] M. McMullan. Patients using the internet to obtain health information: How this affects the patient–health professional relationship. *Patient Education and Counselling*, 63:24–28, 2006.

[17] D. B. Portnoy, L. A. J. Scott-Sheldon, B. T. Johnson, and M. P. Carey. Computer-delivered interventions for health promotion and behavioral risk reduction: A meta-analysis of 75 randomized controlled trials, 1988 – 2007. *Prevention Medicine*, 47:3–16, 2008.

[18] A. J. Roberto, R. S. Zimmerman, K. E. Carlyle, and E. L. Abner. A computer-based approach to preventing pregnancy, std, and hiv in rural adolescents. *Journal of Health Communication: International Perspectives*, 12:53–76, 2007.

[19] Siyakhula Living Lab. Project overview. http://siyakhulall.org/?q=node/42. Online [Last accessed: 31/03/2014].

[20] A. Terzoli. A model for sustainable broadband in poor communities. Presentation to South Africa's National Treasury, Pretoria, June 2012.

[21] U.S. Department of Health and Human Services. Pink book - making health communication programs work. http://http://www.cancer.gov. Online [Last accessed: 21/02/2013].



**Chikumbutso Gremu** received his honours degree in Computer Science in 2012 from Rhodes University and is presently studying towards his Master of Science degree at the same institution. His research interests include e-Learning, Open Source Software, Consumer Health Informatics and Web Technologies



**Acknowledgements**. This work was undertaken in the Distributed Multimedia CoE at Rhodes University, with financial support from Telkom SA, Tellabs, Genband, Easttel, Bright Ideas 39, THRIP and NRF SA (TP13070820716). The authors acknowledge that opinions, findings and conclusions or recommendations expressed here are those of the author(s) and that none of the above mentioned sponsors accept liability whatsoever in this regard.

# Contract-based Web Service Evolution Model

Kudzai. Chiponga, Paul. Tarwireyi, and Matthew O. Adigun
Department of Computer Science
University of ZULULAND, Private Bag X1001, KwaDlangezwa, Empangeni 3886
Tel: +27 35 9026393, Fax: +27 35 9026569
email: {kukuchipo, ptarwireyi, profmatthewo}@gmail.com

**Abstract-Due to the dynamic and changing nature of the Service Oriented Architecture, service based systems are always subjected to changes. These changes emanate from changing consumer requirements, new business rules, infrastructural changes and/or changes due to the competitive drive in service oriented business environments. In SOA, once a service is deployed, consumers search for the service, invoke and bind to its interface and in so doing, create a relation between a consumer and a service provider. This binding is not a physical integration; thus, when service providers decide to change or evolve a service, they may not be aware of the set of consumers that relies on the service. As services are changed and adapted over time, the service consumers using the services always need to be considered, because some changes will have unpredictable impacts in the service consumers. Contracts are used to specify the relation between the service provider and the consumers in SOA. Contracts are the means by which a service provider agrees to deliver a certain service and consumers know what to expect from a service they subscribe to. This paper presents a contracts-based model that can be used to deal with service changes in a controlled manner to allow existing consumers to continue using a service whilst catering for the needs of customers who have new requirements. Using a scenario, we demonstrate the applicability and usability of this model in SOA environments.**

**Index Terms—service evolution, contracts, SOA, Versioning, Compatibility.**

## I. INTRODUCTION

Recent trends in Information and Communications Technology (ICT) suggest that eventually most software capabilities will be delivered and consumed in the form of services [1][2]. Organizations need infrastructure that must quickly evolve and adapt to changing business needs and IT changes[3]. The adoption of the Service Oriented Architecture (SOA) enables organizations to have the ability to quickly respond to the ever changing and competitive business environment. SOA is characterized by loosely coupled, coarse-grained and platform independent services that support business processes [4]. Typically, Simple Object Access Protocol (SOAP) based services in SOA are designed and deployed by a service provider and the provider advertises the services in a registry known as Universal Description Discovery and Integration (UDDI) [5]. A service consumer searches for the service in the registry and binds to the interface of that service.

By publishing the service and its interface, the service provider is committing to support the deployed service. A consumer, in accepting the terms of the service provider as stated in the published web service documents, commits to interact with the service in the manner predefined by the exposed documents from the service provider. The process described above can be considered equivalent to the signing of a contract between the provider and the consumer [6], [7]. Often services are created and deployed targeting a wide range of possible consumers in a particular domain. Due to the loose coupling of web services in SOA , service providers may not be fully aware of the number of consumers using the service [8] such that when the provider decides to perform some maintenance activities, some consumers may break [9]. When a provider introduces a new version of a web service, the general versioning principle is that consumers should not be forced to use the new version immediately. Evolving a service therefore should allow existing service consumers to remain unaffected by service changes as much as possible [8], while allowing customers with new requirements to enjoy the new features of the service. This entails maintaining compatibility between subsequent service versions. One of the challenges in the area is on reducing the complexity of the maintenance tasks. Hence, solutions are still needed.

## II. A MOTIVATING EXAMPLE

We consider a Hotel reservation SOA application, which consists of two services: a reservation service and a banking service. The hotel service handles the reservations of rooms for guests while the banking services offers payment processing for the hotel reservations. An end-user accesses the hotel reservation service through a web browser. The end-user can select the country, in which they need accommodation, specify the range of prices they can afford and the dates they need to be accommodated. The hotel service returns a list of hotels whose offerings match the criterion specified. When the end-user selects a hotel and the room they wish to book, they can continue to a make a reservation. The reservation requires for the end-user to enter the currency they wish to pay in and the banking service is invoked to convert the prices using the current exchange rates.

Due to various factors such as regulatory compliance and competition, entities like the hotels and banks will evolve their service offerings to try and stay a step ahead of their competition, for instance. If due care is not taken, these

changes in the services may break the consumers of the services. This paper explores a way in which the possible service disruptions in consumers can be minimized.

## III. Categories Of Changes

### A. *Changes in service functional behavior*

This refers to changes made to the web service implementation and business process logic. These changes can be as a result of changes in infrastructure and business process requirements [10]. These changes can be implemented without breaking the consumers.

### B. *Changes in non-functional behavior*

Non-functional behavior refers to the changes in the policies which govern the quality attributes of the service. Policies can be used to supplement the web service definitions found in the Web Services Description Language (WSDL) files. This provides the service provider with the room to offer consumers more than just a single policy. Different consumers, like in the case of corporate organizations, have difference internal policies that govern their operations and these may influence changes in the web services offerings by the service providers.

### C. *Changes in service interface*

Service interface changes focus on the changes that happen to web service interfaces in response to changes in service operations or the message structures used to exchange information with the service. For instance, removal of an operation will imply that the service interface no longer supports that operation [10]. For example, in our HotelResv service, the service consumer receives a price for the rooms available from the hotels. In the first version of the HotelResv service, only the rooms and their prices were returned to the requestor (see Figure 2). Now if the requestor needs additional information such as the description of the rooms, then the service provider would have to update the service to accommodate this new functionality.

## IV. Web Service Contracts

For any pair (provider | consumer) to interoperate successfully, there must be a common understanding and agreement of what the provider offers and what the consumer can use. The formal arrangement of the contents of a service, the price, the expected protocols for integration and quality aspects of a service are presented in the form of a contract [11]. A contract is therefore defined as the service schema elements that are expected by the consumer and offered by the provider [11].

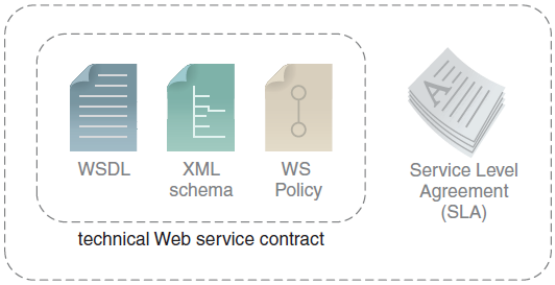



Figure 1. Web Service Contract [12]

Figure 1 shows the contents of a typical web service contract which is a collection of metadata describing the implemented service. The Web Services Description Language (WSDL) is an eXtensible Markup Language (XML) document that may include or refer to an XML Schema. The WSDL document is considered the defacto standard for writing web service contracts [12]. The XML Schema expresses the shared language for defining the structure of documents and helps machines to carry out rules made by people. Web Service (WS)-policy specifies the behavioral expectations of a service and can be used to extend the contract provided by the WSDL and Schema [12]. Service Level Agreements (SLA) though outside the scope of this paper, help establish the conditions and verifiable qualities of a service that a service provider should meet [13][14].

When a service consumer accepts the contract and can implement all parts of the contract to achieve all possible interactions with it, then we say the consumer is compatible with the service.

## V. Compatibility

The notion of contracts gives a mechanism to define compatibility [11]. It is necessary to maintain the same contract running for as long as possible. It is also necessary to ensure that when services are developed and deployed, they work properly [15]. Hence two contracts are compatible if they work together with no need for alterations to achieve interoperability. Proper evolution is not supposed to be destructive to current consumers hence must remain backwards compatible.

### A. *Backwards Compatibility*

Already existing service consumers and service providers develop trust [16] when service consumers and providers become coupled. Developers therefore strive to maintain backwards compatibility to avoid negative consumer impacts. That is to say, the new version of the contract has to continue to support consumers designed to operate with the old version of the contract.

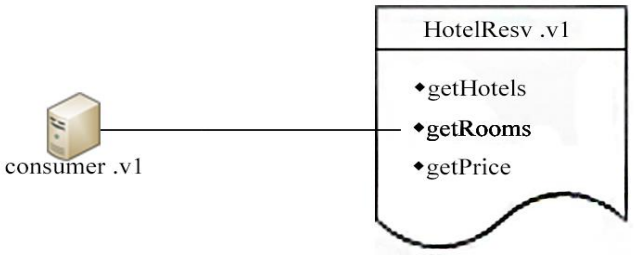



Figure 2. Consumer version 1 compatible with service version 1 Adapted from [12]

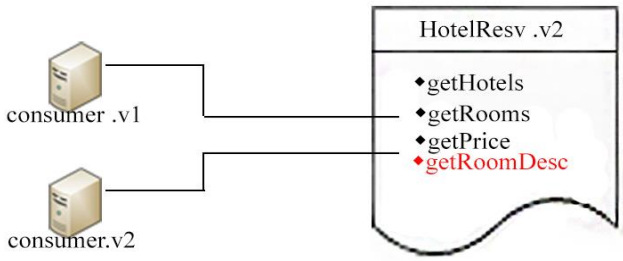



Figure 3. Consumer versions 1 & 2 compatible with service version 2 Adapted from [12]

Figure 2 represents a provider (hotel reservation service version 1) and consumer (consumer version 1) binding, in

which the consumer uses the service in a compatible manner and can display the available rooms and prices of a selected hotel. Initially services deployed and consumers using them are compatible when a consumer can successfully use all or parts of the contract provided by the provider. Figure 3, if a new consumer (consumer version 2) requires a different capability for example, one that displays the full descriptions of each of the rooms available, that would mean the contract would have to be altered. However, replacing the initial *getRooms* element with *getRoomDesc* element in the contract would break the contract with consumer.v1. Therefore changes need to be applied in a way that maintains the backwards interoperability with consumer.v1 such as adding an element (*getRoomDesc*) to the existing contract in service version 2. This can be added as an optional element by the service provider.

### B. *Forwards Compatibility*

Forwards compatibility is difficult to incorporate in service design as there is no easy way to predict future changes that will need to be made [17]. So contracts can be implemented with some degree of forwards compatibility by adding optional elements. This allows for the existing provider contract to seamlessly interoperate with a new version of a consumer with no modifications of the provider contract in any way. This means that the contract is designed to support unknown future consumers' needs. In the HotelResv service for example, the contract.v1, in Figure 4, is designed with an element *AdditionalInfo*, which is not implemented / used by consumer.v1. When this element is used (Figure 5.) in HotelResv.v2, it does not break the contract that is servicing consumer.v1 and it allows for the support of the additional information requirement for consumer.v2.

```
    ….
    <xsd:complexType name="HotelResvRequestType">
          <xsd:sequence>
             <xsd:element name="getRooms"
type="xsd:string"/>
          </xsd:sequence>
          <xsd:anyAttribute/>
       </xsd:complexType>
       <xsd:complexType name="HotelResvResponseType">
          <xsd:sequence>
             <xsd:element name="roomType"
type="xsd:String"/>
             <xsd:element name="price" type="xsd:float"/>
             <xsd:element name="AdditionalInfo"
type="customType:AdditionalInfoType1" minOccurs="0"/>
          </xsd:sequence>
          <xsd:anyAttribute/>
    </xsd:complexType>
…..
```

Figure 4. Contract.v1 designed with forwards-compatibility in mind.

```
    <xsd:complexType name="HotelResvRequestType">
          <xsd:sequence>
             <xsd:element name="getRooms"
type="xsd:string"/>
             ….
       <xsd:complexType>
                <xsd:simpleContent>
                   <xsd:extension base="xsd:String">
                      <xsd:attribute name="getRoomDesc"
type="xsd:string" />
                   …..
             <xsd:element name="price" type="xsd:float"/>
             <xsd:element name="AdditionalInfo"
type="customType:AdditionalInfoType1" minOccurs="0"/>
          </xsd:sequence>
          <xsd:anyAttribute/>
       </xsd:complexType>
       <xsd:complexType name="AdditionalInfoType1">
          <xsd:sequence>
             <xsd:element name="more" minOccurs="0">
                <xsd:complexType>
                   <xsd:attribute name=" getRoomDesc"
type="xsd:string"/>
                </xsd:complexType>
             </xsd:element>
             <xsd:element name="AdditionalInfo"
type="customType:AdditionalInfoType2" minOccurs="0"/>
          </xsd:sequence>
          <xsd:anyAttribute/>
    </xsd:complexType>
```

Figure 5. Contract.v2 using additionalinfo element

## VI. Related Work

Change is inevitable for the continued existence and survival of implemented services, thus there are research efforts towards the evolution of services in SOA. Robinson discusses some challenges in service evolution from a consumer perspective, where consumers play a role in determining the evolution of services based on business expectations [18]. This evolution pattern assumes that the provider and the consumer know each other well enough to negotiate a contract before the implementation of business logic changes. Frank et al [19] introduced the concept of using a service interface proxy to manage the hosting of versioned web-services. In their work, Frank et al mention that during the evolution, the service interface will remain the same while the business logic will be changed. Multiple proxies are then generated per service where an incompatible service exists [19]. Benatallah et al proposed developing adapters for web services integration for resolving interface differences [20]. The chain of adapters is a design technique that was proposed as a possible solution by Kaminski et al [21]. In this technique developers would have to build interface adaptors mapping between subsequent versions of the web service and the chain grows as the web service versions increase.

Treiber et al classified the changes that affect web services and analyzed the causes and effects of the changes on web services [10]. They identified the changes in the implementation, the interface, Service Level Agreements and pre-and-post conditions of a service and some influences that cause the changes.

An in-depth exploration into service compatibility based on set theory was conducted by Andrikopoulos et al, in which they show the reasons and conditions under which services can evolve maintaining compatibility with the current consumers [22]. Andrikopoulos also introduces a contracts based method to eliminate spurious results of uncontrolled changes in the evolution of services. Andrade et al proposed that software development and evolution be centered on the notion of a contract. The aim of contracts was to support more flexible ways of system evolution [23].

Despite several research efforts being undertaken, the evolution of web services is still an open area of research [24][25][26][8]. This work seeks to contribute in these efforts by leveraging contracts, to enable the controlled evolution of web services. The model proposed in this work makes an effort to reduce the complexities associated with maintaining interoperability of subsequent service versions, while minimizing undesired effects on existing consumers.

## VII. Proposed Model For Service Evolution

The previous sections highlighted the view of continuous change to web service contracts. Eventually, this will give rise to evolution of the contracts in which there will be a distinction between changes that break the existing contracts or those that do not have an impact on the existing contracts. The successful evolution of services requires that existing service consumers do not get forced to update to the new service versions immediately after a new version is released. The new service version must remain compatible with the existing consumers and also compatible with the consumers consuming the service in its new state. The proposed model handles Web Service evolution by giving version numbers to contracts, so every time non backward compatible changes are effected, a new web service on a separate namespace is released.

### A. Service Provider

The service provider (Figure 6) is an organization or an individual that provides a service by exposing it to other organizations or individuals such that they can consume it over the internet. The service provider owns the web service and implements the proxy that matches the services they deploy. The provider then registers the service with registry and publishes the contract matching the service deployed.

### B. Registry

The registry allows businesses to find each other and facilitates communication by hosting a catalog of published and available web services. There are private and public registries, and in this work we focus on the use of a public registry, in which the eventual consumers of the web services published by the service provider are not known. In our representation, the registry is the catalog in which the service provider publishes the contracts of the available services.

### C. Consumers

A consumer is a member of the set of users of the web service. The consumer can be a human agent, a web application, software application or another web service. The consumer relies on the contract that they discover in the registry. They bind to the service upon accepting the contract.

### D. Proxy

The proxy is designed and implemented as an adapter service in the Enterprise Service Bus (ESB) performing transformations of SOAP messages from an old consumer to match the requirements of the service. The proxy relies on the published contracts found in the service registry, to verify the version of the service which is required by the consumer. This is done by comparing the version numbers in the published contract to the version that the consumer has indicated in their invocation of the service. The proxy can be implemented as an extension of the ESB or an independent service.

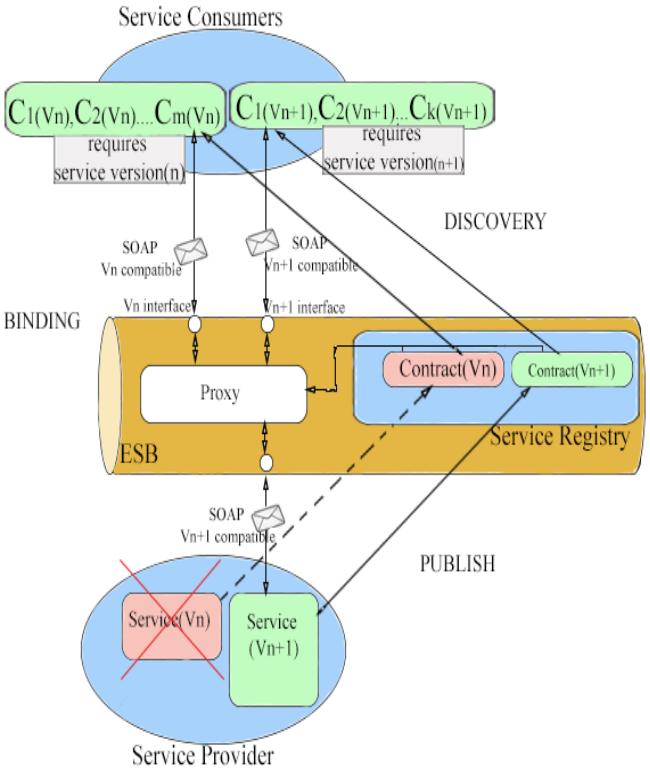



Figure 6. Model for web service evolution

### E. ESB

The ESB is middleware that serves to provide communication medium between consumers and providers. The ESB may provide the hosting of a service registry for the publishing of services by the providers and discovery of advertised services by the consumers. Routing of messages between the consumers and the providers is done in the ESB.

## VIII. How The Model Works

The provider builds and deploys a service Service($V_n$) and registers it in the UDDI. Registration makes available a corresponding contract, Contract($V_n$), for the deployed Service($V_n$). Organizations implement their consumers or consuming applications or services (Consumer($V_n$)), which we consider compatible to Service($V_n$). Contract($V_n$) is binding for as long as the Consumer(Vn) uses Service($V_n$).

Now, let K be the set of all consumers requiring the service version n $K=\{ C_{i(V(n))} \mid i > 0\}$. Let RCi be the set of all requests from consumers using contract version n. RCi = $\{rcv_{(n)} \mid$ SOAP requests from any $C_{i(V(n))}\}$. Let M be the set of all methods (F), exposed by the running web service. $M = \{F_k \mid k > 0\}$, $|F|>0$. We assume the service only exists and is useable iff $|F|>0$, then Fo and Fn are methods exposed by the old and new web service respectively. $R_n$ and $R_{n+1}$ are sets of the service requirements found in the old and new service contracts respectively. Hence $\{R_n \mid$ requirements for service version n$\}$, $\{R_{n+1} \mid$ requirements for service version n+1) and $R_n \subset R_{n+1}$. If a compatible change is made to the Contract($V_n$) then $rcv_{(n)} \subseteq Fo$, and *RCi fulfills* $R_n$. If a breaking change is to be introduced, *RCi does not fulfill* $R_n$ so Contract(Vn+1) is put in place. If *RCi does not fulfill* $R_n$ *and* $rcv_{(n+1)} \subseteq R_{n+1}$ then a proxy is put in place to transform all requests from *K* to match $R_{n+1}$. So we now have $rcv_{(n)}+(proxy) = rcv_{(n+1)}$ and $rcv_{(n+1)} \subseteq Fn$. *RCi then fulfills* $R_{n+1}$.

Figure 7 shows the server-side algorithm for the proxy operations. This algorithm is focused on comparing incoming SOAP request formats with Contract($V_{n+1}$) requirements. If the SOAP request format is not a match, then the proxy transforms the request to the expected format matching Contract($V_{n+1}$) requirements. The transformed request is then passed on to the actual service for processing. The service provider does not wish to keep more than one service running in parallel as this will tie up resources that could otherwise be in use for other purposes. Running more than one service in parallel also brings [23] in unnecessary technical overheads. Thus the proxy service in the ESB is configured to mimic the service interfaces indicated in Contract($V_n$), and applies a transformation to the services requests of Consumer($V_n$) to meet the new service requirements of Contract($V_{n+1}$). Consumer($V_{n+1}$) requests do not need to be transformed so the proxy forwards them to the service.

Marking the Contract($V_n$) as being deprecated can then be done and consumers still using the old contract can be notified. Once all consumers on the old contract have moved to the new version, no changes need to be made by the service provider except to mark the old contract as decommissioned or to deregister that old contract from the registry.

1. **If** not-exists then cache the contracts of the services
2. **If** a request from consumer requires Service($V_n$) create an instance of the proxy:
    a. **Get** SOAP request and compare format with Contract($V_{n+1}$) requirements
    b. **Apply** XSLTransformtion to match Service($V_{n+1}$) format
    c. **Forward** request to Service($V_{n+1}$) and Listen for response
    d. **Get** the SOAP response and apply Transformation to match Contract($V_n$) response
    e. **Send** response to consumer
3. **Else** send request to Service(Vn+1)
    a. **Listen** for response from Service($V_{n+1}$)
    b. **Send** response to requesting consumer

Figure 7. Algorithm for the proxy operations

In our model we assume that a new version of the contract is released whenever a breaking change is introduced in the service contract. As a result, a new service is introduced, and a proxy is built to handle the incoming requests from consumers. We also assume that there is consumer identification and notification mechanism is in place through which we can know and contact the consumer.

## IX. Model Application

For example, in our hotel scenario, at some point the hotel reservation service becomes available and organizations may build web based consumers to the service (Figure 2). A new consumer may request for additional information to be made available about the rooms, and a new operation is added (Figure 3). This is seen as a compatible change made to the hotel reservation service's contract, nothing needs to be done to the existing consumers as the original operations are still available and being serviced [22].

When a change conceived as a breaking change is to be implemented such as renaming *getPrice* to *getPriceList*, to give the consumer a price-list of all available rooms for comparison purposes, we create a new version of the hotel reservation service and publish the new contract. We build and deploy a proxy that mimics the interfaces of both the old version of the service and new version as in figure 6.

```
<xsl:stylesheet version="1.0"
……."
   xmlns:hrs="http://localhos/hotels/hotelResv">
   <xsl:template match="/">
      <xsl:apply-templates
select="soap:Envelope/soap:Body/*"/>
   </xsl:template>
   <xsl:template match="@*|node()">
      <xsl:copy>
         <xsl:apply-templates select="@*|node()"/>
      </xsl:copy>
   </xsl:template>
</xsl:stylesheet>
```

Figure 8. Example XSLT for transforming message format

Figure 8 is an example of the XSLT that is used by the proxy to transform [27] SOAP-Request messages. The stylesheet in Figure 8 transforms the SOAP message to give the contents of the SOAP body without the body element. In a similar fashion another stylesheet is applied to reverse the process for the SOAP-response message in the proxy before being sent back to the requesting consumer.

Once the proxy is in place, there is no need to maintain the old services thus the resources that were being used by that version are freed up to be used elsewhere. We then mark the old service contract as deprecated and notify consumers that relied on the old version. Consumers may choose to remain using the old contract without experiencing any service disruptions up to the time that the service provider ceases to support the old consumer completely. The main idea behind the proposed model is to keep a single latest version of the service instead of having multiple versions. It is envisaged that this will reduce the number of versions of the same service that have to be maintained, hence making web service maintenance tasks less complex.

## X. Conclusion and Future Work

This paper presented a contract-based model for the evolution of a web services. The model aims to minimize the complexity of maintaining multiple versions of a service. It advocates the use of a single proxy to handle all requests and responses to and from consumers. The old version of a service is discarded once the proxy is in place and all requests from consumers are transformed so that they can be serviced by the new service. A hotel scenario was used to illustrate the applicability and usability of the proposed model. Even though the proxy caters for the old consumers, ultimately these consumers will need to update to the new service. The future work lies in sending change notifications to these consumers and setting business rules for updating, so that we do not end up with a chain of proxies to administer.

## XI. References

[1] D. Sprott and L. Wilkes, "Understanding Service-Oriented Architecture," *Understanding Service-Oriented Architecture*, Jan-2004. [Online]. Available: http://msdn.microsoft.com/en-us/library/aa480021.aspx. [Accessed: 20-May-2014].

[2] B. Orriëns, J. Yang, and M. P. Papazoglou, "A framework for business rule driven web service composition," in *Conceptual Modeling for Novel Application Domains*, vol. 2814, Springer, 2003, pp. 52–64.

[3] S. Govardhan and J. Feuerlicht, "SOA: Trends and Directions," in *Proceedings of the 17th International Conference on Systems Integration 2009*, Prague, Czech Republic, 2009, pp. 149–154.

[4] M. P. Papazoglou and W.-J. van den Heuvel, "Service-Oriented Design and Development Methodology," *Int J Web Eng Technol*, vol. 2, no. 4, pp. 412–442, Jul. 2006.

[5] J. Bean, "Chapter 1 - SOA—A Common Sense Definition," in *SOA and Web Services Interface Design*, J. Bean, Ed. Boston: Morgan Kaufmann, 2010, pp. 1–24.

[6] J. Bean, *SOA and Web Services Interface Design: Principles, Techniques, and Standards*. Morgan Kaufmann, 2009.

[7] G. Dai, X. Bai, Y. Wang, and F. Dai, "Contract-Based Testing for Web Services," in *Computer Software and Applications Conference, 2007. COMPSAC 2007. 31st Annual International*, 2007, vol. 1, pp. 517–526.

[8] M. Fokaefs and E. Stroulia, "WSDarwin: Studying the Evolution of Web Service Systems," in *Advanced Web Services*, A. Bouguettaya, Q. Z. Sheng, and F. Daniel, Eds. Springer New York, 2014, pp. 199–223.

[9] M. Fokaefs, R. Mikhaiel, N. Tsantalis, E. Stroulia, and A. Lau, "An Empirical Study on Web Service Evolution," in *2011 IEEE International Conference on Web Services (ICWS)*, 2011, pp. 49–56.

[10] M. Treiber, H.-L. Truong, and S. Dustdar, "On analyzing evolutionary changes of web services," in *Service-Oriented Computing–ICSOC 2008 Workshops*, 2009, pp. 284–297.

[11] M. P. Papazoglou, "The Challenges of Service Evolution," in *Advanced Information Systems Engineering*, Z. Bellahsène and M. Léonard, Eds. Springer Berlin Heidelberg, 2008, pp. 1–15.

[12] T. Erl, A. Karmarkar, P. Walmsley, H. Haas, L. U. Yalcinalp, K. Liu, D. Orchard, A. Tost, J. Pasley, and & 6 more, *Web Service Contract Design and Versioning for SOA*, 1 edition. Upper Saddle River, NJ: Prentice Hall, 2008.

[13] P. Bianco, G. A. Lewis, and P. Merson, "Service level agreements in service-oriented architecture environments," 2008. [Online]. Available: http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA528751. [Accessed: 14-Apr-2014].

[14] C. Ruz and F. Baude, "Enabling SLA monitoring for component-based SOA applications," in *36th Euromicro Conf. on Software Engineering and Advanced Applications*, Lille, France, 2010.

[15] L. Bordeaux, G. Salaün, D. Berardi, and M. Mecella, "When are Two Web Services Compatible?," in *Technologies for E-Services*, M.-C. Shan, U. Dayal, and M. Hsu, Eds. Springer Berlin Heidelberg, 2005, pp. 15–28.

[16] Z. Malik and B. Medjahed, "Trust Assessment for Web Services under Uncertainty," in *Service-Oriented Computing*, P. P. Maglio, M. Weske, J. Yang, and M. Fantinato, Eds. Springer Berlin Heidelberg, 2010, pp. 471–485.

[17] S. Karus, "Forward Compatible Design of Web Services Presentation Layer," Masters Thesis, Faculty of Mathematics & Computer Science, University of Tartu, Estonia, 2007. http://www. cyber. ee/dokumendid/Karus. pdf, 2007.

[18] I. Robinson, "Consumer-Driven Contracts: A Service Evolution Pattern," *Consumer-Driven Contracts: A Service Evolution Pattern*, 2006. [Online]. Available: http://martinfowler.com/articles/consumerDrivenContracts.html. [Accessed: 17-Apr-2014].

[19] D. Frank, L. Lam, L. Fong, R. Fang, and M. Khangaonkar, "Using an Interface Proxy to Host Versioned Web Services," in *IEEE International Conference on Services Computing, 2008. SCC '08*, 2008, vol. 2, pp. 325–332.

[20] B. Benatallah, F. Casati, D. Grigori, H. R. M. Nezhad, and F. Toumani, "Developing Adapters for Web Services Integration," in *Advanced Information Systems Engineering*, O. Pastor and J. F. e Cunha, Eds. Springer Berlin Heidelberg, 2005, pp. 415–429.

[21] P. Kaminski, M. Litoiu, and H. Müller, "A Design Technique for Evolving Web Services," in *Proceedings of the 2006 Conference of the Center for Advanced Studies on Collaborative Research*, Riverton, NJ, USA, 2006.

[22] V. Andrikopoulos, "A theory and model for the evolution of software services," Tilburg University, Open Access publications from Tilburg University 262, 2010.

[23] A. Andrade and J. Luiz, "Evolution by Contract," in *Proceeding of the ACM Conference on Object-Oriented Programming, Systems, Languages, and Applications*, Minneapolis, Minnesota USA, 2000.

[24] G. Lewis, D. Smith, and K. Kontogiannis, "A Research Agenda for Service-Oriented Architecture (SOA): Maintenance and Evolution of Service-Oriented Systems," *Softw. Eng. Inst.*, Mar. 2010.

[25] S. Kijas and A. Zalewski, "Towards Evolution Methodology for Service-Oriented Systems," in *New Results in Dependability and Computer Systems*, W. Zamojski, J. Mazurkiewicz, J. Sugier, T. Walkowiak, and J. Kacprzyk, Eds. Springer International Publishing, 2013, pp. 255–273.

[26] G. A. Lewis and D. B. Smith, "Research Challenges in the Maintenance and Evolution of Service-Oriented Systems," *Migrating Leg. Appl. Chall. Serv. Oriented Archit. Cloud Comput. Environ.*, pp. 13–39, 2013.

[27] H. Suzuki, N. Ishikawa, H. Ueno, and T. Gotoh, "Mobile Content Transformation using XSLT and its Evaluation.," in *12th International WWW Conference Industrial Track*, Budapest, Hungary, 2003.

**Kudzai Chiponga** is currently studying towards his Master of Science degree at the same institution. His research interests include Service Evolution and Versioning.

# Using a Mobile Solution to Support Chronic Disease Management in South Africa

Cainos Mukandatsama, Janet Wesson
Department of Computing Sciences
Nelson Mandela Metropolitan University, P. O. Box 77000, Port Elizabeth
email: {Cainos.Mukandatsama, Janet.Wesson }@nmmu.ac.za

**Abstract- This paper discusses the design and evaluation of a mobile solution to support patients suffering from chronic conditions. The mobile solution aims to improve medication compliance of patients suffering from chronic diseases. Such patients take medication on a daily basis and also need support on the management of their chronic conditions. Chronic diseases are different from other diseases in that chronic diseases cannot be cured completely, but can be controlled by taking medication on a regular or daily basis. It is important for patients with a chronic disease to comply with their medication schedule, to prevent complications or a negative impact on their health. A user study was previously done to investigate the usability of the proposed solution using a convenience sample of staff and students. A field study was recently conducted to obtain feedback from patients with chronic diseases and determine if the solution met their expectations. This paper outlines the design of the proposed mobile solution, the design of the field study and the results of the field study.**

**Key words— chronic disease management, medication compliance, self-care, self-regulating system, m-health, m-health initiatives**

## I. INTRODUCTION

Healthcare management is an essential part of people's welfare, especially for patients with diseases that linger over a long period of time. When a patient gets ill, they look for ways to make them feel better and this usually happens by getting some medication. Chronic diseases are especially difficult to deal with since they persist over a long period of time. Patients with chronic diseases need to have medical attention to manage their conditions. According to the World Health Organization (WHO), chronic diseases are conditions or diseases that require on-going management over a period of several years [6]. The problem of chronic disease management has been aggravated by the global population increase with the main consequence being the increase in the number of people with chronic diseases, disability and a high level of medication dependency [7]. Chronic diseases are responsible for many deaths and are also a major contributor to the rising costs in healthcare [8].

Many elderly people suffer from chronic diseases. The National Centre for Health Statistics determined in 2012 that 81.6% of elderly people have at least one chronic disease [9]. In order to minimize the effects of these diseases, patients need to take pills on a daily basis [10]. Long-term conditions require daily attention. Chronic disease management aims to reduce the disease burden of a chronic condition [11]. To improve chronic illness care and medication compliance, patients should be empowered and engaged in healthcare self-management [12]. This paper will discuss how the problem of chronic disease management can be reduced by making use of mobile technology.

To control chronic diseases, the patient needs to take the correct pills every day in the right dosage and at the scheduled time. The effects of not adhering to the schedule have proven to have a negative impact on the patient's health and should be avoided [1]. Patients with chronic diseases therefore require some form of support mechanism in order to help them to adhere to the medication schedules that they are meant to follow. Typically, such patients, especially the elderly, would have caregivers who will be responsible for helping them in their daily activities including taking of pills, but the number of caregivers is significantly lower than the number of patients [3].

It is necessary to produce a solution that can be embedded within technology that a lot of people are currently using. This would reduce the complications of the users having to learn how to use new devices. Many people are currently using mobile phones. According to recent studies, more than five billion people are using mobile phones of which approximately 170 million of these are smart phones [4]. Mobile technology has been piloted in a range of health-related areas, and has been used to improve the dissemination of public health information diagnosis and treatment, the distribution of health information to doctors and nurses, patient management, public health monitoring and the increased efficiency of administrative systems. Research suggests that mobile phones can play a significant role in health management [5].

It is therefore relevant to investigate the viability of using a mobile application to help users with chronic disease management. Some work has been done in this field, but the existing work suffers from several limitations. Also some of these applications can only run on phones like the Apple iPhone or the Windows Mobile phone, which are expensive in the South African context.

The remainder of this paper is structured as follows: Section II outlines the related work derived from literature study and Section III discusses the proposed mobile solution. Section IV discusses the design of the field study to evaluate the proposed mobile solution and Section V outlines the results. Section VI consists of the conclusion of the paper and future work.

## II. RELATED WORK

### A. Chronic Disease Management

Many authors have proposed various definitions of chronic disease management. The Disease Management Association of America proposed a comprehensive definition and outlined the key aspects of the definition as: supporting the relationship and plan of care between the care provider and the patient, emphasizing on avoiding the disease or condition getting worse by using evidence-based guidelines and practices as well strategies that empower the patient [13]. This definition also includes the aspect of on-going assessment of the results with the aim of improving the overall health of the patient with a chronic condition. The goals of chronic disease management and prevention are: reducing the severity of the effects of the chronic diseases, improving the quality of an individual's health and duration of the individual's life, prevention of the occurrence of the chronic disease and delaying the start of disease and disability [14].

### B. Medication compliance

Medication compliance can be described as the degree or extent to which a patient conforms to the recommendations of their day-to-day treatment by the health provider with respect to the time, dosage and frequency [15]. Compliance with conditions that require medications for a short period is generally considered to be higher than that of long period medication regimens [16]. Medication compliance for chronic diseases can be regarded as low since they fall in the long-term category.

### C. Self-Care

According to World Health Organization, self-care can be defined as the ability for an individual, caregiver and the community at large to support health, to prevent and control diseases and sustain health to manage sickness and disabilities with or without aid by healthcare providers [17]. According to Dorthea Orem, self-care is behavior that an individual learns for a particular purpose and follows a sequence of patterns and actions [18]. Self-care reduces the dependency of the patient on the health professionals and empowers the patient. Increase in self-care and reduction of patient-health professional dependencies are important in chronic disease management to increase medication compliance by patients as they manage their health status on a daily basis. The proposed mobile solution attempts to promote medication compliance by encouraging self-care.

### D. Barriers to self-care

1. **Health care providers** - In cases where the health care providers are not able to provide sufficient information, the patient's self-care ability will be affected and thus health care providers will be regarded as barriers to self-care [19].
2. **Social and cultural barriers** - The involvement of a patient's family can either support or hinder patient self-care behaviour. It is helpful to take into account the role of spirituality, participation in community religion, and cultural practices such as traditional healing to understand how family and social context can affect self-care directives [20].
3. **Cognitive barriers** - A patient's knowledge about a specific chronic condition can influence his or her ability to perform optimal self-care. It has been reported that people with low literacy levels have more difficulty learning self-care skills [21].
4. **Physical barriers** - Sometimes the adoption and success of self-care can be influenced by the physical state of the patient. Chronic conditions (especially for the elderly) often result in physical disability due to reduced and lessened strength, vision or sensation [22]. Physical incapacity can be a barrier to the patients' ability to take care of themselves and manage their illness.
5. **Access to resources and facilities** – Lack of access to health care resources and services for the community and primary care, can be a significant barrier to optimal self-care [19]. If this access is low, self-care will not be optimal since the patients and their caregivers will not be able to consult professionals or get some advice, knowledge and medical equipment easily.

### E. Mobile health

In recent years, a number of researchers have increasingly used mobile phones as platforms for delivering health interventions [23]. Mobile phones have grown to be an important platform for delivering health interventions. This is mainly because of the following:

1. Mobile phones have been widely adopted in the world [24];
2. People tend to carry their mobile phones everywhere. A study by Patel *et al.* [23] found that individuals were within arm's reach of their phones on average 58% of the time;
3. People have a tendency of being attached to their mobile phones [25]; and
4. Mobile phones can exhibit context-awareness by connecting to other data sources [26].

Mobile health (m-health) is an umbrella term that covers areas of networking, mobile computing, medical sensors and other communication technologies within healthcare. The concept of m-health refers to "*mobile computing, medical sensors and communications technologies for health care*" [28]. Studies done by Chang Liu *et al.* [29] conclude that medical compliance applications on the Apple App store are less than 5% of the available m-health apps. There is therefore a need to improve medication compliance for patients, especially those suffering from chronic diseases, by optimizing the capabilities of mobile phones.

### F. Existing systems

This section looks into several applications that serve to remind people of their medication schedule. These applications have some similarities in the basic functionality and had their results published in literature.

A Smartphone-based Medication Self-management System (SMSS) was designed by Hayakawa et al [30]. SMSS provided SMS reminders when a patient forgets to take medication by using a wireless pillbox for detection. The system also kept history of how the patient was taking their

medication. Setting up and operating such a system is expensive. Receiving SMS'es would require the patient to manually delete them so that the inbox will not get full. The system does not provide visual cues to assist the patient to identify the pills

SIMPILL is an interactive pill holder/container that uses mobile phone technology to assist patients to remember to take their pills at the prescribed time [31]. The SIMPILL container consists of an ordinary pills' container with a SIM card and transmitter attached to it. SIMPILL reminds patients by sending an SMS to the patient should they not take their medication. There is no assistance in identification of the pills and it does not support placing an order at the pharmacy before the pills get used up.

## III. MY PILL REMINDER

A mobile application, called My Pill Reminder, was designed and developed for patients with chronic diseases to remind them to take their pills on time. My Pill Reminder is the proposed tool for supporting chronic disease management by improving medication compliance and self-care in the South African context. This mobile application is a self-regulatory medication compliance mechanism because it strives towards making the patient develop a pattern of taking their medication at the prescribed times. This application was designed in a way that it does not require expensive external devices and the user does not incur any costs in receiving the reminders. My Pill Reminder's main objective is to remind the patient to take the right pills at the right time in the right amount.

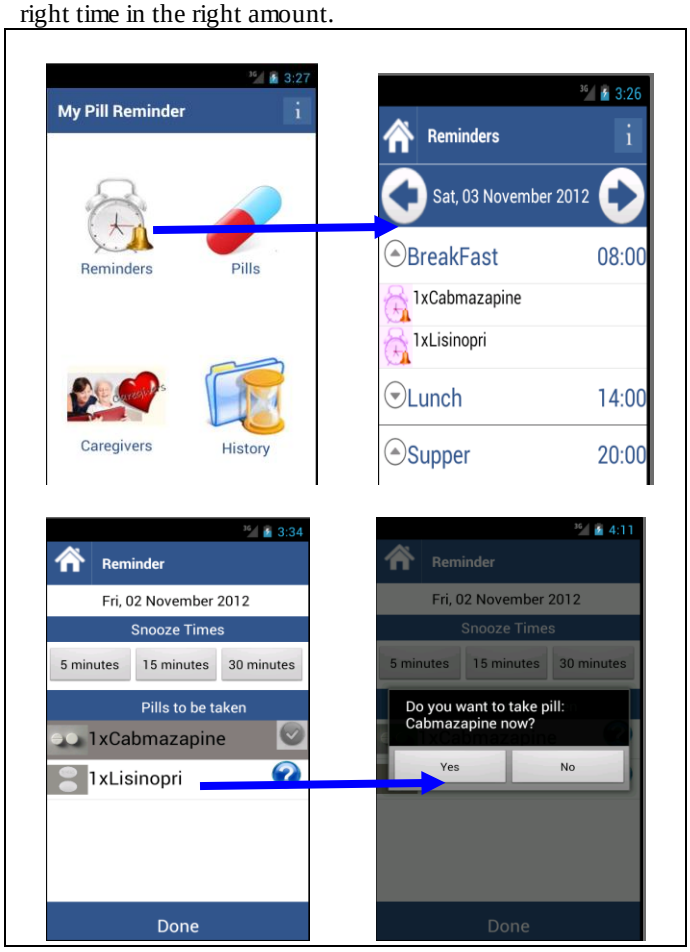



Figure 1: My Pill Reminder home screen and main screens

The design of the application is shown in Figure 1. Figure 1 shows the home screen and how the patient can navigate to view the reminders. The figure also shows the steps involved in taking pills. The My Pill Reminder application assists patients, especially elderly people, in identifying the pills by providing visual cues whilst the patient is taking their pills. My Pill Reminder reminds the user using text, pictures and audio whenever the user needs to take the pills and also ensures that they take the right pills. The application has a functionality of tracking the quantities of the pills which are in the patient's prescription and automatically reordering pills from the pharmacy. The application also tries to prevent emergencies as a result of the user not complying with the medication schedule by notifying family members and caregivers.

### A. *User Study*

A user study was conducted on the Mobile Pill Reminder application in a controlled environment to investigate how usable the application was. The major objective was to measure the performance of the application, that is, its efficiency in terms of time, and effectiveness in terms of task completion. The study used a convenience sample of staff and students at NMMU to adhere with ethical constraints. Figure 2 shows the overall usability scores obtained from the user study. The averages of the ratings for all the participants were all more than 85%, which indicates that the application was regarded as highly usable. The overall usability score was calculated to be 88.94%, which shows that the participants managed to use the application with ease and were also highly satisfied with it.

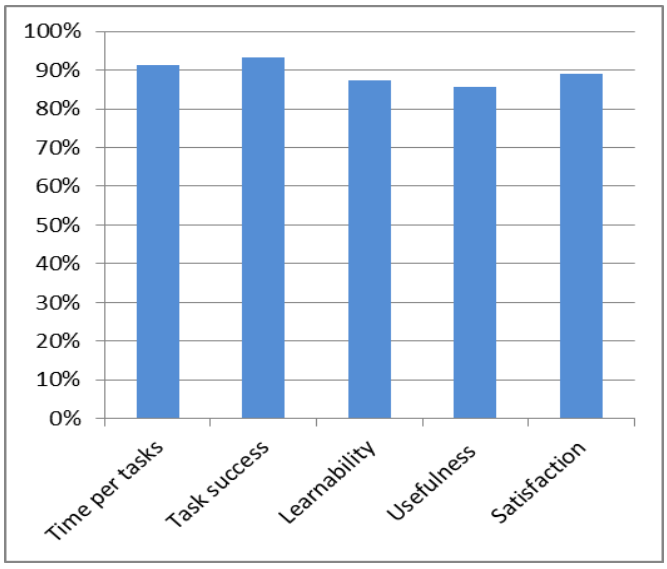



Figure 2: Graph illustrating the overall scores of the different usability metrics (n=20)

## IV. FIELD STUDY

### A. *Aim of the evaluation*

The mobile solution focuses on being easy to learn and use. The main aim of the field study was to investigate if patients with chronic diseases felt that the mobile solution was easy to learn and use. Other objectives of the field study can be summarized as follows:

- To investigate if the users feel the application is useful;
- To identify usability issues of the application; and
- To identify possible changes according to the users.

### B. *Instruments*

The same version of the application was used to enable the mobile application to be evaluated consistently. For users who own Android phones, the study was carried on their phones by installing the Mobile Pill Reminder application. However in cases where the participant did not own Android phone, the participants were issued a Samsung Galaxy S4 mobile phone with the Mobile Pill Reminder application installed. The mobile phones also consisted of some basic tutorials which explained the basic Android patterns and the corresponding interactions relevant for using the application. In order to be able to track the usage and how the participants were taking their medication, logging software was incorporated in the application, and the data was exported at the end of each field study.

The Post-Study System User Questionnaire (PSSUQ) was used for obtaining feedback from the participants after using the My Pill Reminder application. PSSUQ was chosen as an instrument because it provides a comprehensive insight into the overall satisfaction of the users after using the mobile application. The feedback from the participants was used to determine the extent to which they felt the application was useful.

### C. *Participant selection*

Choosing the participants was an integral part of this field study. Obtaining suitable participants willing to take part in the study was very difficult. This was mainly because most of the people had the perception that the study would also evaluate their mobile phone competency. Since the My Pill Reminder application was being evaluated, it was preferred that the participants had some experience with using a mobile phone. The main prerequisite used for selecting a participant was that the participant had to have at least one chronic condition that required taking medication every day. Some of the characteristics considered whilst choosing the participants were as follows:

- **Age** - 18 years old or older;
- **Health status** - suffering from a chronic condition;
- **Attitude toward technology**- uses mobile phones regularly;
- **Location** – resides in Port Elizabeth; and
- Familiarity with the medical terminology and exposure to medication procurement.

The final sample size of the study was six participants. A larger sample size was considered preferable, but several potential participants cancelled due to either a lack of interest or resistance to technology. The field study was conducted over a period of three days.

### D. *Demographics*

The demographics of the participants used in the field study were influenced by the fact that the participants had to have a chronic condition. The demographical data captured included: age, gender and the level of mobile phone expertise of the participant. Five of the participants were male and one was female. There were four age groups which comprised the 26-40 years group, which had one participant, the 41-65 years group, which also had one participant, and the 66 or above age group, which had four participants. In terms of mobile phone expertise, one participant was a beginner and the other five participants were intermediate users. All of the participants had at least one chronic condition.

## V. RESULTS AND ANALYSIS

After using the application in the study, the participants provided feedback by completing a questionnaire. The participants were also interviewed to get an in-depth view of how they felt whilst using the application. The data from both the completed questionnaires and interviews were categorized and analyzed qualitatively. Due to the small sample size, the results were not sufficient or representive enough to be analyzed quantitatively. Owing to this, the questionnaires' results were only analyzed qualitatively and categorized.

### A. *Results*

From the feedback, several themes were derived and the frequency of the comments was also noted. The themes with the highest frequency were also noted and further critically analyzed. The comments that were common from the majority of the participants relating to the problems they encountered in using the application were regarded as usability issues. The results are tabulated in Tables 1 and 2. Comments which fell within the same theme were assigned the same color and white was used to indicate that a theme was not assigned to a comment.

Table 1: Summary of positive aspects (n=6)

| Participant comment | Frequency | |
|---|---|---|
| It is useful and helps me | 5 | |
| Very easy to use | 5 | |
| It was very easy to learn | 6 | |
| History facility enabled viewing trends in taking pills | 2 | |
| Use of different colors highlighted important information | 2 | |

Table 2: Summary of negative aspects (n=6)

| Participant comment | Frequency | |
|---|---|---|
| Reminder needs to be more intrusive and must be louder and longer | 4 | |
| Light must be associated with the reminder | 2 | |
| Flexibility needed in taking the pill (before time) | 1 | |
| Reminding you to take the pills before you leave home | 1 | |
| Going direct into the Reminder screen | 3 | |
| Eliminating the use of the Notification area | 3 | |
| Touch screen phones were complicated | 2 | |

| | | |
|---|---|---|
| Reversal of actions in taking pills | 2 | |

*B. Discussion*

From Table 1 and 2, there were several themes and corresponding comments that were identified from the participants' feedback. The themes are as follows:

1. **The application was easy to learn and use**–The participants managed to learn and use the application by themselves after being given a short tutorial on key Android features and interactions. This theme is indicated by green color in Table 1.

2. **The application was useful** – The participants also highlighted that they felt that mobile application could be useful in managing their chronic conditions. Two of the participants liked the functionality of being able to view the history of how they have been taking their pills. This theme is indicated by the black color in Table 1.

3. **The reminder must be more noticeable** - most of the negative comments were in the "reminder theme". The reminder was designed to be consistent with the user's settings on the phone, which could be soft if the user has set the settings accordingly. Participants' comments suggested that the reminder should be louder, the light must flash and it must repeat several times to increase the chances of the participant noticing the reminder. The comments also highlighted that the participants would like more flexibility and options in taking their medication. Participants would like the inclusion of a facility to take their pills before the time set on the application, in the event that they want to take their medication earlier. One of the participants also wanted to be reminded to take medication with him to his workplace, so that he would not forget the medication at home. This theme is indicated by the red color in Table 2.

4. **Eliminating the use of the notification area** - Another usability identified was the use of the Notification area to access the reminder. This comment was as a result of those participants who were not familiar with using Android phones or with Android patterns. Participants would prefer the reminder to appear as a pop-up dialog which is inconsistent with Android patterns. This theme is indicated by the purple color in Table 2.

There were also other comments that were made by participants but were not allocated to themes. Some of these comments came from one or two participants and were mainly dependent on the demographics of the participant. For example, two of the participants felt that the touch-based mobile phones were complicated and difficult to use. These comments could be regarded as influenced by the level of mobile experience, particularly, with touch-based smart phones.

This research had several limitations. Firstly, the field study had a sample size of six participants. Using a larger sample size would have increased the confidence level. Secondly, giving the participants a second phone affected the study because they were not using the phones they use every day. Being given a smart phone also introduced a novelty effect and excitement, which possibly affected the feedback they gave at the end of the study. Also, the duration of the field study was short and having the participants use the application for a longer period could have yielded more accurate and conclusive results.

## VI. CONCLUSIONS

The paper discussed how the My Pill Reminder application can support chronic disease management. The My Pill Reminder application was successful in reminding the participants to take their pills and encouraging self-care and medication compliance as shown by the results of the earlier user study. The results of the field study highlighted that there is a need to reach a compromise between being user-centric and following the standard design principles and patterns.

Future work can involve addressing the concerns of the participants. The quality of the results can also be increased by having a longer field study with a bigger sample size. There is potential of extending the research by linking the mobile solution to intelligent devices, which monitor the state of health of the patient. A similar study could also be conducted with other groups such as HIV patients or people who are taking birth control pills, which need to be taken at fixed times.

## REFERENCES


[1] A. Lorenz and R. Oppermann, "Mobile health monitoring for the elderly: Designing for diversity," *Pervasive and Mobile Computing*, vol. 5, no. 5, pp. 478–495, Oct. 2009.

[2] K. Elgazzar, M. Aboelfotoh, P. Martin, and H. S. Hassanein, "Ubiquitous Health Monitoring Using Mobile Web Services," *Procedia Computer Science*, vol. 10, pp. 332–339, Jan. 2012.

[3] Jing, G. and Koronios, A. 2010 Mobile Application Development for senior citizens [Online]. Available: http://www.pacis-net.org/file/2010/S05-03.pdf [Accessed: 03 April 2014].

[4] Butler, K., McDaniel, P. and Ongtang, M. 2010. Porscha: Policy Oriented Content Handling in Android. [In ACSAC '10: Proceedings of the 26th Annual Computer Security Applications Conference] ACM.

[5] Kinkade S, Verclas K. Wireless technology for social change. Washington, DC, and Berkshire, UK; 2008 .Available from http://mobileactive.org/files/MobilizingSocialChange_full.pdf, [Accessed 18April 2014]

[6] World Health Organisation (2012). [Health statistics and health information systems –elderly people]Available online: http://www.who.int/healthinfo/survey/ageingdefnolder/en/index.html [Accessed: 02 April 2014]

[7] García-Sánchez, P., González, J., Mora, A. M., & Prieto, A. (2013). Deploying intelligent e-health services in a mobile gateway. Expert Systems with Applications, 40(4), 1231–1239. doi:10.1016/j.eswa.2012.08.068

[8] Sunyaev, A., & Chornyi, D. (2012). Supporting chronic disease care quality. Journal of Data and Information Quality, 3(2), 1–21. doi:10.1145/2184442.2184443

[9] National Centre for Health Statistics. 2012. Definition of Chronic Diseases [Online]. Available: http://www.medterms.com/script/main/art.asp?articlekey=33490 [Accessed: 03 April 2014]

[10] Australian Department of Health. 2010. Chronic Disease.[Online]. Available: http://www.health.gov.au/internet/main/publishing.nsf/content/chronic [Accessed: 03 April 2014]

[11] Rijken M, Jones M, Heijmans M. Supporting self-management. In: Nolte E, McKee M, eds. Caring for People With Chronic Conditions: A Health System Perspective. Berkshire: Open University Press; 2008:16-142.

[12] Chen, G., Yan, B., Shin, M., Kotz, D., & Berkel, E. (2012). MPCS : Mobile-Phone Based Patient Compliance System for Chronic Illness Care

[13] McKenna M and Collins J. Current Issues and Challenges in Chronic Disease Control. IN: Chronic Disease Epidemiology and Control, 3rd Edition. American Public Health Association, Washington DC, 2010.

[14] McKenna M and Collins J. Current Issues and Challenges in Chronic Disease Control. IN: Chronic Disease Epidemiology and Control, 3rd Edition. American Public Health Association, Washington DC, 2010.

[15] Cramer, J., Roy, A., Burrell, A., & Fairchild, C. (2008). Medication compliance and persistence: terminology and definitions. Value in Health, 11(1), 44–47. doi:10.1111/j.1524-4733.2007.00213.

[16] Morris, L. S., & Schulz, R. M. (1992). Patient compliance-an, 283–295.

[17] World Health Organisation (2012). [Health statistics and health information systems –elderly people]Available online: http://www.who.int/healthinfo/survey/ageingdefnolder/en/index.html [Accessed: 02 April 2013]

[18] Orem, DE. Nursing: Concepts of Practice. 6th ed. St. Louis: Mosby; 2001. World Health Organization. Report on chronic disease. http://www.who.int/topics/chronic_diseases/en/.

[19] Williams, L. (2012). overcome barriers to self-care, 37(3), 32–38.

[20] Truter I. African traditional healers: cultural and religious beliefs intertwined in a holistic way. SA Pharmaceutical Journal. 2007:56-60.

[21] Macabasco-O'Connell A, Crawford MH, Stotts N, Stewart A, Froelicher ES. Self-care behaviors in indigent patients with heart failure. J Cardiovasc Nurs. 2008;23(3):223-230.

[22] Pun SPY, Coates VE, Benzie IFF. Barriers to the self-care of type 2 diabetes from both patients' and providers' perspectives: literature review. J NursHealthc Illn. 2009;1(1):4-19.

[23] Klasnja, P., & Pratt, W. (2012). Healthcare in the pocket: mapping the space of mobile-phone health interventions. Journal of biomedical informatics, 45(1), 184–98. doi:10.1016/j.jbi.2011.08.017

[24] Pew Internet & American Life Project. Internet, broadband, and cell phone statistics; 2010 [Online]. Available: http://www.pewinternet.org/Reports/2010/Internetbroadband-and-cell-phone-statistics.aspx?r=1 [Accessed: 23 April 2013]

[25] Hong, J., Suh, E., Kim, S., 2009. Context-aware systems: a literature review and classification. Expert Systems with Applications 36 (4), 8509–8522

[26] Ventä L, Isomursu M, Ahtinen A, Ramiah S. My Phone is a part of my soul – how people bond with their mobile phones. In: Proc UbiComm '08. IEEE ComputerSociety; 2008. p. 311–7

[27] International Data Corporation Woldwide quarterly mobile phone tracker; 2012. http://www.idc.com/getdoc.jsp?containerId=IDC_P8397

[28] Istepanian, R., Laxminarayan, S., Pattichis, C.S., 2006. Introduction to mobile mhealth systems. In: Evangelia Micheli-Tzanakou (Ed.), M-health: Emerging Mobile Health Systems. Springer, Nantwich, 3–3

[29] Liu, C., Zhu, Q., Holroyd, K. a., & Seng, E. K. (2011). Status and trends of mobile-health applications for iOS devices: A developer's perspective. Journal of Systems and Software, 84(11), 2022–2033. doi:10.1016/j.jss.2011.06.049

[30] M. Hayakawa, Y. Uchimura, K. Omae, K. Waki, H. Fujita, and K. Ohe, "A smartphone-based medication self-management system with realtime medication monitoring.," Applied clinical informatics, vol. 4, no. 1, pp. 37–52, Jan. 2013.

[31] SIMpill. "The SIMpill Medication Adherence Solution.(2012), http://www.simpill.com/thesimplesolution.html [Accessed: 19 June 2014].

# SerPro: A Mashup tool for Enhanced Usability for Novice Users

Sabelo Yalezo and Mamello Thinyane Telkom
Centre of Excellence in ICT for Development,
Department of Computer Science
University of Fort Hare, P. O. Box 1314, Alice 5700
Tel: +27 40 6022464, Fax: +27 40 6022464
Email: {syalezo, mthinyane}@ufh.ac.za

**Abstract- Mashups are platforms that allow integration of various components (e.g. Application Programming Interfaces (APIs)) to create envisioned composite application. Their primary goal is to share and reuse the available APIs on the web. They also enable a seamless integration of public web APIs. Mashup tools like Yahoo! Pipes and Dapper have been regarded as some of the most usable Mashup tools available. This has led to an investigation on the metrics and the universal use of Mashup tools, in particular considering their utilization by novice users. Subsequent to the investigation we define usability of Mashup tools based on the comparison study that has been carried out. The findings from the study are then used to guide the architecture and development of a new Mashup tool targeted specifically for novice users. Furthermore, we recommend some of the techniques that can be used to enhance the usability of Mashup tools.**

**Index Terms—usability-Factors, EUP, Mashups-architecture, usability metrics**

## I. INTRODUCTION

Information Communication Technologies (ICTs) are increasingly being deployed in developing countries, thereby creating new ways of doing business and delivering e-services to the end-users. Similarly the increasing availability of online services has led further developments in service composition and consumption with one of the key enablers of this being service Mashups. Mashups are developed for both novice users and advanced-users. As a results their usability, which is coupled to the profile of the users, is one of the major challenges affecting their design and implementation. According to ISO 9241-11, usability is defined as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" [1]. Mashup tools are said to be usable when more users are able to use them (i.e. the more a system is used by end-users, the more likely it will be usable). Although designers produce primary usability objectives throughout the requirements gathering process, Mashup tools are still not usable [2]. In this research we define the usability based on the following hypothesis: designing and implementing a customizable and satisfying Mashup tool can improve the usability for novice users.

Hence, this research targets novice user as it's subject to investigate on how to define the usability of Mashups. With the notion that if the proposed Mashup tool can be usable to novice users it should also be usable to advanced users. Other system objectives are not ignored but rather follow the system usability. The research hypothesis was answered by the following objectives in order to attain a usable Mashups:

- To identify usability shortfalls of Mashup tools
- To define usability metrics for Mashup tool.
- To design a Mashup tool architecture based on usability guidelines obtained.
- To evaluate the effectiveness of the proposed tool.

By achieving the above objectives, the research outcomes will be a usable Mashup tool will be available to novice users. In addition a guideline on usability measuring variable was provided to help other Mashups developers to enhance or improve the already existing ones.

This paper is structured as follow: section II highlights some of the weaknesses, challenges and strengths currently faced by Mashup tools, section III defines the methodology used to produce the proposed Mashup tool; section IV, addresses the usability attributes deduced from the gathered requirements, section V presents the proposed Mashup tool (SerPro) and finally section VI and section VII focuses on results and a conclusion, respectively.

## II. STATE OF ART

In this section we further review some of the existing Mashup tools. The main purpose of this section is to identify the causes of high usability in Mashup tools and also some of the issues or challenges that impede the usability of Mashup tools. We describe these tools, first, based on design architecture and then the composition layer. The Mashup tools examined include: Microsoft Popfly, Intel mash maker, marmite, Yahoo! Pipes, MobiMash and dapper.

### *A. Design Architecture*

Design architecture defines what the system does and how the system is implemented. In this sub-section we focus on the design architecture of the currently available Mashup tools.

#### *1) Microsoft Popfly*

In Popfly, users build Mashups using basic programming constructs called *blocks*. Each block performs a set of *operations* such as data retrieval and data display. Each operation takes *input* parameters to allow customization. Blocks are connected to form a network in which the output of a block can be used as input for adjacent blocks [3].

*2) Yahoo Pipes*

Yahoo! Pipes is a web-based visual programming language for constructing data Mashups [4]. The Yahoo! Pipes editing environment consists of four main regions: a navigational bar across the top, the toolbox on the left, the work canvas in the center, and a debug-output panel at the bottom. The toolbox contains modules, the building blocks of the Yahoo! Pipes visual language. In the graphical language of Yahoo! Pipes, modules (operators) are laid out on a design canvas.

*3) Intel MashMaker*

Intel MashMaker is an interactive tool that tracks what the user is doing and tries to infer what information and visualizations they might find useful for their current task [5]. The Intel MashMaker client is currently implemented as an extension to the Firefox web browser. Mash Maker adds a toolbar to the browser that shows buttons representing enhancements that Mash Maker believes the user might want to apply to the current page. An enhancement might combine the data on the page with data from another source, or visualize data in a new way.

*4) MobiMash*

MobiMash is a platform for the construction of mobile Mashups, characterized by a lightweight composition paradigm, mainly guided by the notion of visual templates. The composition canvas consists of two main panels: the data panel on the left that displays the data retrieved by querying some selected data services; the visual template panel on the right that shows selected visual template, i.e., a representation of the User Interface (UI) of the final app. The data items selected in the data panel are associated to UI elements and that updates the visual template panel with a preview of the association items [6].

***5)** Marmite*

Marmite lets end-users create Mashups that repurpose and combine existing web content and services [7]. Marmite supports a data flow architecture, where data is processed by a series of *operators* in a manner similar to Unix pipes. More specifically, Marmite lets end-users: "Easily extract interesting content from one or more web pages" [7].

*6) Dapper*

Dapper stands for Data mapper. Dapper is a web application that visually runs in a wizard mode asking the user to fill-in some field at each step in order to create a "dapp" (data imported). The user interface is very minimalist, but it gets the things done [8].

In the next sub-section we discuss composition and service consumption based on the above Mashup tools.

*B. Composition Layer and Service Consumption*

Composition layer is where different components are integrated to produce the requested composite application. Service consumption details on how end-user interact with system in order to get their envisioned Mashups. This sub-section focuses on how the different APIs or services are composed and consumed by the end-users.

*1) Microsoft Popfly*

In Popfly, blocks are listed in different categories, which users can search. Additionally, users may share their Mashups with others for reuse and modifications. Shared Mashups can be retrieved using a textual search.

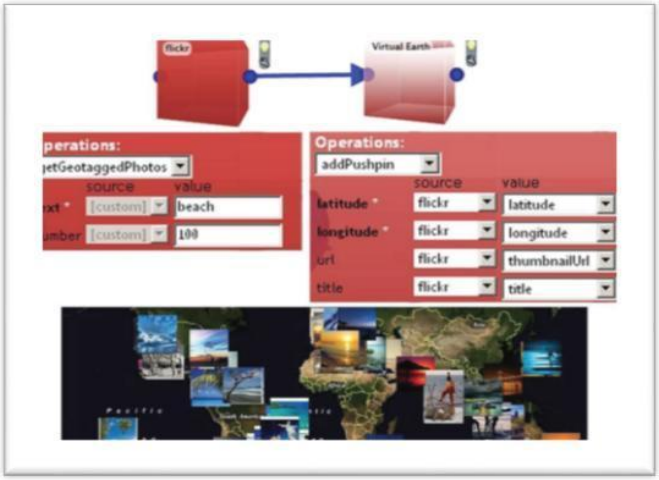



**Figure 1.MS Popfly** [3]

Figure 1.MS Popfly **[3]** shows a Mashups example in which the Flickr API sends a list of images about "beaches" with their geographical coordinates to the Virtual Earth block. The results are displayed on a map (i.e. Figure 1 at the bottom).

*2) Yahoo! Pipes*

The input and output ports are wired together, representing the flow of data through the application. Selecting an output port, highlights all the compatible input ports to which the user may connect it. There are a number of data types within Yahoo! Pipes, which determine what inputs and outputs are compatible. In the most general terms, there are simple scalar data *values*, and *items*, which are sets of data objects (e.g., items in an RSS feed) [4]. Values include types like text, urls, locations, numbers, dates, and times. The widgets and pipes are shown in Figure 2 below:

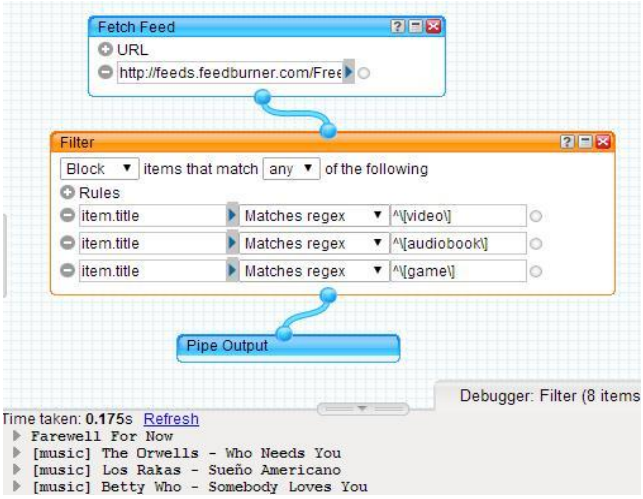



**Figure 2.Yahoo pipes** [4]

*3) Intel MashMaker*

Mash Maker is intended to be an integral part of the way the user browses information, rather than being a special tool that a user uses when they want to create Mashups [5].

In order to create Mashups from normal websites, Mash Maker must first extract structured data from them. Intel MashMaker uses structured data from existing web sites to create new "mashed up" interfaces combining information from many sources.

*4) MobiMash*

The composition process of MobiMash is characterized by an End-User Development (EUD) Web environment, where a visual composition paradigm, based on the completion of visual templates, allows the users to easily configure the fusion of contents coming from different data sources, and the synchronization of such core contents with both remote APIs and local services available on the mobile device. The so-created applications are devices' native applications that, in contrast with Web Mashups, do not need the Web browser as an execution environment - the access to mobile device services is therefore enhanced. The composition paradigm generates an application schema that is based on domain specific language addressing dimensions for data integration and service orchestration, and at runtime that guides the dynamic instantiation of the final mobile app [6].

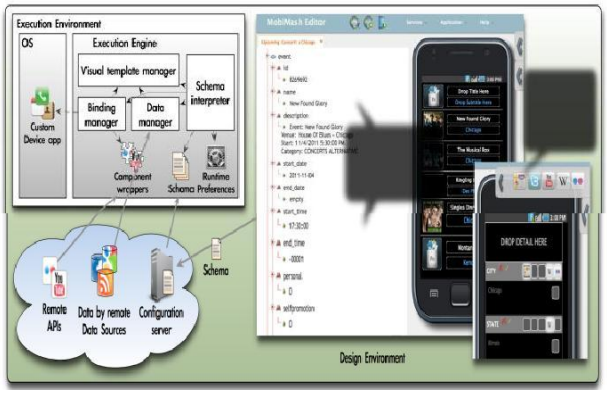



**Figure 3.MobiMash [6]**

*5) Marmite*

The composition process follows a data flow manner, for example filtering out values or adding metadata. It integrates with other data sources, both from local or remote databases and from other existing web pages or services [7]. In addition it directs the output to a variety of sinks, such as databases, map services, text files, web pages, or compliable source code that can be further customized. In other words, the linked view shows both the program and the data simultaneously. Nevertheless some users are able to use this system and to construct programs with web services quickly without difficulties.

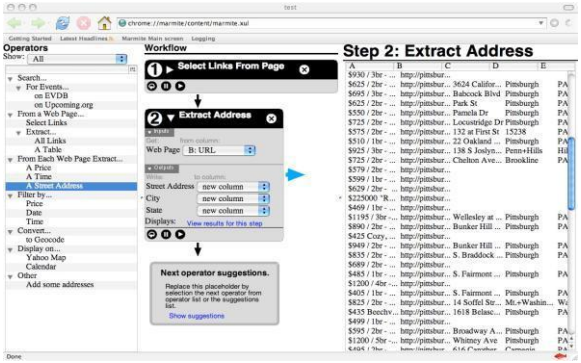



**Figure 4.Marmite** [7]

*6) Dapper*

The main purpose of this Mashup tool is to convert any type of content into a standard form that can be reused (RSS, XML)[8]. It also has a set of publishing features that turn that content into Google Gadget, Netvibes Module, Flash widgets and so on.

*C. Challenges experienced by Mashup tools*

The design and service consumption layer in the above mentioned Mashup tools have been reviewed. The purpose of this subsection is to summarize the usability challenges deduced through the review of literature. From the above sections following were the usability shortfalls: the Mashup tools need a background of programming before use, since they user interface was unnecessary complex; they are not easy to learn; their design is complex, as a result they do not look appealing to users.

Most of the examined Mashup tools used a data flow architecture and widgets (i.e. small application that represents APIs). With the composition layer, different techniques were used in different Mashups tool. In this research we used a data flow architecture and widget. The section focuses on the methodology used.

## III. METHODOLOGY

In this research, Mashups usability was determined through the review of relevant literature. User requirements were determined using focus group and SUS. In the following subsections each method is explained as to how it contributed to the research. Using the mixed methods (qualitative and quantitative) we defined our guidelines on how to measure the usability of a Mashups. The mixed methods was used because focus group is qualitative (i.e. raw users response) and SUS is quantitative (i.e. SUS score is a numbers). The low fidelity prototype was used, for instance paper prototypes for Mashup tool that are commercial was used (e.g. MobiMash).The workflow of Mashups between Mashup tool and Mashups editor was designed based on usability guidelines obtained. In all experiments carried out 12 postgraduates were used of which 4 of them masters and 8 of them being honours students. In addition ethical clearance was obtained, and users response we confidential and voluntary.

*A. Focus Groups*

A focus group is defined as an informal method for collecting in-depth information regarding the needs, judgments and feelings of typical users about an interface [9] [10]. In a focus group, about 7 users were selected to discuss the usability of Mashups, to identify the weaknesses of Mashups and what they think can be the solution. This method allowed diverse and relevant issues to be raised; it brought out users' spontaneous reactions, comments and ideas through their interaction [10]. The focus group method was used to obtain users opinions and preference towards Mashup tools. Moreover the usability metrics were obtained using focus group. In this research focus group was used as an informative technique to design and implement the proposed Mashup tool.

### B. SUS - the System Usability Scale

The System Usability Scale (SUS) is a simple, ten-item scale giving a global view of subjective assessments of usability [11]. The SUS scale is generally used after the respondent has had an opportunity to use the system being evaluated, but before any debriefing or discussion takes place. The respondents were asked to record their immediate response to each item, rather than thinking about items for a long time. Twelve postgraduate students from the university were selected as test subjects. The purpose of the SUS was to determine the usability and learnability of the Mashup tools.

These research methods allow the aggregation of all users' behaviours and perceptual functions into a single value for an entire task, using a heuristic. Furthermore, they do not require a fully working system; they can be used with low-fidelity prototypes such as paper prototypes. Next we discuss the research usability variables obtained from formative design using the currently existing Mashups.

### C. System requirements

This section addresses the usability metrics that will guide the design and implementation of SerPro. These usability research variables will guide this research to create a usable Mashup tool. Based on the observed and deduced usability requirements the following lists the research usability metrics:

- Familiarity Mashup tools must match to the real world objects. In order words they should be least surprising.
- Workflow: the visibility of task sequence being executed. That is tasks performance must be showed to users.
- Customizable design: users should be able to add their own components, and ambiguous components must be handled.
- Error prone less and error handling. Possible errors must be reduced. And in case an error takes place, it must be handled without disturbances.
- Documentations tutorials should be quick and easy and API documentation should be clear.

In other words, in order to say a Mashup tool is usable it must adhere to the aforementioned usability attributes. In deeply exploring the Mashup tools usability requirements, Sarraj proposed a set of guideline for Mashup tool designers and a useful framework or model for identifying the specific usability impact factors [12]. Sarraj, further identified 3 parts that a Mashup tool must adhere to: visual support, user interaction support and functional support [12]. The *Visual Support* part is concerned with the usability of the user interface of the Mashups Maker: layout of components, size, colour, metaphors, etc. The *User Interaction Support* part addresses the usability of the Mashups Maker from a user interaction perspective. It groups usability aspects such as cognitive and intuitive interaction support. The *Functional Support* part considers how the users' functional requirements are supported by the Mashups Maker. Next is the design and implementation of SerPro.

The next section discusses the architecture and the prototype of the proposed Mashup tool.

## IV. SERPRO MASHUP TOOL

Service Provisioner web tool (SerPro)-, is the Mashup tool proposed in this research with the aim of enhancing the usability of Mashup tools to novice users. This tool was inspired by Microsoft Popfly and Yahoo! Pipes. The architecture and the prototype of SerPro are discussed in this section.

### A. Architecture

The architecture describes the components that are used by SerPro. The SerPro architecture is divided into three layers: discovery layer, the annotation layer and the composition layer. Each layer is described in the following listing:

#### 1) Discovering layer

In the discovering layer, it is where various web resources are discovered from the web. In this layer we use netflix curator which is Java framework for zookeeper. Netflix curator allows us to publish resources discovered in a repository, which is later used to populate resources/ APIS in the Mashups editor. As resources are discovered they are semantically annotated through annotation layer.

#### 2) Annotation layer

In this layer various components are annotated. The literature has showed that there are mainly two types of web services (sometimes referred as APIs): REST and SOAP. REST web services are described in terms of HTML documentations while on the other side SOAP web services are described in the well-structured form known as XML. It is known that the HTML language is not in a structured format but when well-designed it can be structured. That's the reason why annotation layer uses SWEET to annotate REST-based web services and WSMO4j to annotate SOAP-based web services. These libraries result in a smooth integration of web APIs [13]. When resources are well annotated they undergo the composition process.

#### 3) Composition Layer

The composition layer is where integration of desperate web APIs takes place. This layer is divided into two: visual composition and programmatic composition. Visual composition takes place in the Mashups editor and it is done through widgets and wirings. Programmatic composition takes place on server side (i.e. Mashups engine)

##### a) Visual Composer Layer

Visual composer editor provides the Mashups designing canvas to the user. It shows component list, from which users can drag and drop components onto the canvas in order to connect them. The editor implements the domain syntax. From the editor, it is also possible to launch the execution of a composition through a run button and hand the Mashups over to the Mashups engine for execution.

##### b) Server side composition

Composition data mappers parse component and composition descriptors to represent them in the composition editor at design time and to bind them in the Mashups engine at run time. This tool also comes with a component registration interface for developers, which aids them in the setup and addition of new components to the

platform. On the server side, we have a set of Restful web services, i.e., the repository services, and components services.

*4) Widget design*

This section focuses on how widgets are designed. The widgets design depends on the procedure on the algorithm that is discussed. In Figure 5 below, a widget in the left side and another in the right hand side is shown. The one in the left hand side corresponds to the design time while the one the right side corresponds to runtime. The difference between these binding instants is that in runtime the textboxes that were in design time are converted to two select boxes: first one for source widgets and second one populates response of the selected source.

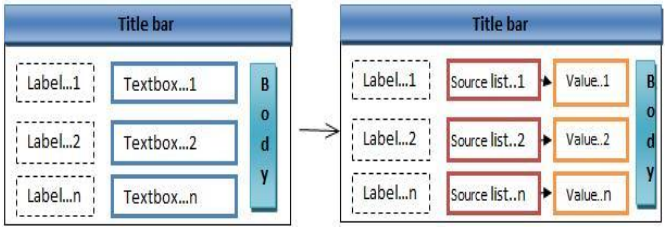



**Figure 5.Widgets at Runtime**

There are two methods to bind runtime and design time; the first automated and second one is manual. The automated binding requires some Artificial Intelligence (AI) algorithms to compose services and the manual approach does not require a specialized algorithm. In this we wish to develop both approaches to meet end-users goals.

Figure 6, below depicts the architecture of SerPro.

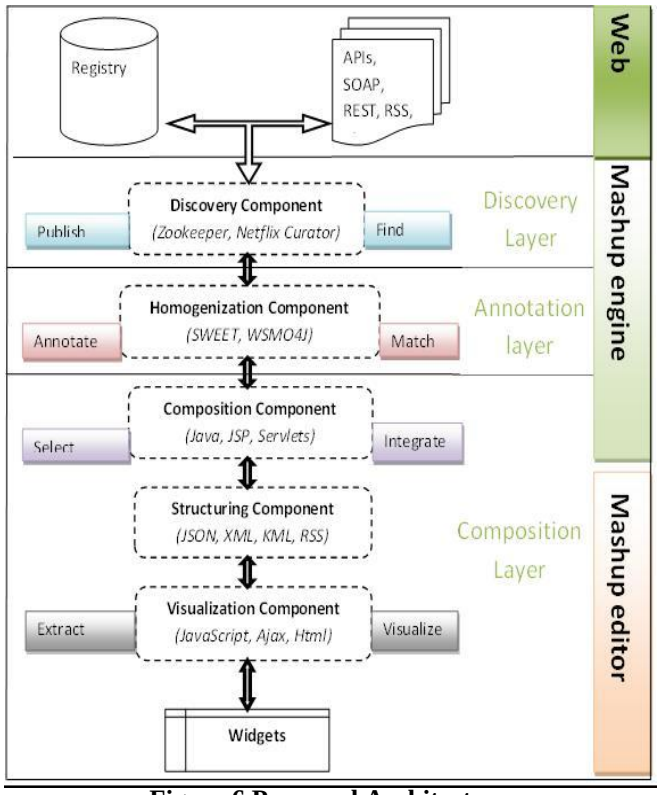



**Figure 6.Proposed Architecture**

Using the architecture in Figure 6, the first prototype of SerPro was implemented and it discussed in the following subsection.

*B. Prototype*

This subsection presents the first prototype of SerPro, which has been designed and developed based on usability requirements and the proposed architecture.

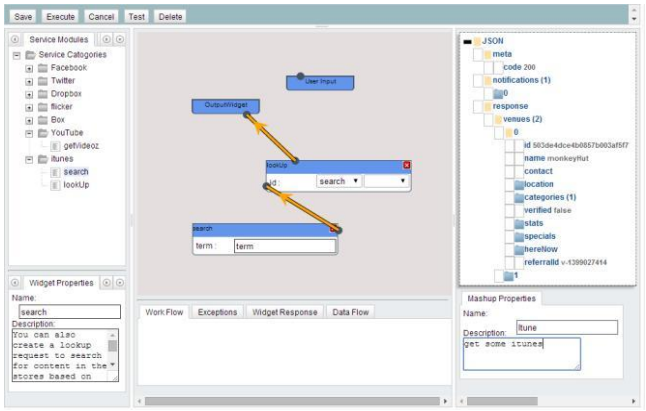


**Figure 7.Mashups Editor**

## V. Experimentation and Results

This section details on the process gathering data and later presents the results obtained from data analyses.

*A. Comparative analysis*

This subsection analyses various Mashup tools together with the proposed Mashup tool based on the formulated usability attributes. In the listing below **"X"** signifies that a Mashup tool supports the usability attribute and "dash (-)" means that the Mashup tool does not support usability attribute. The following are some abbreviations used in the table: RV1=Familiar, RV2=Workflows, RV3=Customizable, RV4=Error handling and RV5=Documentation.

**Table 1.Comparative analysis**

| | RV1 | RV2 | RV3 | RV4 | RV5 |
|---|---|---|---|---|---|
| **Microsoft Popfly** | X | - | X | X | - |
| **Yahoo Pipes** | - | - | X | X | - |
| **Intel MashMaker** | X | - | X | - | X |
| **MobiMash** | X | - | - | - | X |
| **Marmite** | - | X | X | X | X |
| **Dapper** | X | X | X | - | X |
| **SerPro** | X | X | - | X | X |

In terms of usability, SerPro, Dapper and Marmite portrayed better results than the rest, however there were still some usability attributes missing. SerPro, the proposed Mashup tool possessed a room for improvement since it managed to meet some of the research metrics, except customizability. The transparency and visibility of a Mashup tool is one of the factors that can promote. The usability testing carried on the SerPro first prototype using SUS yielded the following results.

*B. SerPro SUS evaluation*

A SUS was used as a method to test its usability. About 12 postgraduate students were given a task to complete using the proposed Mashup tool. In the task they were given a list of REST APIs, and had to choose their preferred ones. After choosing, they were asked to create a simple Mashups application and then record they response on the SUS survey. The minimum threshold for the SUS score is 68, in order to regard a system as a usable and acceptable system. The following figure depicts the obtained SUS scores:

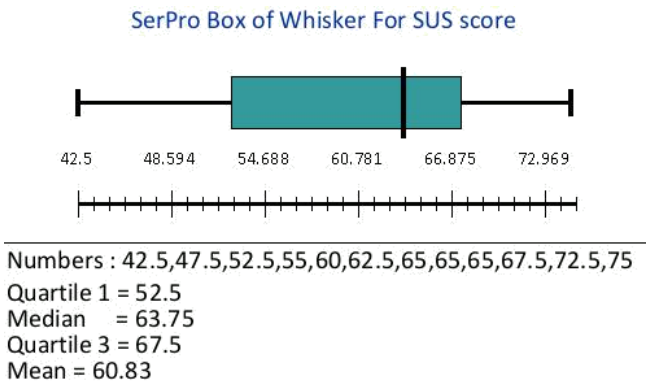



**Figure 8.SerPro Box of Whisker**

It can be noted from the diagram that 25$^{th}$ percentile is 52.5, which means that 25% of the individual SUS scores are below 52.5 and 75% of the scores are above 52.5. It shows that most users were happy with system performance although a SUS score of 68 was not achieved. The interquartile range is 10 which show that 8 out of 12 scores where inside the box (i.e. they were between 52.5 and 67.5). The following table summarizes the SUS score of the evaluated Mashup tools:

**Table 2.Mashup tools SUS scores**

| Mashup tool | SUS score |
|---|---|
| **Yahoo Pipes** | 40.21 |
| **Dapper** | 42.92 |
| **Microsoft Popfly** | 48.33 |
| **IntelMashMaker** | 52.92 |
| **MobiMash** | 45.58 |
| **Marmite** | 57.29 |
| **SerPro** | 60.83 |

As compared to the SUS on its counterparts (i.e. Yahoo Pipes 40.21 and Dapper 48.74), the proposed tool performed well. In future the final prototype of the tool will incorporate some of the shortfalls that had been identified in order to create a complete usable Mashup tool. In addition the SerPro has to be customizable and satisfying to users.

### *C. Research Limitation*

While conducting the research there was no exhausted information on the currently exiting Mashup tools. This evaluation was based on the first prototype, hopefully in the subsequent prototypes there will be an improvement. The publicly available annotation tools and discovery tools were not accessible.

## VI. CONCLUSION

In this paper we compared already existing Mashup tools to the proposed tool (SerPro). Research methods were used to observe actual users evaluating a product and collecting information about the specific ways in which the product is easy or difficult for them. The usability evaluation of the prototype was analyzed and showed that there is still room for improvement. The research hypothesis together with research objectives were answered although there were slight complications. As a result the SUS score of SerPro channelled a great score compared to its counterparts.

In future, since most web resources are manually discovered due to the lack of semantic annotations, an annotation layer which would provide a well-defined structure to APIs will be integrated into the system. Subsequent to that a usable tool that can serve its purpose with ease and flexibility will be achieved.

## VII. ACKNOWLEDGEMENTS

This work is based on the research undertaken within the Telkom CoE in ICTD supported in part by Telkom SA, Tellabs, Saab Grintek Technologies, Easttel, Khula Holdings, THRIP and National Research Foundation of South Africa (UID : 84006). The opinions, findings and conclusions or recommendations expressed here are those of the authors and none of the above sponsors accepts no liability whatsoever in this regard.

## VIII. REFERENCES

[1] ISO 9241-11, “Ergonomic requirements for office work with visual display terminals ( VDTs ) — Part 11 : Guidance on usability,” 1998.

[2] C. Abras, D. Maloney-krichmar, and J. Preece, “User-Centered Design,” *w.Encyclopedia of human coputer interaction*. Thousands oaks:sage publication, pp. 1–14, 2004.

[3] T. Loton, *Introduction to Microsoft Popfly, No Programming Required.* Lotontech Limited, 2008.

[4] M. C. Jones and E. F. Churchill, “Conversations in Developer Communities : a Preliminary Analysis of the Yahoo ! Pipes Community,” in *In Proceedings of the fourth international conference on Communities and technologies*, 2009, pp. 195–204.

[5] R. Ennals, E. Brewer, M. Garofalakis, M. Shadle, and P. Gandhi, “Intel Mash Maker: join the web,” *ACM SIGMOD Rec.*, vol. 36, no. 4, pp. 27–33, 2007.

[6] C. Cappiello, M. Matera, M. Picozzi, A. Caio, M. T. Guevara, and P. Milano, “MobiMash : End User Development for Mobile Mashupss,” in *In Proceedings of the 21st international conference companion on World Wide Web (pp. 473-474). ACM.*, 2012, pp. 473–474.

[7] J. Wong and J. I. Hong, “Making Mashupss with marmite: towards end-user programming for the web.,” in *In Proceedings of the SIGCHI conference on Human factors in computing systems*, 2007, pp. 1435–1444.

[8] B. H. Sigelman, L. Andr, M. Burrows, P. Stephenson, M. Plakal, D. Beaver, S. Jaspan, and C. Shanbhag, “Dapper , a Large-Scale Distributed Systems Tracing Infrastructure,” *Google Res.*, no. April, 2010.

[9] J. Nielsen, L. A. Blatt, J. Bradford, and P. Brooks, “Usability Inspection,” in *conference companion chi ’94*, 1994, pp. 413–414.

[10] J. Hom, *The Usability Methods Toolbox Handbook.* 1998, pp. 2–50.

[11] J. Brooke, “SUS : A Retrospective,” *J. usabilty Stud.*, vol. 8, no. 2, pp. 29–40, 2013.

[12] W. Al Sarraj and O. De Troyer, “Web Mashups makers for casual users: a user experiment.,” in *In Proceedings of the 12th International Conference on Information Integration and Web-based Applications & Services*, 2010, pp. 239–246.

[13] P. Andrews, I. Zaihrayeu, and J. Pane, “A Classification of Semantic Annotation Systems,” *Semant. Web*, vol. 3,, no. 3, pp. 223–248, 2012.

**SabeloYalezo** received his degree in 2009 from the University of Fort Hare and is presently studying towards his Master of Science degree at the same institution. His research interests include web 2.0, web services, web applications, ICTD applications and internet security.

# Transforming learning: a web-based m-learning system for ad-hoc learning of mathematical concepts amongst first year students at the University of Namibia

Hannah Thinyane[1], Ingrid Siebörger[1] and Maria Ntinda[1,2]
Department of Computer Science
Rhodes University[1], P. O. Box 94, Grahamstown 6140
Tel: +27 46 6038291, Fax: +27 46 6361915
and Department of Computer Science
University of Namibia[2]
email: [h.thinyane, i.sieborger]@ru.ac.za[1]; pewantinda@gmail.com[2]

**Abstract- In the last decade, there has been an increase in the number of web-enabled mobile devices, offering a new platform that can be targeted for the development of learning applications. Worldwide, developers have taken initiatives in developing mobile learning (M-learning) systems to provide students with access to learning materials regardless of time and location. The purpose of this study was to investigate whether through the proliferation of broadband networks and mobile handset access it is viable to support first year students enrolled at the University of Namibia (UNAM) to use mobile phones for ad-hoc learning of mathematical concepts. A system, EnjoyMath, aiming to assist students in preparing for tests, examinations, review contents and reinforce knowledge acquired during traditional classroom interactions was designed and implemented. Two pre-intervention user studies were conducted in order to ascertain students level of mobile phone use and desire to use mobile phones for ad-hoc learning purposes. The results of these studies are available in Thinyane, Sieborger and Ntinda (2014, in press) and were used to inform the design of a mobile mathematics application for first year university students. In this paper we discuss the design and implementation of the developed mobile mathematics application, EnjoyMath. The results of a user study conducted with participants at UNAM, ascertained the participants' perception of the usability of the EnjoyMath system and are detailed here. The EnjoyMath system was well received by the first year students at UNAM, resulting in a recommendation for its inclusion in supporting e-Learning initiatives both at UNAM and in general in SADC countries.[1]**

**Index Terms— m-learning, ICT4D, mobile applications**

[1]This work was undertaken in the Distributed Multimedia CoE at Rhodes University, with financial support from Telkom SA, Tellabs, Genband, Easttel, Bright Ideas 39, THRIP and NRF SA (TP13070820716). The authors acknowledge that opinions, findings and conclusions or recommendations expressed here are those of the author(s) and that none of the above mentioned sponsors accept liability whatsoever in this regard.

## I. INTRODUCTION

Information Communication Technologies (ICTs) have advanced significantly over time, proving to be powerful drivers that can enhance living conditions and opportunities around the globe [1]. This in turn has led to numerous investigations regarding how ICTs could be utilised in multiple disciplines including that of Education. To date, access to the Internet in educational environments has increased, supported by less expensive devices such as low-end computers and mobile phones. The advancement in technology in education globally provides access to information through the use of Electronic learning (E-learning) systems. E-learning systems have been used in education for the past few decades, enabling students to access information and content pertaining to the courses they are studying outside of the formal classroom environment. Although computers have been the common device used for accessing information in educational environments, mobile phones are in greater abundance than computers [2, 3, 4]. Currently, there are more than 84 million mobile phones with Internet capabilities in Africa [5]Furthermore, seven out of ten mobile phones are predicted to be Internet-enabled by 2014 in Africa [5]. As a result, researchers are investigating and initiating methods pertaining to how mobile phones could be used in education even though still in its infancy [6].

Although ICTs have advanced globally, countries such as Namibia are faced with the challenge of access to ICT infrastructure. There are a limited number of landlines, and a general lack of electricity and bandwidth in Namibia [7], making it difficult to utilise ICT infrastructure [8]. However, mobile network coverage is accessible by most people in Namibia regardless of their socio-economic level. Namibia has a 65% mobile network coverage [7], and the mobile penetration amongst the Namibian population is high. Mobile Telecommunications Limited (MTC), which is the biggest mobile network provider in the country, ended the year 2010 with a total number of 1.53 million active SIM-card subscribers (International Telecommunication Union, 2011) (although the number of unique subscribers is unknown). As such, a significant number of individuals can connect to the Internet via mobile phones [7]. Given that the

Namibian total population is around two million citizens, this is a significant level of mobile phone penetration. The proliferation of mobile phones and access to mobile phone networks in Namibia provide a platform to support people accessing information for a number of endeavours including educational purposes. To address the problem of limited computing resources in Namibia and especially in educational settings, mobile phones could be used to compliment computers in education.

Every year, students from different educational backgrounds enrol at the University of Namibia (UNAM), some of which enter the university with a poor academic background, especially in Mathematics [9]. As a result, many students experience difficulties in understanding Mathematics at the tertiary level due to a lack of understanding of the prerequisites taught at the primary and secondary levels [10]. One of the biggest challenges faced by tertiary institutes in Namibia is the poor quality of Mathematics and science education at primary and secondary school levels [11]; requiring greater intervention at a tertiary level in order for students to master the required concepts. This coupled with the near ubiquitous access to mobile phones of tertiary education students in Namibia make mobile phones a potential means of extending learning to outside the classrooms; complimenting traditional learning taking place at UNAM.

This paper describes a study undertaken to investigate the use of mobile phones for informal learning of mathematics concepts for use by first year students at UNAM. The paper begins by describing related mobile learning (m-learning) studies in order to determine the state of art in this area. It then briefly describes the design and implementation of a system, EnjoyMath, which UNAM first years made use of. A user study, conducted with the UNAM first years in order to ascertain their perceptions of the EnjoyMath application, is then described and the results of which are discussed. The paper concludes with some of the lessons learned and potential future work

## II. M-LEARNING

A number of researchers define and conceptualize m-learning differently in terms of devices, technology, the mobility of the learners, the mobility of learning and the learners' experience of learning with mobile devices [12]. For example, O'Malley, Vavoula, Glew, Taylor, Sharples, and Lefrere [13] define m-learning as any sort of learning that happens when the learner is not at a fixed, predetermined location, or learning that happens when the learner takes advantage of the learning opportunities offered by mobile technologies. In this paper, m-learning is referred to as learning that takes place regardless of the time and location via a mobile phone.

Significant research has been done in the field of m-learning where different mechanisms have been used to prototype and implement m-learning systems depending on learners' needs. Most of these efforts were centered on text based systems that allowed student-to-student and student-to-lecturer communication using SMS. For example Markett [4] designed two interfaces for in-class and outside class interaction. This was done in order to support interactivity and to facilitate the students' learning. Other approaches explored for m-learning include the use of quizzes with randomly generated questions and the use of games [14, 15].

Mobile applications have been developed to support learning in various curricular areas at different educational levels. For example, Mahamad, Ibrahim Foad and Mohd Taib [16] developed an m-learning system using Open Source Software for primary school learners. It allows students to use quizzes to learn and the students' progress and performances were tracked using graphs. In South Africa, a Web base application to assist secondary school learners (grade 8-9) in mathematics was developed using a MySql database & PHP on the server side and using J2ME and/or mobile browser on the Client side [17].

Developing m-learning applications remains challenging because developers need to support the wide variety of devices that are available. Variance across handsets includes: screen size, input technique, battery power, connectivity supported by the device and the device operating system. When designing m-learning systems, it is advisable that students use their own mobile phones for testing and lots of interaction should be encouraged [18].

Several requirements for m-learning applications are typically highlighted in most m-learning research projects. The requirements include [18, 19, 20]:

- the mobile devices to be used in the study should be identified prior to development and it is advisable to use the students' own mobile devices;
- those parts of the activities to be supported by mobile technologies and those to be supported by other technologies should be identified;
- encourage interactivity amongst the participants; and
- students' receptivity should be identified - receptivity should be relative to the proposal of using mobile devices for educational purposes.

## III. ENJOYMATH

Human-centred design (HCD) was employed in the design of the EnjoyMath system, which uses an iterative approach to design and implementation; cycling through context, requirements, design and testing. Pre-intervention questionnaires were conducted with students from UNAM as well as Rhodes University in order to help inform the design of the application. A detailed discussion and analysis of these interventions and their findings are available in Thinyane, Siebörger and Ntinda [21]. For our purposes here it is suffice to say that we found 79% of the participants in the pre-intervention study made use of their mobile phones in order to access the Internet. In addition, of those surveyed, only 7% had phones which were not Internet enabled. Furthermore, research has highlighted the benefit of having participants use their own mobile phones in mobile application research work [22, 23], because users are more familiar with their own phones and it further avoids additional costs in terms of training, support and provision of phones to participants. As a result, it was decided that a web-based application could be developed in

order to avoid having to develop multiple applications for all phone types used amongst first year students.

An initial protype of the application was developed using Flowella, a rapid prototyping tool for mobile designers that enables usability to be tested and designs to be refined prior to coding using screen mock-ups like images and pencil sketches. The prototype was presented to the mathematics lecturers at Rhodes and UNAM, randomly selected first year mathematics students and a group of randomly selected students with knowledge in mathematics for comments and improvement to the user interface design. The feedback was incorporated into the design of the application. Furthermore, the application design took into consideration the screen sizes of mobile phones and their battery lives. This impacted on the arrangement of the content and resulted in avoiding the use of unnecessary navigations, Javascript, CSS and flash elements in order to minimize the processing power needed and thereby maximizing the battery power of the mobile phone.

HTML, JavaScript (supported by Ajax and jQuery), PHP and a MySQL database were used to implement the system. HTML, Ajax and jQuery were used on the client side, while PHP was used on the server side and a MySQL database was used for data storage.

## IV. User Study: EnjoyMath Evaluation

The post-intervention study aimed at ascertaining the mobile phones used by the participants to access the EnjoyMath system and the perceived ease-of-use (users' satisfaction) of the system. In addition, factors that prevented the participants from accessing the EnjoyMath system were investigated.

### *A. Participants mobile phone models*

To ascertain the mobile phones used to access the EnjoyMath system, participants were asked to state the make and model of the mobile phones they used to access the M-learning system. This was important to know in order to improve on the user interface in case the participants had problems accessing or viewing the EnjoyMath system. Of the 23 participants, only 14 listed the mobile phones they used to access the M-learning system. Participants used different mobile phones, with screen sizes ranging from 240x320 to 320x480.

### *B. System usability*

One of the factors ascertained during the post-intervention study was the usability of the system. Usability of the system measures the perceived ease-of-use (ascertaining whether the designed system was: learnable, efficient, memorable, had a low error rate or easily recovered from errors, and was satisfying [24]). The SUS questionnaire was used to measure the overall performance of the system; adopted because it was identified to be a quick way of measuring the overall usability of a system [25]. Statements in the SUS questionnaire cover aspects such as the need for support, training and complexity of use [26]. Moreover, the SUS questionnaire was found to be the best at revealing the usability of a system compared to other usability questionnaires in previous research studies investigated [27]. Participants were asked to complete the SUS questionnaire, which consists of Likert scale type questions with five options ranging from strongly disagree to strongly agree (1=strongly disagree, 2= disagree 3=neutral, 4=agree and 5=strongly agree). The mean and the standard deviation of each statement were calculated and are shown in Table 1. Statement 1 ("I think that I would like to use this system frequently.") received the highest mean of all responses with a mean score of 4.56. A mean of 4.56 implies that on average, the participants strongly agreed that they would use the EnjoyMath system frequently. In contrast, statement 2 ("I found the system unnecessarily complex.") received the lowest mean with a score of 1.44. This implied that on average, participants strongly disagreed that the EnjoyMath was unnecessarily complex.

| SUS Statement | Mean | Std dev |
|---|---|---|
| 1. I think that I would like to use this system frequently. | 4.56 | 0.70 |
| 2. I found the system unnecessarily complex. | 1.44 | 0.86 |
| 3. I thought the system was easy to use. | 4.11 | 0.96 |
| 4. I think that I would need the support of a technical person to be able to use this system. | 2.39 | 1.61 |
| 5. I found the various functions in this system were well integrated. | 4.11 | 1.08 |
| 6. I thought there was too much inconsistency in this system. | 1.83 | 0.79 |
| 7. I would imagine that most people would learn to use this system very quickly. | 4.11 | 1.23 |
| 8. I found the system very Awkward to use. | 2.56 | 1.65 |
| 9. I felt very confident using the system. | 4.33 | 0.69 |
| 10. I needed to learn a lot of things before I could get going with this system | 2.61 | 1.58 |

**Table 1: SUS statements and the mean and standard deviation**

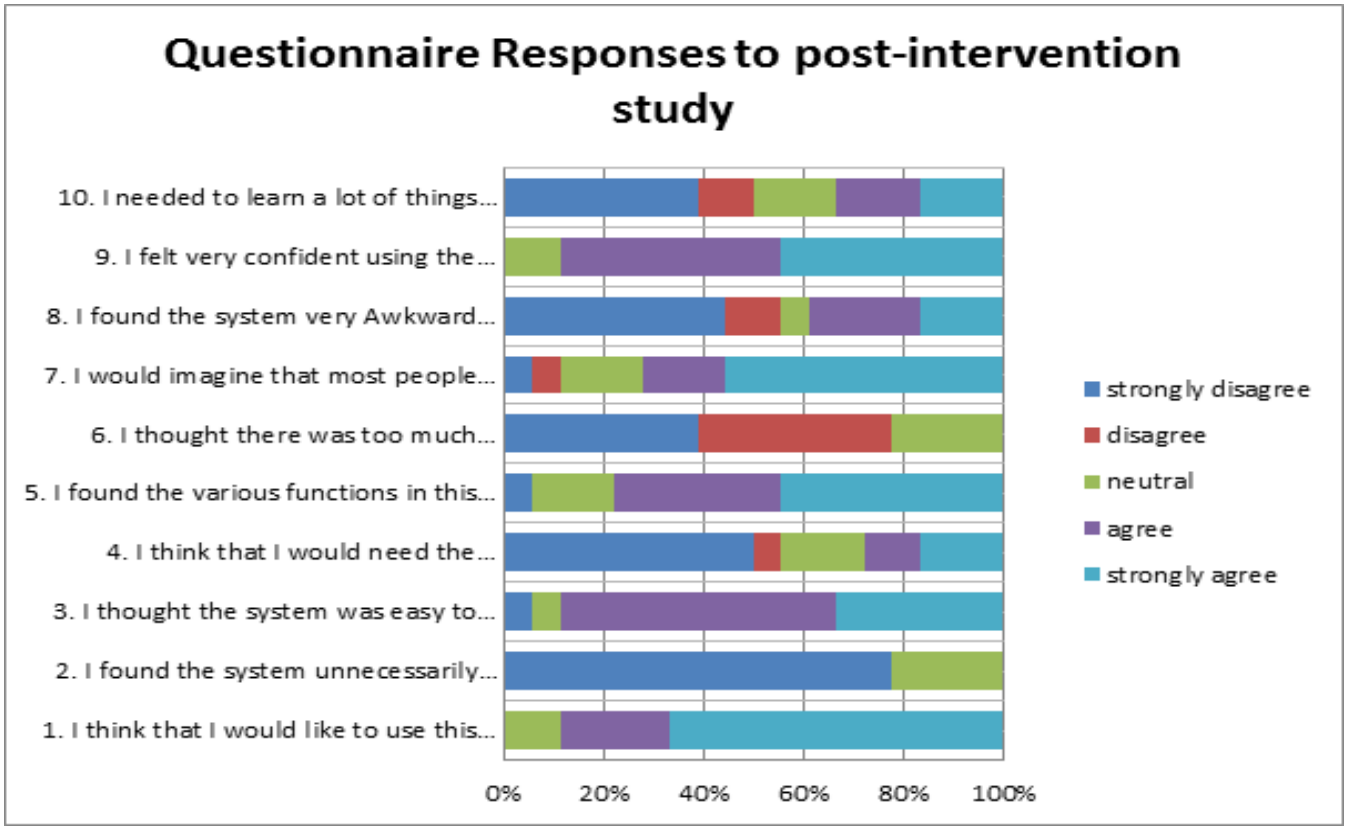



**Figure 1: Mean of the responses from the post intervention study**

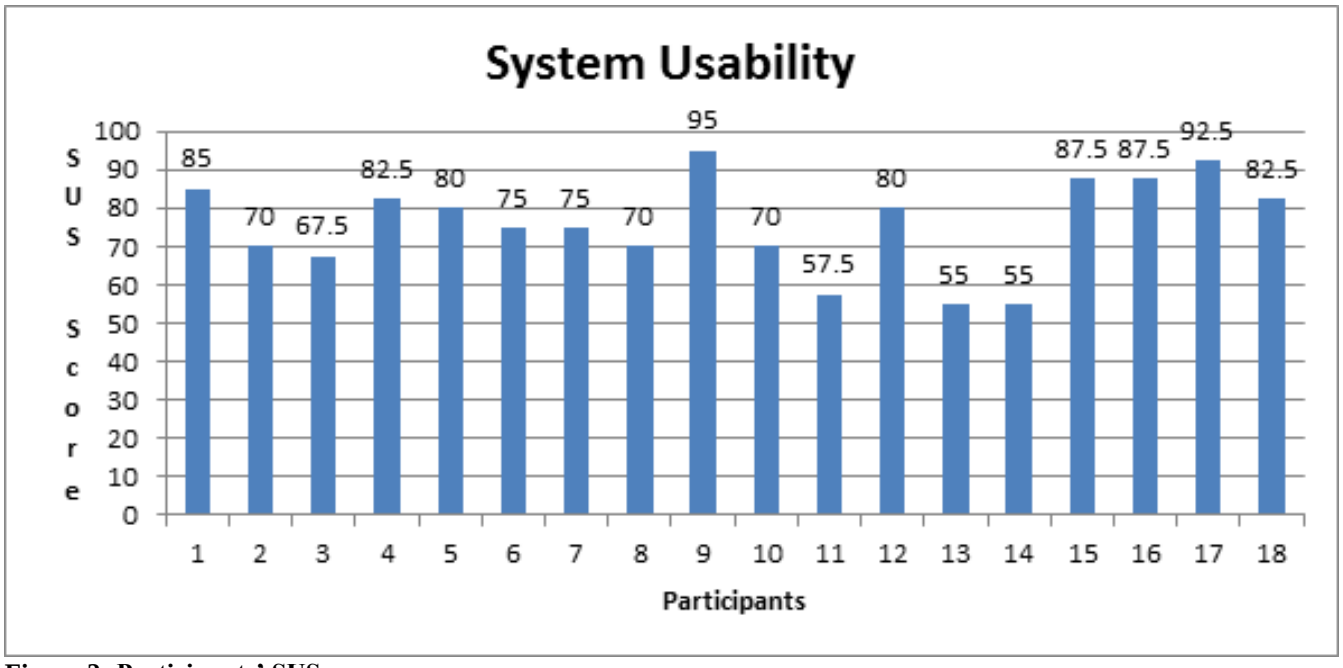



**Figure 2: Participants' SUS scores**

Data derived from the questionnaire were grouped and are shown in Figure 1. It can be seen that the participants tended to strongly agree with positively worded items and strongly disagree with negatively worded items. Based on these results, it can be inferred that the participants were satisfied with the developed M-learning system.

The SUS scores of the participants were calculated and then used to calculate the mean SUS score. The mean SUS score was used to conclude whether the EnjoyMath system was usable or not. As recommended by Lewis and Sauro [25] all questions omitted by the participants were given three points because 3 is the neutral point Brooke (personal communication, November 14, 2012). This was done to avoid discarding participants' data that had completed most of the SUS items [28] Brooke [28, p5] stated that "All items should be checked". The score of each statement was obtained by subtracting one from the scale position (scale positon-1) for even-numbered questions (negatively worded items) and subtracting the scale position from 5 (5-scale position) for odd-numbered questions (positively worded items) [28]. The SUS score of each participant was then obtained by multiplying the sum of the statement score contribution by 2.5 [28]. The effect of the 2.5 multiplier is to get a scale ranging from 0 to 100 rather than 0 to 40

(Brooke, personal communication, November 14, 2012) [29]

Little literature on the SUS mean score's sufficiency has been published [30]. However, Bangor, Kortum and Miller [31] conducted a comprehensive analysis on products that used the SUS questionnaire. This study provided details on what constituted an acceptable SUS mean score and found that a product with a SUS mean score of less than 50 is considered to be cause for significant concern and is judged to be unacceptable. In addition, a product with a SUS mean score between 50 and 70 is considered to be a candidate for scrutiny and should be improved. A product with a SUS mean score above 70 is considered passable and a true superior product should score better than 90. The method used to interpret the average mean SUS score in this study was adopted from (Bangor et al., 2008). Figure 2 shows the SUS score of the participants. It was found that seven participants had a SUS score below 75 and eleven participants had a SUS score of 75 or above. The lowest recorded score of a participant was 55 and the highest was 95. The overall mean SUS score for all participants was 76, indicating that the EnjoyMath system was generally perceived by the participants to be passable.

## V. Conclusion

The purpose of this study was to investigate whether it is viable for first-year students enrolled at UNAM to use mobile phones for ad-hoc learning of mathematical concepts, as a compliment to traditional learning. Numerous studies investigated the use of mobile phones in education and found that M-learning enhances students' learning at the tertiary level [32, 33]. In this study, participants showed interest in the developed prototype and 78% were enthusiastic about using new technologies (e.g. mobile phones) in education. This is a common thread found in most M-learning studies, for example in Bradley and Holley's reserach [23]. Participants (78%) in this study suggested that the M-learning materials would enhance students' understanding and assist them with regard to focusing on practicing their skills in Mathematics. These results encourage the use of mobile technology (mobile phones) solutions for learning at UNAM.

It was not possible in this study to determine whether the M-learning system improved the students' overall results in Mathematics as the M-learning system was implemented for only a short period of time. Furthermore, it is difficult to correlate changes in Mathematics performance directly to a particular intervention. That said, the general positive response to using the system could indicate increased practice and support of mathematical concepts for the participants, which would hopefully result in an improved mathematical ability in the future. In this project, participants indicated that an M-learning system is viable. This was based on the results obtained via the SUS questionnaire. Although the results obtained from the SUS questionnaire did not provide insight into where to focus improvements to the system, the score of the SUS questionnaire provided feedback on whether the EnjoyMath system was usable. This system was found to be passable with a SUS score of 76. This project did not reveal any negative effects in the use of mobile phones in education. Therefore, mobile phones could be used as a technology for educational reform and to increase access to educational material.

## VI. References


[1] World Economic Forum. (n.d., 8 May). *Global information technology*. Available: http://www.weforum.org/issues/global-information-technology

[2] D. Bhatia, A. Bhavnani, R. W. W. Chiu, S. Janakiram, and P. Silarszky. (2008, 8 May). *The role of mobile phones in sustainable rural poverty reduction*. Available: http://siteresources.worldbank.org/EXTINFORMATIONANDCOMMUNICATIONANDTECHNOLOGIES/Resources/The_Role_of_Mobile_Phones_in_Sustainable_Rural_Poverty_Reduction_June_2008.pdf

[3] T. H. Brown, "Towards a model for m-learning in Africa," *International Journal on E-learning*, vol. 4, pp. 299-315, 2005.

[4] C. Markett, I. A. Sánchez, B. Tangney, and S. Weber, "Using short message service to encourage interactivity in the classroom," *Computers & Education*, vol. 46, pp. 280-293, 2006.

[5] C. Fripp. (2011, 8 May). *Africa: 84 million mobile devices are internet enabled*. Available: http://www.itnewsafrica.com/2011/10/africa-84-million-mobile-devices-are-internet-enabled/http://www.itnewsafrica.com/2011/10/africa-84-million-mobile-devices-are-internet-enabled/

[6] L. F. Motiwalla, "Mobile learning: A framework and evaluation," *Computers & Education*, vol. 49, pp. 581-596, 2007.

[7] R. Gomez. (2009, 16 September). *Namibia public access landscape study*. Available: http://faculty.washington.edu/rgomez/projects/landscape/country-reports/Namibia/1Page_Namibia.pdf

[8] D. G. Alemneh and S. K. Hastings, "Developing the ICT infrastructure for africa: The influence on global scholarship," *Journal of Education for Library and Information Science*, vol. 47, p. 4, 2006.

[9] C. Academics without Borders. (2011, 9 May). *Academics without Borders Canada - Namibia*. Available: http://www.awbc-usfc.org/projects/namibia/

[10] Jason. (2009, 8 May). *How to paginate data with php*. Available: http://net.tutsplus.com/tutorials/php/how-to-paginate-data-with-php/

[11] Haufiku.A., "An investigation of lower primary teachers' content knowledge of mathematics in Ohangwena region in Namibia," Rhodes University, Grahamstown, 2008.

[12] J. Traxler, "Defining, discussing and evaluating mobile learning: The moving finger writes and having writ.." *The International Review of Research in Open and Distance Learning*, vol. 8, 2007.

[13] C. O'Malley, G. Vavoula, J. Glew, J. Taylor, M. Sharples, and P. Lefrere. (2004, 9 May). *Guidelines for learning/teaching/tutoring in a*

*mobile environment*. Available: http://www.mobilearn.org/download/results/guidelines.pdf
[14] Q. Sun, C. Ardito, P. Buono, M. F. Costabile, R. Lanzilotti, T. Pederson, and A. Piccinno, "Experiencing the past through the senses: an m-learning game at archaeological parks," *IEEE Multimedia*, vol. 15, pp. 76-81, 2008.
[15] A. Purohit, N. Bhatia, and S. Arumugam. (8 May). *Matheasy: An application for m-learning in mathematics*. Available: http://matheasy.webstarts.com/uploads/MathEasy_Report.pdf
[16] S. Mahamad, M. N. Ibrahim, I. A. M. Foad, and S. M. Taib, "Open source implementation of m-learning for primary school in Malaysia," *International Journal of Social Sciences*, vol. 3, pp. 309-313, 2008.
[17] M. Matthee and J. Liebenberg, "Mathematics on the move: Supporting mathematics learners through mobile technology in South Africa," presented at the mLearn Conference, 2007.
[18] C. F. Batista, P. A. Behar, and L. M. Passerino, "Mobile learning environments and applications m-learning in mathematics: mapping requirements," in *International Conference on Interactive Collaborative Learning*, 2010.
[19] H. Singh, "Leveraging mobile and wireless internet," *Learning Circuits*, vol. September, 2003.
[20] D. Parsons, H. Ryu, and M. Cranshaw, " A study of design requirements for mobile learning environments," in *Sixth International Conference on Advanced Learning Technologies*, pp. 96-100.
[21] H. Thinyane, I. Sieborger, and M. Ntinda, "First year students' use of mobile phones and perceptions of m-learning in two SADC countries," in *EDULEARN*, Barcelona, Spail, 2014.
[22] L. Dyson, A. Litchfield, E. Lawrence, and A. Zmijewska, "Directions for m-learning research to enhance active learning," in *Proceedings of the ASCILITE-ICT: Providing choices for learners and learning*, 2007, pp. 587-596.
[23] C. Bradley and D. Holley, "How students in higher education use their mobile phones for learning.," in *Proceedings of mLearn Conference*, 2010.
[24] J. Nielsen, *Usability engineering*. San Diego: Morgan Kaufann, 1993.
[25] J. Lewis and J. Sauro, "The factor structure of the system usability scale," in *Human Centered Design*, ed, 2009, pp. 94-103.
[26] A. H. S. Chan, H. W. C. Lo, and A. W. Y. Ng, "Measuring the usability of safety signs: a use of system usability scale (SUS)," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, 2011.
[27] J. N. Stetson and T. S. Tullis, "A comparison of questionnaires for assessing website usability," in *Usability Professional Association Conference*, 2004.
[28] Brooke, "SUS-a quick and dirty usability scale," in *Usability evaluation in industry*, ed, 1996, pp. 189-194.
[29] Y. Laouris and Eteokleous, "We need an educationally relevant definition of mobile learning," in *Proceedings of mLearn*, 2005.
[30] S. Michaels. (2012, 8 May). *Usability: Using the system usability scale (SUS) in practice.* Available: http://www.fusebox.com/2012/02/usability-using-the-system-usability-scale-sus-in-practice/
[31] A. Bangor, P. T. Kortum, and J. T. Miller, "An empirical evaluation of the system usability scale," *Intl. Journal of Human–Computer Interaction*, vol. 24, pp. 574-594, 2008.
[32] V. Kalloo and P. Mohan, "Mobilemath: An innovative solution to the problem of poor mathematics performance in the Caribbean," *Caribbean Teaching Scholar*, vol. 2, 2012.
[33] K. Whattananarong, "An experiment in the use of mobile phones for testing at King Mongkut's Institute of Technology North Bangkok, Thailand," in *International Conference on Making Education Reform Happen: Learning from the Asian Experience & Comparative Perspectives*, 2004.

**Hannah Thinyane** received her PhD from the University of South Australia in 2006 in Augmented reality and is now an associate professor at Rhodes University. Her research interests include mobile computing and ICT for devleopment (ICT4D)

**Ingrid Sieborger** received her Masters degree in Computer Science with Distinction in 2006 from Rhodes University. She is presently studying towards her PhD at the same institution. Her research interests include the use of ICTs in education, ICTs for development (ICT4D) and computer networks.

**Maria Ntinda** received her Masters degree in Computer Science with Distinction in 2014 from Rhodes University. Her research interests include mobile computing and ICTs for development (ICT4D).

# LIMITED RANGE COMMUNICATIONS

# Comparison of Energy-based Leader Selection Algorithms in Wireless Mesh Networks

Olukayode Oki, Pragasen Mudali, Nathi Zulu and Matthew Adigun
Department of Computer Science
University of Zululand,
South Africa
Email: okikayode@gmail.com

**Abstract- Wireless Mesh networks (WMNs) are gaining popularity as a scalable replacement for Wired Network infrastructure. The increasing popularity of WMNs has prompted the development of security mechanisms. The newly-ratified IEEE 802.11s mesh networking standard specifies a security mechanism that builds upon the IEEE 802.11i security standard meant for wireless local area networks. The IEEE 802.11s security mechanism specifies the existence of a single Mesh key Distributor (MKD) which helps to authenticate new nodes that join the network. However, there is no mechanism for selecting a new MKD if the current MKD is unreachable or has failed. This scenario can arise due to the dynamic nature of WMN backbone topologies, wireless link variability in deployed networks and battery depletion in battery-powered WMNs. MKD selection in energy-scarce WMN deployments can be performed by adapting energy-based Leader Selection Algorithms from Wireless Sensor Networks. This paper evaluates the influence of heterogeneous and homogeneous energy-based leader selection algorithms on MKD selection when subjected to different rounds and network sizes. The evaluation showed that the heterogeneous-based LSAs (EECS and UDAC) outperform the homogeneous-based LSAs (LEACH and EECHA) in the achieved performance for communication overhead cost and the energy consumption rate. Whilst the homogeneous-based LSAs outperform the heterogeneous LSAs in terms of leader selection delay.**

***Keywords — Authentication, Heterogeneous, Homogeneous, Mesh key Distributor, Wireless Mesh Network***

## I. INTRODUCTION

Wireless mesh networks (WMNs) provide a scalable architecture for deploying networks in areas without prior networking infrastructure [1]. Thus, WMNs are useful in rural scenarios. A rural African WMN deployment often means that mesh devices are battery-powered due to the lack of stable electrical supplies.

A typical WMN (See Figure 1) is comprised of two classes of devices: backbone devices and client stations [2]. The backbone devices consist of Mesh Points (MPs) and Mesh Access Points (MAPs). The backbone of a WMN is a self- configuring network, in which all MPs and MAPs can route traffic either directly to a destination (if possible) or

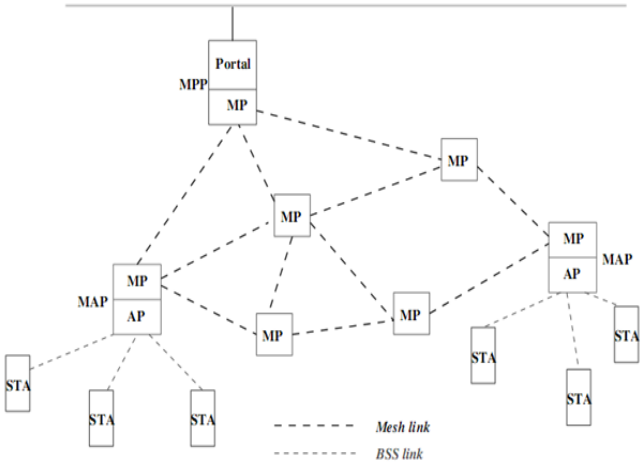



**Figure 1: Wireless Mesh Network Architecture [15]**

via a multi-hop path [3]. The WMN topology is dynamic in nature as both backbone devices and client stations can enter and exit the network. Network exits may be the result of battery drainage in rural areas. Backbone devices may also experience a temporary lack of connectivity due to the transient nature of wireless links when WMNs are deployed [4],[5],[6].

The security aspects of WMNs have not received as much attention as routing protocols and energy-efficiency [14]. The scalable and ad-hoc nature of a WMN increases its vulnerability to different security threats like impersonation [7] where devices can leave the network and an impersonator can re-join the network under pretense. Thus, the authentication of devices is a critical feature.

Authentication must occur between backbone devices as well as between MAPs and their associated stations. These authentication mechanisms are specified in the IEEE 802.11s standard for mesh networking using IEEE 802.11 technology. The IEEE 802.11s authentication mechanism is adapted from the IEEE802.11i standard designed for Wireless Local Area networks. A key feature of the 802.11s authentication mechanisms is the presence of a Mesh Key Distributor (MKD). The MKD is a centralised repository for the upcoming 4-way handshake that occurs between the MAPs and the mesh stations. The authentication process is wholly reliant on the presence and availability of the MKD but, due to the ad-hoc nature of the WMN backbone, the transient nature of wireless links when deployed and battery drainage (particularly in rural areas); there may be times when the MKD is neither present nor available. This scenario requires the efficient selection of a new MKD so

that the device authentication process is not compromised. To the best of our knowledge, there is no proposed mechanism for the selection of a replacement for the unavailable MKD.

In Africa and most other developing countries where electricity supplies are not reliable, the MKD can become unavailable due to power outage, battery depletion or transient wireless links. Currently, there is no leader selection protocol for selecting a new MKD if the current MKD fails or dies [7]. Hence, this study seeks to evaluate the performance of energy-based leader selection algorithms in the context of MKD selection in WMNs.

The evaluation of four energy-based LSAs (EECS, UDAC, LEACH and EECHA) shows that, some of them can be utilised in the context of finding a new MKD in a WMNs. Heterogeneous energy-based LSAs (EECS and UDAC) outperformed the homogeneous energy-based LSAs (LEACH and EECHA) when communication overhead and the energy consumption rate are considered. The homogeneous energy-based LSAs outperform the heterogeneous energy-based LSAs when the MKD Selection Delay is considered.

The remainder of this paper is organised as follows. Section III presents a review of existing studies in the energy based leader selection algorithms for wireless sensor network and how it can be adopted in MKD selection for WMNs. Section IV details the simulation setup employed while Section V discusses the measurement methodologies used in this study. Section VI discusses the performance evaluation results for the simulation, whilst the paper is concluded in Section VII.

## II. IEEE 802.11s SECURITY

There are two types of security key holders: a Mesh Key Distributor (MKD) and Mesh Authenticators (MA) [7]. A Mesh Point (MP) can assume the role of the MKD and a MA at the same time. Both roles are optional. The MKD is the centre for key generation and authentication, delegating some of its work to the MAs. MPs are regular stations which have to be authenticated by an MA or the MKD before they can participate in the network. A MP with MA functionality plays the 802.1X authenticator role and a MP without the MA functionality plays the 802.1X supplicant role. An MKD and MA can be co-located with MA, and can manage authentication and key distribution for both MA and a supplicant. In a 802.11s WMN, there exists one MKD, multiple MAs and supplicants. A supplicant can become an MA after it passes security key holder association with the MKD. Considering an MP in an IEEE 802.11s secure network, when the MP needs to establish a secure link with a peer MP, a peer link setup procedure is first executed (step 0 in Figure 2). In this initial step, the role of an MP is determined and security policy is selected. Whether an MP and its peer MP are an 802.1X authenticator or supplicant MP is determined in the peer link management. As shown in this architecture, there is only one MKD with which multiple MAs are associated. A supplicant performs security authentication through MAs. The set of MAs, supplicants, and the single MKD form an MKD domain (MKDD). Optionally, the MKD is connected to an AS through which 802.1X authentication is executed.

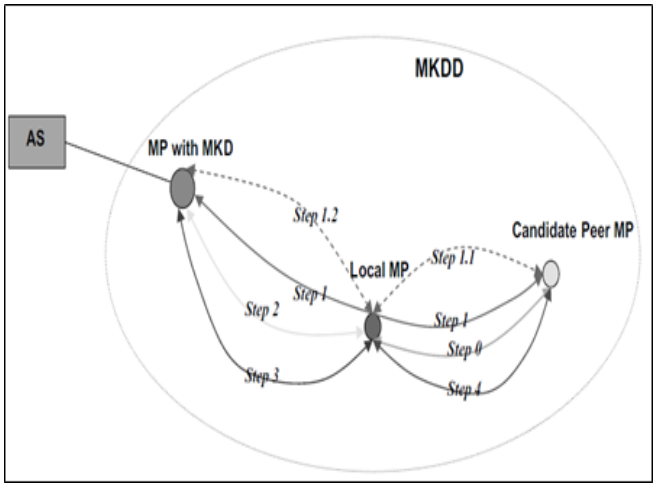



**Figure 2: Major function blocks of 802.11s Mesh Security [8]**

## III. RELATED WORKS

Leader Selection Algorithms (LSAs) have been conceived for the domain of Wireless Sensor Networks. Sensor networks typically comprise many hundreds of devices transmitting to a sink where sensor data is captured [See Figure 3]. LSAs are used to ease the management of sensor devices and to reduce the communication overhead created by the large network size. LSAs create hierarchical networks by selecting Cluster Heads to form a communication backbone to the sink. Cluster Heads (CHs) are selected from groups of sensor devices in close proximity to each other and the CH selection criteria can vary widely.

In this paper, attention is paid to LSAs that employ energy consumption or battery levels in their selection criteria. Such energy-based LSAs are typically classified as Homogeneous and Heterogeneous energy based selection. In heterogeneous leader selection, the node with higher remaining energy becomes the cluster head and those with lower remaining energy become ordinary cluster members. Whilst in homogeneous leader selection, it is assumed that all the nodes in the network are having equal remaining energy; hence the selection of a new cluster head is done stochastically.

### *A. Homogeneous Energy-based LSAs*

[9] presents a new approach of an Energy Efficient Homogeneous Clustering Algorithm (EEHCA) for WSN in which the lifespan of the network is increased by ensuring a homogeneous distribution of nodes in the network. In EEHCA, a cluster leader is randomly selected initially. This study was simulation based and MATLAB was used as the simulation tool. In [9], power consumption was used as a performance metric and the result shows that the proposed algorithm extends the network lifetime.

In [10], a hierarchical cluster algorithm for sensor network called Low Energy Adaptive Clustering Hierarchy (LEACH) was proposed. The LEACH algorithm introduced in [10] set the whole network into small cluster and selects a leader for each cluster. Generally, cluster leader loses their energy faster compare to other nodes because cluster leader require more energy to transmit data to the base station (BS).

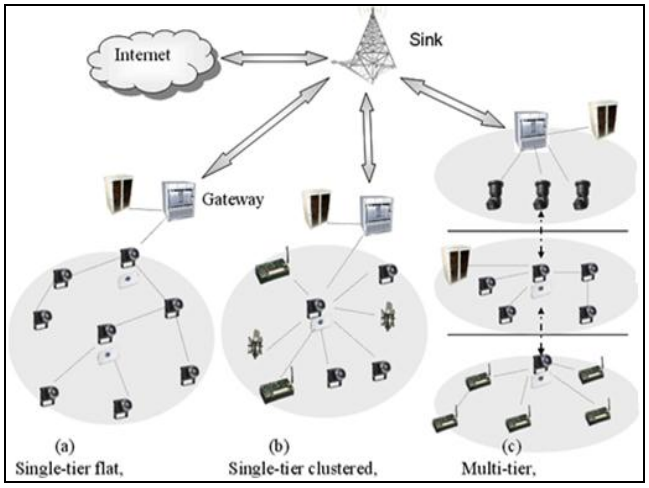



**Figure 3: Wireless Sensor Networks [13]**

Hence, LEACH uses random selection as criteria to interchange cluster leaders. This study was simulation based and the simulation was done using NS2. Three performance metrics were considered; network lifetime, number of cluster leader per round and energy consumption. The results of this study shows that LEACH is energy efficient and only 5% of total number of node can be cluster leader per round.

### *B. Heterogeneous Energy-based LSAs*

The energy-based leader selection algorithms have been well studied in the context of wireless sensor networks. In [11], an EECS novel clustering Schema for WSN was presented. The study utilized the remaining energy as the criteria for selecting a cluster head using local radio communication. EECS introduces a novel technique to balance the load among the cluster leaders. EECS is a distributed and energy efficient algorithm in nature which makes it more suitable for larger network such as WMNs. In EECS algorithm, each cluster has it cluster leader. The study was simulation based and the tool that was used for simulation is MATLAB. Two performance metrics were considered; network lifetime and total energy consumption. The results of this study show that EECS prolong the network lifetime.

In [12], a novel clustering algorithm which maximizes the network lifetime by reducing the number of communication among sensor nodes was presented. The study also incorporates new distributed cluster formation method that enables self-organization of huge number of nodes and this feature will make this study to be more suitable for evaluation in the context of MKD selection for WMNs. The UDAC approach maintains constant number of clusters by prior selection of cluster leader and rotating the role of cluster leaders to even distribute energy load among all sensor nodes. In [12], remaining energy was used as criteria for selecting a cluster leader in a distributed manner. The study was a simulation based and the tool that was used for simulation is NS2. Three performance metrics that were considered includes: energy dissipation rate, number of cluster head per round and network lifetime. The results of this study show that UDAC reduce energy consumption by employing clustering techniques.

## IV. EXPERIMENTAL SETUP

Leader Selection Algorithms (LSAs) are originally intended to select many CHs. For the purposes of this study, the selected LSAs were amended so that only one CH is selected. This constraint allows for the selection of only one mesh backbone device to serve as a replacement MKD. Thus, the IEEE 802.11s specification that there exist only one MKD will not be violated.

The Network Simulator version 2.35 (NS2) software running on Ubuntu operating system was employed as the simulation tool for this study. Various network sizes, ranging from 50 to 500 stationary WMN backbone nodes were distributed over a rectangular 1000m x 1000m flat space for 1000s of simulated time. Table 1 contains the additional simulation setup details that were used for all the experiments in this study.

**Table 1: SIMULATION SETUP DETAILS**

| | |
|---|---|
| Simulation Time | 1000 Seconds |
| Number of Nodes (nn) | 50 - 500 nodes |
| Network Area | 1000m x 1000m |
| Mac protocol | IEEE 802.11 |
| Nodes movement | Static |
| Initial Energy | 5.0 Joule |
| Transmission Power | 0.6 W |
| Receiving Power | 0.3 W |
| Idle and Transition Power | 0.2 W |

## V. MEASUREMENT METHODOLOGY

In this study, four energy-based Leader Selection Algorithms (EEHCA, EECS, LEACH and UDAC) originally developed for Wireless Sensor Networks were simulated. These algorithms were evaluated using the *Communication Overhead*, *Leader Selection Delay* and *Energy Consumption Rate* metrics. Each simulation was repeated five times and the average result is presented.

The following measurement procedure were used for each of the metrics been measured.

### *A. Communication Overhead*

Communication overhead is the sum of the total number of control packets sent and the total packets received between the nodes in the network, during the process of selecting a cluster leader. This metric is used to compute the total communication cost between the nodes in the network. The lower the communication overhead value, the better the algorithm performance. The communication overhead is calculated using the formula below:

$$Communication\ Overhead = \sum_{1}^{n} SentMessage_n + ReceivedMessage_n$$

Where n = number of backbone nodes

### B. *Leader Selection Delay*

Leader selection delay is the time taken for Leader Selection Algorithms to successfully select one node as an MKD. It is calculated based on the time taken for all events to exchange messages between nodes on the network. This time ends when the selected node sends an advertising message for an MKD. Leader selection delay metric will help us to know which Leader Selection Algorithm will take the minimal time to select a leader which is very important for selecting an MKD in WMN, since MKD is meant to perform security measures such as authentication of new nodes on the network. The lower the leader selection delay value, the better the algorithm performance.

### C. *Energy Consumption Rate*

Energy consumption rate refers to the rate at which energy is being consumed by the nodes in the course of selecting a new MKD. Different network node states such as sleeping state, active state and idle state also consumes a certain amount of energy. The lower the energy consumed by the node, the better the performance. The energy consumption rate is calculated using the formula below:

$$Rate = \frac{EnergyConsumption}{time}$$

## VI. EVALUATION RESULTS

The results of the performance evaluation of both heterogeneous (EECS and UDAC) and homogeneous (LEACH and EEHCA) energy based leader selection algorithms when respectively subjected to different leader selection rounds and various network sizes are presented in this section. Two hundred nodes were used for leader selection rounds, while 50 to 500 nodes were varying for network sizes.

### A. *Communication Overhead*

The purpose of this experiment was to determine the communication overhead cost for both the heterogeneous and homogeneous energy-based LSAs when respectively subjected to different leader selection rounds and various network sizes. Communication overhead is considered in order to find out which Leader Selection Algorithm incur low communication cost among the nodes in the network, while selecting an MKD leader. Figures 4a and 4b depict the results of the communication overhead cost for both the heterogeneous and homogeneous energy-based LSAs when respectively subjected to different leader selection rounds and various network sizes.

In Figure 4a, it can be observed that both the homogeneous- and heterogeneous-based LSAs communication overhead increases gradually as the number of rounds increases. It can also be observed that the Heterogeneous-based (EECS and UDAC) outperform the Homogeneous-based LSAs. The poor performance achieved by the homogeneous-based LSAs can be attributed to the random selection of leader by the homogeneous-based LSAs

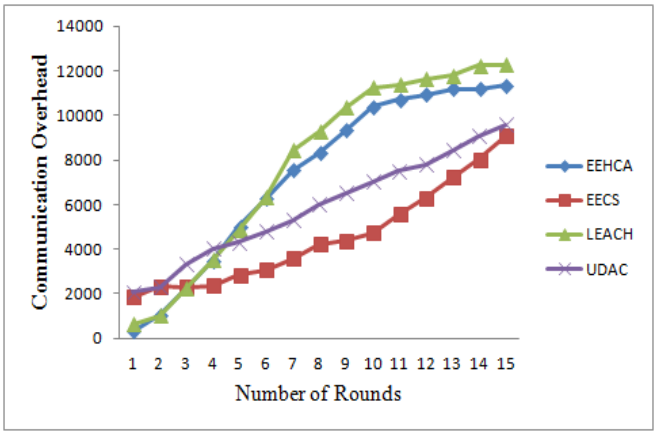



**Figure 4a: Effect of Rounds on Communication Overhead**

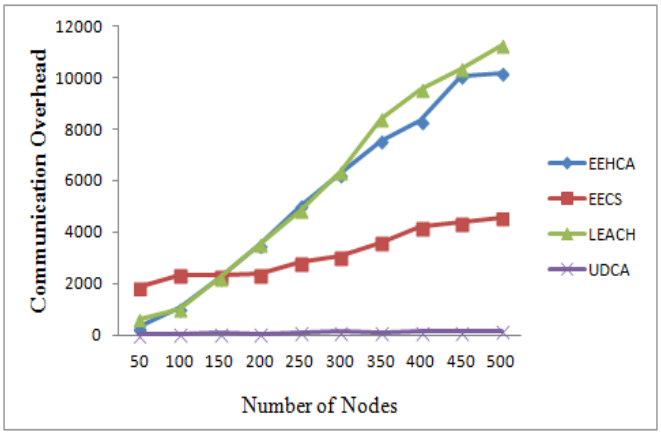



**Figure 4b: Effect of Network Sizes on Communication Overhead**

Based on random selection of leader by homogeneous-based LSAs, nodes with low remaining energy can be selected as a leader. Hence, it compromises the entire network reliability and increases the network communication overhead cost, since a new leader has to be selected every time the current leader fails. In Figure 4b, it can be observed that the Heterogeneous-based (EECS and UDAC) outperform the Homogeneous-based LSAs. It can also be observed that when the numbers of nodes are 50 and 100, the communication overhead for homogeneous-based algorithms were lower than that of EECS, however, as the number of nodes increases, the homogeneous algorithms began to incur higher communication overhead than other LSAs considered. The poor performance of the homogeneous-based LSAs can be attributed to the random selection of its leader.

### B. *Leader Selection Delay*

Leader selection delay measures the time taken for Leader Selection Algorithms to successfully select one node as an MKD leader. Figures 5a and 5b depict the Leader Selection Delay for both the heterogeneous and homogeneous energy-based LSAs when respectively subjected to different leader selection rounds and various network sizes. This metric was measured in seconds (s). In Figure 5a, it can be observed that both the LEACH and EEHCA of Homogeneous-based leader selection algorithms outperform Heterogeneous-based algorithms. The better

performance average delay value achieved by the homogeneous-based leader selection algorithm can be attributed to the fact that the leader are been selected at random, which does not require any specific process as against that of Heterogeneous-based, which consist of three different phases for selecting a leader. Hence, each of these phases introduces some delay, which accumulates and lead to the poor and inconsistent behaviour of the heterogeneous-based leader selection algorithms. In Figure 5b, it can be observed that the EECS of heterogeneous-based leader selection algorithm outperform the other three algorithms (UDCA, LEACH and EEHCA) considered. Whilst UDCA and LEACH algorithms behave exactly the same way, EEHCA also behave almost exactly as these two algorithms. The better performance achieved by EECS leader selection algorithm in terms of low average delay value can be attributed to the optimal cluster head selection behaviour of the four energy-based LSAs considered. EECS always create more clusters as the need arise, which reduces the burden on the cluster head while transmitting the control message to the base station. However, the remaining three leader selection algorithms (UDCA, LEACH and EEHCA) always optimize the number of clusters by creating a fewer number of clusters and this leads to overburden on the cluster head through congestion, which eventually led to the high delay value incur by the three algorithms.

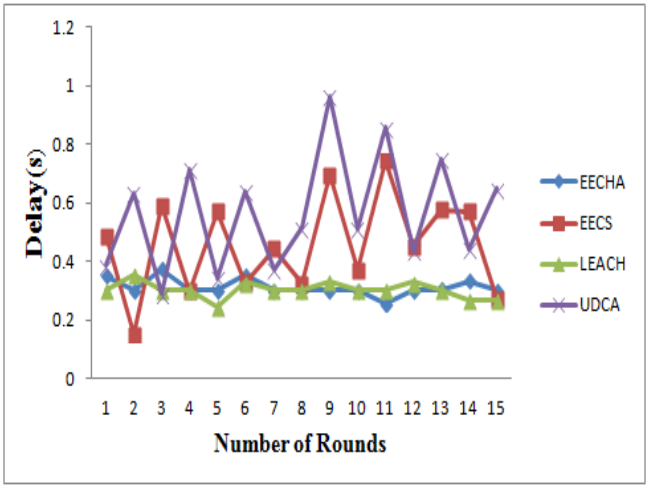



**Figure 5a: Effect of Rounds on Leader Selection Delay**

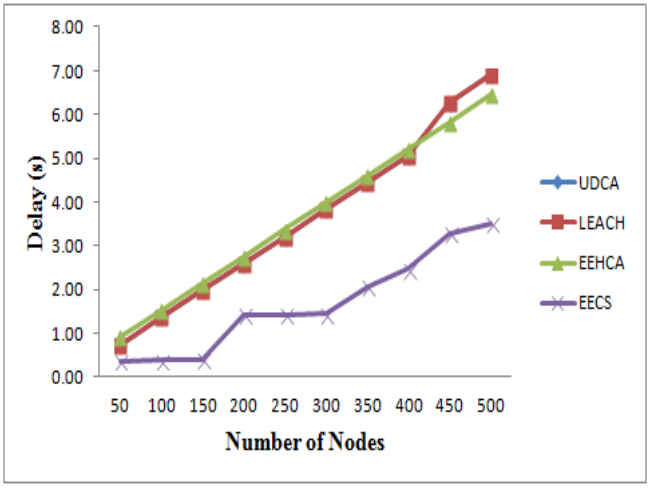



**Figure 5b: Effect of Network Sizes on Leader Selection Delay**

*C. Energy Consumption Rate*

Figures 6a and 6b depict the results of the network energy consumption rate for both the heterogeneous and homogeneous energy-based LSAs when respectively subjected to different leader selection rounds and various network sizes. In Figure 6a, it can be observed that the UDCA of Heterogeneous-based leader selection algorithms outperform other Energy-based leader selection algorithms considered. Although, there are inconsistencies in the performance of the LEACH and EECHA, but EECS is the least performing algorithm among all the LSAs considered in terms of energy consumption rate. The high energy consumption rate of the EECS can be attributed to its poor clustering optimization. In EECS leader selection process, it creates high number of clusters, which led to high number of cluster heads and this in turn reduce the energy efficiency of the network. The LEACH and EECHA (homogeneous-based leader selection) inconsistency behaviour can be attributed to their stochastic selection of leader. Based on stochastic selection of leader by homogeneous-based leader selection algorithm, nodes with low remaining energy can be selected as a leader. Hence, it dissipates the entire network energy faster, since a new leader has to be selected every time the current leader fails. In Figure 6b, it can be observed that both the UDCA and LEACH leader selection algorithm outperforms other Energy-based leader selection algorithms considered. Also, it can be observed that they both have similar behaviour in most scenarios. Whilst EECS is the least performing algorithm among all the LSAs considered in terms of energy consumption rate. The poor performance of EECS algorithm in terms of energy consumption rate can be attributed to its network clustering process. In EECS leader selection process, the number of clusters normally increases as the network grows and each cluster's has cluster head. Due to the continuous increase in the number of cluster heads and the communication among those cluster heads and base station, more energy are being consumed, which in turn reduce the energy efficiency of the network. The low energy consumption rate achieved by both the UDCA and LEACH algorithms can be attributed to their clustering optimization process. In this clustering optimization, few clusters are normally created with few cluster heads; hence, the lower the number of clusters, the lower the cluster heads and the lower the cluster heads, the lower the energy consumption rate.

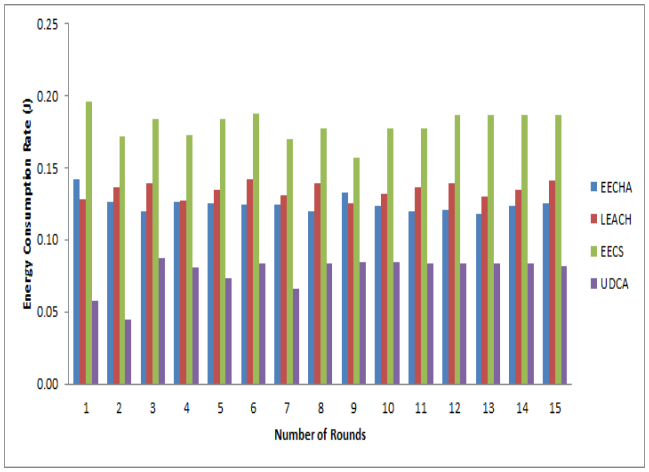



**Figure 6a: Effect of Rounds on Network Energy Consumption Rate**

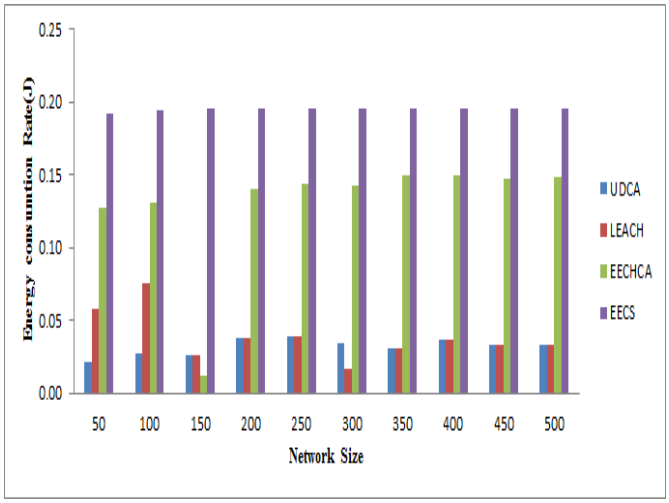



**Figure 6b: Effect of Network Sizes on Network Energy Consumption Rate**

## VII. CONCLUSION

This study considered the scenario where the current Mesh Key Distributor (MKD) becomes unavailable for use in a Wireless Mesh Network. The unavailability of the MKD breaks the node authentication mechanism defined in the IEEE 802.11s standard for mesh networking. The selection of a new MKD can be automated by adopting Leader Selection Algorithms (LSAs) from the domain of Wireless Sensor Networks.

In this paper, the performance of four energy-based LSAs was evaluated to determine their suitability for MKD selection in Wireless Mesh Networks. The LSAs were evaluated using NS2 and the evaluation showed that the heterogeneous energy-based LSAs (EECS and UDAC) outperformed the homogeneous energy-based LSAs (LEACH and EEHCA) for communication overhead cost and the energy consumption rate. The homogeneous energy-based LSAs outperform the heterogeneous energy-based LSAs in terms of Leader Selection Delay. Although there are few variations in the behaviour of both the heterogeneous and homogeneous scenarios, homogeneous algorithms are not recommended for MKD selection in WMNs, due to its stochastic selection of a leader which can lead to compromised network reliability. Hence, based on the results of this study, we argue that the heterogeneous based LSAs of wireless sensor network can be adopted for MKD selection in WMNs. We intend to repeat the experiments on the WMN testbed being planned on the University campus.

## REFERENCES

[1] Akyildiz I.F and Wang. X, "Wireless Mesh Networks," Wiley: Chichester, 2009.

[2] Akyildiz I.F, A Survey on Wireless Mesh Networks, *IEEE Radio Communications*, September 2005, pp.525-530.

[3] Salem N.B. and Hubaux J.P, Securing Wireless Mesh Networks, in IEEE Wireless Communication, Volume 13, Issue 2, April 2006 pp. 50 - 55.

[4] Allen W., Martin A, Rangarajan A." Designing and deploying a rural ad hoc community Mesh Network Testbed. Proc. IEEE Conf. on local Computer Networks; November 2005.

[5] Lundgren H, Ramachandran K, Belding-Royer E, Almeroth K, Jardosh A. "Experiences from the Design, Deployment and usage of the UCSB Meshnet Testbed". IEEE Wireless Communications 2006: 13(2):18-29.

[6] Camp J, Robinson J, Steger C, Knightly E. Measurement Driven Deployment of a Two-tier Urban Mesh Access Network. 4th Proc. Intl Conference on Mobile systems, applications and services; June, 2006.

[7] Guido R. Hiertz A," IEEE 802.11s Wireless Mesh Networks*", Submitted to 802.11s Wireless architecture Sub Group*, November, 2005.

[8] Kuhlman. D, Moriarty.R, Braskich. T, Emeott.S, and Tripunitara. M, A Proof of Security of a Mesh Security Architecture," Cryptology ePrint Archive, Report 2007/364, 2007.

[9] Zengwei Z., Zhaohui W., and Huaizhong L., "An Event-Driven Clustering Routing Algorithm for Wireless Sensor Networks". Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems. Sendai, Japan, Sep. 28 - Oct. 2, 2004.

[10] Loscrì.V, Morabito.G and Marano.S "A Two-Levels Hierarchy for Low-Energy Adaptive Clustering Hierarchy". IEEE Vehicular Technology Conf. Sept. 25-28, 2005.

[11] Otgonchimeg B, Kwon. Y, IMS and IDC, "EECED: Energy Efficient Clustering Algorithm for Event-Driven Wireless sensor Networks" Fifth Intl Joint conf., Aug. 25-27, 2009.

[12] Wang P, Dai R. and Akyildiz I.F, *"Collaborative Data Compression Using Clustered Source Coding for Wireless Multimedia Sensor Networks,"* Proc. of INFOCOM 2010, March 2010.

[13] Baccarelli. E, Chlamtac I., Gumaste A., and Szabo C., "Broadband Wireless Access Networks: a Roadmap on Emerging Trends and Standards," *Broadband Services: Business Models and Technologies for Community Networks*, Eds., Wiley, Mar. 2005.

[14] Salem N.B, Hubaux J.P, "Securing Wireless Mesh Networks", in IEEE Wireless Communication, Volume 13, Issue 2, April 2006 pp. 50 - 55.

[15] Wang X, and Lim A.O, "*IEEE 802.11s Wireless Mesh Networks: Frameworks and Challenges*," Ad Hoc Networks 6(6), 970-984 (2008)

# Bandwidth reduction using correlated source compression for smart grid meters with feedback

Reevana Balmahoon and Ling Cheng
Department of Electrical and Information Engineering
University of the Witwatersrand, 1 Jan Smuts Avenue, Braamfontein 2000
Tel: +27 73 0090536, Fax: +27 86 5857443
email: 763088@students.wits.ac.za, ling.cheng@wits.ac.za

Abstract – **Correlated source compression has the advantage of decreased bandwidth as shorter messages are sent across a channel. Here, we develop a model for correlated sources across a wiretap channel and investigate feedback for the model. Further, we present a theorem specifying the upper bound for transmission rate when feedback is employed. This work shows that a reduction in bandwidth is achieved when correlation and feedback is implemented. Correlated meters for a smart grid is the considered application; we show that smart grid meters can be considered as correlated sources and that this work has room for practical use.**

**Index Terms – Correlated Sources, Slepian-Wolf, Feedback.**

## I. INTRODUCTION

Correlated source coding incorporates the lossless compression of two or more correlated data streams. Here, we use smart grid meters as a correlated source system. The smart grid is a type of electrical grid that functions to predict and intelligently respond to the behaviour of the users connected to it [1]. It is capable of making the electricity grid work more efficiently, securely and reliably through bidirectional flows of power and communication [2]. The two-way communication may be implemented using AMR (automatic meter reading), where the smart meter is a key component. These meters take readings of electricity consumption. An investigation has been done to prove that these meter readings generally have correlation.

A concept that is related when looking at correlated sources is that of side information. The side information is traditionally used to help the decoder or a third party retrieve the transmitted message, for e.g. a scheme presented by Villard and Piantanida [3]. This correlated side information may be considered as a source that is correlated to another source in the network. This concept of side information has been earlier analysed by Yang *et al.* [4] where correlated side information is presented to generalise the decoding procedure.

The techniques for source compression mentioned in this paper aim to achieve the Slepian-Wolf bound. The Slepian-Wolf theorem gives a bound on the channel rates required for transmission so that the receiver is able to decode the transmitted messages with minimal error [5] .The rate bounds that relate to the Slepian-Wolf theorem are presented herein. According to Kurkoski and Wolf [5], an important aspect of the Slepian-Wolf theorem is that the encoders can achieve better compression rates by exploiting the correlation in the data streams. The result is that Slepian-Wolf coding can achieve the same compression rate as an optimal single encoder that has all correlated data streams as inputs [5].

Correlation between sources contributes to a security risk and source compression is a method that has come about to help to increase the security. A method for correlated sources is the use of raptor codes, which have been researched in works by Cheng and Ferreira [6] and Cheng *et al.* [7]. These schemes operate in a different network layer to the scheme presented in this paper and are hence not applicable here. We look at correlated sources by the use of meters (similar to sensors that have been used in [8]), where the correlation results from certain protocols that are considered as predetermined information. An interesting study of correlated sources has been done by Gunduz *et al.* [9].

The compression also provides a more secure system, as the message uncertainty is increased when the message is compressed. This is of concern when an adversary (i.e. an eavesdropper) is observing the transmission over the communication links. In practical communication systems links are prone to

eavesdropping and as such this work incorporates wiretapped channels, more specifically the Wiretap Channel II. The mathematical model for this Wiretap Channel is given by Rouayheb *et al.* [10]. A key characteristic of this wiretap channel of type II is that it is error-free. In other work by the authors [11], a correlated source model across a wiretap channel II has been developed and analyzed in terms of information leakage. In this work we look at employing feedback to a correlated source model.

There has been interesting work on feedback, ranging from feedback for multiple sources [12] [13] to multiterminal access [4] [14] where feedback for single and multiple encoder cases are addressed. Yang *et al.* [4] present a scenario that is related to this work, which is feedback for a source that has correlated side information. Here, we consider feedback as an aid for determining the outstanding information needed by the decoder to retrieve the transmitted message.

In this paper, the correlation between meters is shown using a real case scenario and a model incorporating correlated sources and feedback is presented. Section II provides details on correlated sources and presents the correlation between meter readings taken from a smart grid meter. Section III gives an overview of the Slepian-Wolf theorem and its implication for correlated sources. Section IV details a model incorporating feedback for two correlated sources. The paper is concluded in Section V.

## II. CORRELATED SOURCES

The messages from correlated sources have some similarity (measure of correlation) between them. Encoding methods where each of the correlated streams are encoded separately and the compressed data from all these encoders are jointly decoded by a single decoder as shown for two correlated streams (and depicted in Figure 1) have been established in investigations mentioned in this paper.

Below we show that smart grid meters have the property of correlation and hence the coding methods that apply to correlated sources will also apply to smart grid meters.

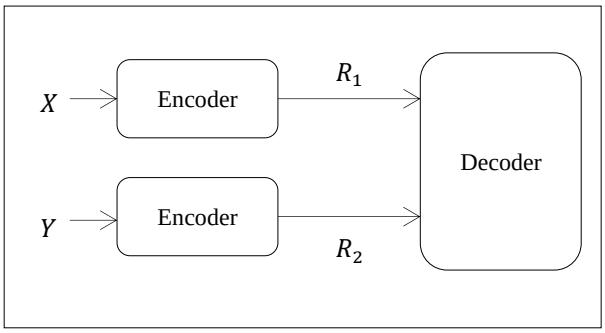



Figure 1 - Correlated data streams

Readings were taken from two real time meters for a smart grid application and the correlation between sources gave rise to the following diagram:

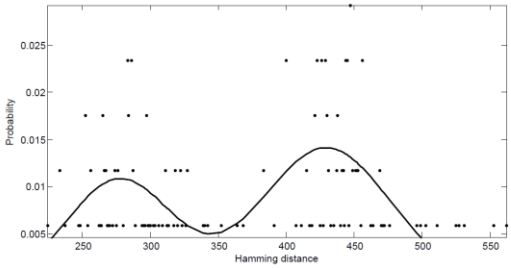



Figure 2 - Diagram showing Gaussian fit for correlation between readings from two meters

The correlation between the sources in the above figure is evident; they have a Gaussian distribution. The curve, $f\ (x)$ that fits this Gaussian distribution is:

$$f\ (x) = 0.0176\, e^{-\frac{(x-429.3)^2}{65.95^2}} + 0.01344\, e^{-\frac{(x-277.2)^2}{53.98^2}} \quad (1)$$

The Hamming distances between readings from the sources are indicated on the $x$ axis. Their probability, given by: $P = \frac{N}{T}$, where $P$ is the probability, $N$ is the number of occurrences and $T$ is the total number of readings, is given on the $y$ axis.

The correlation is an advantage because a compressed form of the information to send needs to be transmitted across the channel; this comes from the fact that it is not necessary to transmit the correlated information. However, it poses a threat when the correlated bits are known as an eavesdropper is able to gain additional information about the source.

The Slepian-Wolf theorem for correlated sources $X$ and $Y$, may be represented diagrammatically as follows:

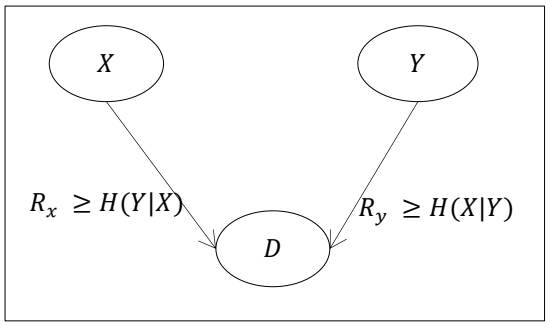



Figure 3 - Diagram showing rate allocation for correlated sources X and Y, with destination D

$$R_x + R_y \geq H(Y|X) + H(X|Y) \geq H(X,Y) \quad (2)$$

where $R_x$ denotes rate allocation for source $X$, $R_y$ indicates rate allocation for source $Y$ and $H(Y|X)$ and $H(X|Y)$ represent the conditional entropy of $Y$ given $X$ and conditional entropy of $X$ given $Y$ respectively.

From equation (2), the overall rate allocation $(R_x + R_y)$ for the correlated sources is thus $H(X,Y)$. This means that $X$ and $Y$ need to have a rate allocation of $H(X,Y)$ to ensure that the received messages can be decoded correctly.

## III. FEEDBACK MODEL

We consider the feedback source network depicted in Figure 4. The independent, identically distributed (i.i.d.) sources $X$ and $Y$ are mutually correlated random variables represented by source alphabets with cardinality greater than or equal to 2. They transmit messages (in the form of syndromes) to the receiver along wiretapped links and the receiver transmits some information to the transmitters (i.e. the feedback). The scenario is a typical two way Gaussian channel where the communication channel can be reduced to two independent Gaussian links. Here, we have developed a generalised model, based on Yamamoto's model [15]. We assume that the links may be wiretapped by an eavesdropper.

Assume that the sources are encoded into two syndromes ($T_{X,k}$ and $T_{Y,k}$), where $T_{X,k}$ and $T_{Y,k}$ are $X$'s and $Y$'s syndrome at time $k$ respectively. We can write $T_{X,k} = F(X^K, T_{X,k-1})$ and $T_{Y,k} = F(Y^K, T_{Y,k-1})$. Here, $T_{X,k}$ and $T_{Y,k}$ are characterised by $(V_{X,k}, V_{CX,k}) = F'(T_{X,k})$ and $(V_{Y,k}, V_{CY,k}) = F'(T_{Y,k})$, where $V_{X,k}$ and $V_{CX,k}$ is $X$'s private and common information at time $k$ respectively and $V_{Y,k}$ and $V_{CY,k}$ is $Y$'s private and common information at time $j$ respectively. Here, $V_{X,k}$ and $V_{Y,k}$ represent the private information of sources $X^K$ and $Y^K$ respectively and $V_{CX,k}$ and $V_{CY,k}$ represent the common information between $X^K$ and $Y^K$ generated by $X^K$ and $Y^K$ respectively, at time $k$. Here, we represent the codeword set for $X$ and $Y$ similarly to Yamamoto's [15], where the common and private information position in the sequence associated with $X$ are specified by integers $l$ and $i$ respectively. Similarly, for $Y$ the common and private information positions are specified by $m$ and $j$ respectively. In order to retrieve $X$, both $i$ and $l$ are required and similarly for $Y$, both $j$ and $m$ are required.

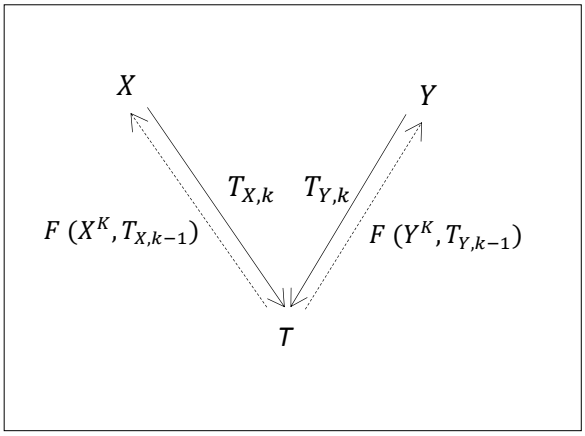



Figure 4 - Two correlated source model with feedback

The decoder determines $X$ and $Y$ after receiving $T_{X,j}$ and $T_{Y,j}$. In order to decode a transmitted message, a source's private information and both common information portions are necessary. If there is insufficient information for decoding then the feedback method enables the decoder to request for the necessary information to find the private/common information required.

We assume that there is a possibility that the Slepian-Wolf bound is not met; hence there may be more than one possible codeword. The purpose of the bit transmitted as feedback is to indicate to the encoders whether there had been one or more than one codewords retrieved and to request for the outstanding information.

We further assume that the correlation is initially underestimated. This means that the actual model in reality is higher in entropy than the original model with which we begin our algorithm. Initially $X$ and $Y$ have joint entropy $H(X,Y)$, and $H(\hat{X},\hat{Y}) > H(X,Y)$ where $\hat{X}$ and $\hat{Y}$ are the actual sources and $X$ and $Y$ are sources of the original model.

We now put forth our algorithm for the feedback source network.

*Algorithm:*

*Step 1:* Let $k$ be a counter initialised to 0.

*Step 2:*
*At X:* If $k = 0$, $X$ sends the Elias codeword of $i$ (i.e. $b_i = \lfloor \log i \rfloor + 1$) to $T$. Upon receiving the Elias codeword of $i$, the decoder knows that $i$ is an integer in the set $I_{i,k} = \{2^{b_i-1}, 2^{b_i-1}+1, \dots, 2^{b_i-1}-1\}$.
If $X$ receives 01, $X$ partitions $I_{i,k-1}$ into $2^{\lfloor \sqrt{l} \rfloor}$ non-overlapping subsets of size $|I_{i,k}| = \left\lceil \frac{|I_{i,k-1}|}{2^{\lfloor \sqrt{p} \rfloor}} \right\rceil$. This follows the same principle employed by Raymond [4]. Then, $X$ transmits the partition index where $i$ is contained, to $T$.

*At Y:*
If $k = 0$, the encoder sends the Elias codeword of $j$ (i.e. $b_j = \lfloor \log j \rfloor + 1$) to $T$. Upon receiving the Elias codeword of $j$, the decoder knows that $j$ is an integer in the set $I_{j,k} = \{2^{b_j-1}, 2^{b_j-1}+1, \dots, 2^{b_j-1}-1\}$.
If $Y$ receives 01, $Y$ partitions $I_{j,k-1}$ into $2^{\lfloor \sqrt{l} \rfloor}$ non-overlapping subsets of size $|I_{j,k}| = \left\lceil \frac{|I_{j,k-1}|}{2^{\lfloor \sqrt{p} \rfloor}} \right\rceil$. Then, $Y$ transmits the partition index where $j$ is present, to $T$.

*At the decoder side:* If there is a unique index in $I_{i,k}$ and $I_{j,k}$ then the private portions of $X$ and $Y$ are determined and $T$ transmits 11 to both $X$ and $Y$.

If there is no unique element in $I_{i,k}$, the decoder sends 01 to $X$ and increments $k$ by 1. If there is no unique element in $I_{j,k}$, the decoder sends 01 to $Y$ and increments $k$ by 1.

*Step 3:*
*At X:* If X receives 01, repeat step 2 (at X) until there is a unique element in $I_{i,k}$.
If X receives 11, the encoder sends the Elias codeword of $i$ (i.e. $b_l = \lfloor \log l \rfloor + 1$) to $T$. Let $k = 0$. Upon receiving the Elias codeword of $l$, the decoder knows that $l$ is an integer in the set $I_{l,k} = \{2^{b_l-1}, 2^{b_l-1}+1, \dots, 2^{b_l-1}-1\}$.

*At Y:* If Y receives 01, repeat step 2 (at Y) until there is a unique element in $I_{j,k}$.

*At Decoder*: If there is a unique index in $I_{l,k}$ then the common portion of $X$ is determined and $T$ transmits 00 to $X$. At this point the decoder has $i, l$ and $m$ and can fully determine $\hat{X}$, which is defined by $i$ and $l$. If while determining $\hat{Y}$ there are many possibilities, then the decoder sends 10 to $Y$ and increments $k$ by 1.

If there is no unique element in $I_{l,k}$, the decoder sends 10 to $X$ and increments $k$ by 1.

*Step 4:*
*At X:* If $X$ receives 10, $X$ partitions $I_{l,k-1}$ into $2^{\lfloor \sqrt{l} \rfloor}$ non-overlapping subsets of size $|I_{l,k}| = \left\lceil \frac{|I_{l,k-1}|}{2^{\lfloor \sqrt{p} \rfloor}} \right\rceil$. Then, $X$ transmits the partition index where $l$ is present, to $T$. This partition and step 3 is repeated until there is a unique element in $I_{l,k}$.
*At Y:* If Y received 11, Y sends the Elias codeword of $m$ (i.e. $b_m = \lfloor \log m \rfloor + 1$) to $T$. Upon receiving the Elias codeword of $m$, the decoder knows that $l$ is an integer in the set $I_{m,k} = \{2^{b_m-1}, 2^{b_m-1}+1, \dots, 2^{b_m-1}-1\}$.

*At the decoder:* If there is a unique index in $I_{l,k}$ then $T$ transmits 00 to $X$. If there is a unique index in $I_{m,k}$ then $T$ transmits 00 to $Y$. If there is no unique index in $I_{m,k}$, the decoder sends 10 to $Y$ and increments $k$ by 1.

*Step 5:*
*At Y*: If Y receives 10, $Y$ partitions $I_{m,k-1}$ into $2^{\lfloor \sqrt{l} \rfloor}$ non-overlapping subsets of size $|I_{m,k}| = \left\lceil \frac{|I_{m,k-1}|}{2^{\lfloor \sqrt{p} \rfloor}} \right\rceil$. Then, $Y$ transmits the partition index where $m$ is present, to $T$. This step and step 4 is repeated until there is a unique element in $I_{m,k}$. Thus, the decoder now has j and m and can determine $\hat{Y}$.
It can be seen from the algorithm above that the syndromes for each time instant are dependent on the previous partition and the feedback.

$$T_{X,k} = F(I_{i,k-1}, F_{TX,k-1}, I_{l,k-1}) \quad (2)$$

$$T_{Y,k} = F(I_{j,k-1}, F_{TY,k-1}, I_{m,k-1}) \quad (3)$$

where $F_{TX,k-1}$ is the feedback transmitted from $T$ to $X$ at time $k-1$, and similarly for $Y$, $F_{TY,k-1}$ is the feedback transmitted from $T$ to $Y$ at time $k-1$. Further the sources need not know the rate pair $(R_X, R_Y)$ beforehand, and the decoder is able to select the rates with which the encoders transmit information after the first iteration.

We now present an example to make the algorithm clearer.

Assume the indexes represented by $i, l, j$ and $m$ are 3, 4, 2 and 2 respectively. Following the steps as explained in the algorithm above we have:

*Step 1:* $k = 0$

*Step 2:*
*At X:* $X$ sends 2 (the Elias codeword of $i = 3$) to T. This means that $I_{i,0} = \{2, 3\}$
*At Y:* $Y$ sends 3 (the Elias codeword of $j = 4$) to T. This means that $I_{j,0} = \{4, 5, 6, 7\}$

*At the decoder:* There is more than one element in $I_{i,0}$ and $I_{j,0}$ so the decoder sends 01 to $X$ and 01 to $Y$.

*Step 3:*
*At X*, 01 is received and $I_{i,0}$ is partitioned into {2},{3}. These partitions are represented by 0 and 1 respectively. The partition that $i$ lies in is 0, hence a 0 is transmitted from $X$ to $T$. This need not be repeated as the unique element in $I_{i,0}$ has been identified. The decoder sends 11 to X.

*At Y*, 01 is also received and $I_{j,0}$ is partitioned into $\{4, 6\}, \{5,7\}$ (i.e. $I_{j,1}$). These partitions are represented by 0 and 1 respectively. The partition that $j$ lies in is 0, hence a 0 is transmitted $Y$ to $T$. This is repeated as the decoder is now presented with 2 elements in $I_{j,1}$. The decoder sends a 01 to $Y$ as more information about the private information is required. Then $Y$ partitions $I_{j,1}$ into the new $I_{j,2}$ to give $\{4\}, \{6\}$. The partition that $j$ lies in is 0, hence a 0 is transmitted from $Y$ to $T$. The decoder sends 11 to Y.

The decoders have both received 11 at this stage, and this means the common information can now be transmitted. $X$ sends 2 (the Elias codeword of $l = 2$) to T. This means that $I_{l,0} = \{2, 3\}$. At the decoder there is more than one element in $I_{l,0}$ hence the decoder sends 10 to $X$.

*Step 4:*
*At X*: Since 10 has been sent to $X$, $I_{l,0}$ is partitioned into $\{2\}, \{3\}$ and the partition that $l$ lies in is 0, hence $I_{l,1} = \{2\}$ and a 0 is transmitted from $X$ to $T$. Since there is a unique element in $I_{l,1}$ the decoder will send 00 to $X$.

The decoder has $i, j$ and $l$ so the indices for $X$ are found and the codeword may be retrieved however $m$ is required in order to find $Y$.

*At Y:* $Y$ sends 2 (the Elias codeword of $m = 2$) to $T$. This means that $I_{m,0} = \{2, 3\}$. At the decoder there is more than one element in $I_{m,0}$ hence the decoder sends 10 to $Y$.

*Step 5*:
*At Y*: Since 10 has been sent to $Y$, $I_{m,0}$ is partitioned into $\{2\}, \{3\}$ and the partition that $m$ lies in is 0, hence $I_{m,1} = \{2\}$ and a 0 is transmitted from $Y$ to $T$. Since there is a unique element in $I_{m,1}$ the decoder will send 00 to $X$.

The decoder now has access to $j$ and $m$, and can thus retrieve the indices for the codeword for $Y$. The result is the correct $\hat{X}$ and $\hat{Y}$ for the actual model.

As a result, to determine the upper bound for the rate required when feedback is used we have developed the following theorem:

*Theorem 1: For two correlated sources X and Y with feedback as described in the above algorithm, the transmission rate is upper bounded by the joint entropy and mutual information between the sources:*

$$R < H(\hat{X}, \hat{Y}) + I(X; Y)$$

where $R$ is the rate required for the transmission of messages.

This theorem presents the rate for the worst case scenario of transmitting $m$ entirely to $T$. It can be proven in a straight forward manner. Here, the rate required for correct decoding of the sources is $H(X, Y)$, which conforms to the Slepian-Wolf theorem. The addition for this case is a request for $m$, where we can see from step 5 that all of $m$ is transmitted and the rate required is $I(\hat{X}; \hat{Y})$, which represents the mutual information of the original, underestimated model. Here, we thus need to transmit a maximum of the joint entropy of the actual sources and the mutual information of the underestimated sources for correct decoding.

This algorithm shows how to transmit blocks of information between two correlated sources and a destination node. From step 4 and 5, we can see that this algorithm presents the worst case scenario if $Y$'s common information portion is outstanding. This portion contributes to the decrease in bandwidth as we only transmit $m$ if necessary. It is possible to send across a portion of $m$ in order to limit the bandwidth, and this can be achieved by the decoder guessing the

codeword (this has not been implemented but is considered as future work). The bandwidth used in a network is largely determined by the packet size of a message. If a portion of $m$ is sent as a response to the feedback then the packet size is reduced as compared to sending the entire message across the link. This method will reduce the bandwidth required to transmit a message as the entire message will not have to be resent to the decoder.

Correlation is an advantage because the compression allows for less information to be transmitted across the channel, and this results in a reduction in the bandwidth required to transmit and receive messages because a syndrome (compressed form of the original message) is sent across the communication links instead of the original message. Correlation combined with sending a portion of $m$ across the transmission will reduce the bandwidth required as less information is transmitted than if no correlation and feedback was used.

## IV. CONCLUSION

This paper investigated feedback for a wiretapped channel that transmits information from two correlated sources to a receiver. A theorem showing the upper bound required for the proposed algorithm has been presented. An investigation involving meter readings for a smart grid system show that a correlated sources approach has practical use. The use of feedback enables the decoder to request only the remaining bits necessary to determine the correct message and the source does not need to retransmit the entire message. This method may be used in smart grid systems to decrease the bandwidth involved in transmitting meter readings to a control station.

## V. REFERENCES


[1] PG Del and C Landi, "Real-time smart meter with embedded web server capability ," in *IEEE Internation Conference on Instrumentation and Measurement Technology*, 2012, pp. 682 - 687.

[2] J Xia and Y Wang, "Secure Key Distribution for the Smart Grid," *IEEE Transactions on Smart Grid*, vol. 3, no. 3, pp. 1437 - 1443, September 2012.

[3] P Piantanida and J Villard, "Secure Multiterminal Source Coding With Side Information at the Eavesdropper," *IEEE Transactions on Information Theory*, vol. 59, no. 6, pp. 3668 - 3692, June 2013.

[4] E Yang, D Tomohiko Uyematsu, and R.W Yeung, "Universal Multiterminal Source Coding Algorithms With Asymptotically Zero Feedback: Fixed Database Case," *IEEE Transactions on Information Theory*, vol. 54, no. 12, pp. 5575 - 5590, December 2008.

[5] B.M Kurkoski and J.K Wolf. (2008) Slepian Wolf Coding. [Online]. http://www.scholarpedia.org/article/Slepian-Wolf_coding

[6] L Cheng and H Ferreira, "Time-Diversity Permutation Coding Scheme for Narrow-Band Power-Line Channels," in *IEEE International Symposium on Power Line Communications and Its Applications*, 2012, pp. 120 - 125.

[7] L Cheng, T Swart, and H Ferreira, "Adaptive Rateless Permutation Coding Scheme for OFDM-based PLC," in *IEEE 17th International Symposium on Power Line Communications and Its Applications*, 2013, pp. 242 - 246.

[8] J Barros and S Servetto, "Network Information Flow With Correlated Sources," *IEEE Transaction on Information Theory*, vol. 52, no. 1, pp. 155 - 162, January 2006.

[9] D Gunduz, E Erkip, A Goldsmith, and H Poor, "Source and Channel Coding for Correlated Sources Over Multiuser Channels," *IEEE Transactions on Information Theory*, vol. 55, no. 9, pp. 3927 - 3944, September 2009.

[10] S Rouayheb, E Soljanin, and A Sprintson, "Secure Network Coding for Wiretap Networks of Type II," *IEEE Transactions on Information Theory*, vol. 58, no. 3, pp. 1361 – 1371, March 2012.

[11] R Balmahoon and L Cheng, "Information Leakage of Correlated Source Coded Sequences over Wiretap Channel," *eprint arXiv:1401.6264*, January 2014.

[12] L Ong and M Motani, "Coding Strategies for Multiple-Access Channels With Feedback and Correlated Sources," *IEEE Transactions on Information Theory*, vol. 53, no. 10, pp. 3476 - 3497, October 2007.

[13] A Jain, D Gunduz, S.R Kulkarni, and H.V Poor, "Energy-Distortion Tradeoff with Multiple Sources and Feedback ," in *Information Theory and Applications Workshop*, 2010, pp. 1 - 5.

[14] A Abrardo, G Ferrari, M Martalo, and F Perna, "Joint Channel Decoding with Feedback Power Control in Sensor Networks with Correlated Sources," in *International Symposium on Wireless Communication Systems*, 2009, pp. 274 - 278.

[15] H Yamamoto, "Coding Theorems for Shannon's Cipher System with Correlated Source Outputs, and Common Information," *IEEE Transactions on Information Theory*, vol. 40, no. 1, pp. 85 - 95, January 1994.

[16] D Gunduz and E Erkip, "Lossless Transmission of Correlated Sources over a Multiple Access Channel with Side Information," in *IEEE Data Compression Conference*, 2007, pp. 83 - 92.

# Energy Minimization in WSNs: Empirical Study of Multicast Incremental Power Algorithm

Adeyemi Abel Ajibesin*; Neco Ventura*; Alexandru Murgu* and H. Anthony Chan†
Department of Electrical Engineering
University of Cape Town*, South Africa, Rondebosch 7701
Tel: +27 78 1295025, Fax: +27 21 6503465
And Huawei Technologies†, Plano, Texas, USA
email: {ajbade001, neco.ventura, alexandru.murgu}@uct.ac.za*; h.a.chan@ieee.org†}

**Abstract- This paper studied the existing multicast incremental power algorithm (MIP) and proposes a new approach considering the data envelopment analysis (DEA) methodology to further reduce the multicast energy in wireless sensor networks (WSNs). In order to achieve this mission, an empirical model based on input-orientation with Banker, Chanes and Cooper (BCC) approach is developed. This research has shown how to evaluate efficiency ratings of WSNs and identify the inefficient WSNs with their magnitude at which they deviate from the best practice frontier. Furthermore, the results show how the inefficient WSNs can reduce their input energy so that they also become efficient. Thus our model is able to further reduce the multicast energy by 5% over MIP method if all the WSNs sampled were operating efficiently.**

**Index Terms— Performance evaluation, multicasting, energy efficiency, WSNs, network modeling and analysis**

## I. INTRODUCTION

Multicasting allows users to send the same information from one or more WSN nodes to a selected subset of other nodes in a consumer communication network. In other words, multicasting reduces the transmission overhead it takes for all the WSN nodes in the subset to receive the information [1]. In WSN, transmission energy is one of the major challenges; therefore its efficient usage should be considered [2]. Many works have addressed the problem of minimum energy multicast but most of these works have failed to address the relative efficiency evaluation [3], [4]. Also, most of the performance evaluations were carried out using single-factor metric, making the output results inappropriate for efficiency analysis [5]. In this work, a new approach that explores the real efficiency evaluation of WSNs resources (e. g. energy) is proposed. This approach extends the existing simulation method that attempt to minimize energy consumptions by the nodes in WSN. The empirical method, which is employed in this work, is considered for the data envelopment analysis (DEA) methodology with input-orientation approach. Also, this empirical model, which we developed, is extended beyond the basic Chanes, Cooper and Rhodes (CCR) DEA model and therefore appropriate for energy minimization problem. Thus the proposed model is implemented for input reduction with a focus on energy minimization in WSNs.

In the literature, the essence of minimum energy multicast is to provide solution to the problem of high energy transmission over the network. This was achieved using the minimum energy multicast scheme, but minimum energy multicast problem has been proved to be NP-complete [6]. However, alternative solutions using polynomial-time based heuristics approach have been considered [7], [8]. One of these solutions known as Multicast Incremental Power (MIP) algorithm has been widely applied to wireless networks with outstanding performance [9], [10].

We study the MIP algorithm and investigate its performance for real efficiency evaluation. This is necessary because MIP evaluation is based on effective performance. Our aim is to develop alternative method based upon BCC model to minimize the multicast energy by the WSN nodes. Using the input-oriented BCC model approach, the performance of the WSNs is expected to be improved by comparing each of the WSNs with the best practices WSNs. We obtained the data set for our model from the traditional simulation of WSNs, which is an engineering approach to performance evaluation. By contrast, the DEA methodology, which we adopted, relies on the linear programming technique for optimization [11]. In addition, the DEA method is different from other methods because it is a powerful technique for performance evaluation based on optimal solution of relative efficiency rather than averages that measures effective performance [12]. In today's technology, with very fast development, no network can afford to be an average performance especially in a competitive technology market [13].

These challenges motivate a new model formulation to improve network performance. As a result, we formulate an alternative model for energy savings in WSNs. Specifically, our model called *input-oriented BCC/VRS model* is formulated for multicast power minimization. This model is capable of assessing the network relative performance, and identified those that are efficient and those that are inefficient including the percentage or level of their efficiency. Furthermore, we formulated the slacks model so as to project the identified inefficient WSNs unto their efficient frontier. The model developed is very appropriate whenever there is need for improvement in saving particular input resources such as energy. The performance of the empirical model, which we formulated, is compared with the existing model. In particular, we found that the proposed input-oriented BCC/VRS model outperformed the existing MIP algorithm in terms of energy saving. In networks management, our proposed models are unique in the sense that they are capable of providing more information about the network without affecting their performance [14].

The remainder of this paper is presented in sections. Section II presents the proposed system architecture. In

section III, input-oriented BCC/VRS model based on the DEA methodology developed. Section IV discusses the implementation and results of the proposed model. Section V presents the analysis of the results while section VI summarily concludes the paper.

## II. Proposed System Architecture

This section first discusses the minimum energy multicast framework, based on the simulation method and then presents the proposed method, based on the DEA method.

### A. Existing Minimum Energy Multicast Method

Figure 1 summarizes the approach that is considered in the existing minimum energy multicast method (simulation method) and the proposed empirical method (DEA method). As it could be observed from Figure 1, the first part, which is the simulation method requires that simulation is set up according to the MIP algorithm requirements and the inputs parameters such as *node* and *sinks* are configured. The algorithm is run based on the parameters set for each WSN and the optimal value of the multicast energy is obtained. Then the average of these values is evaluated using statistical mean.

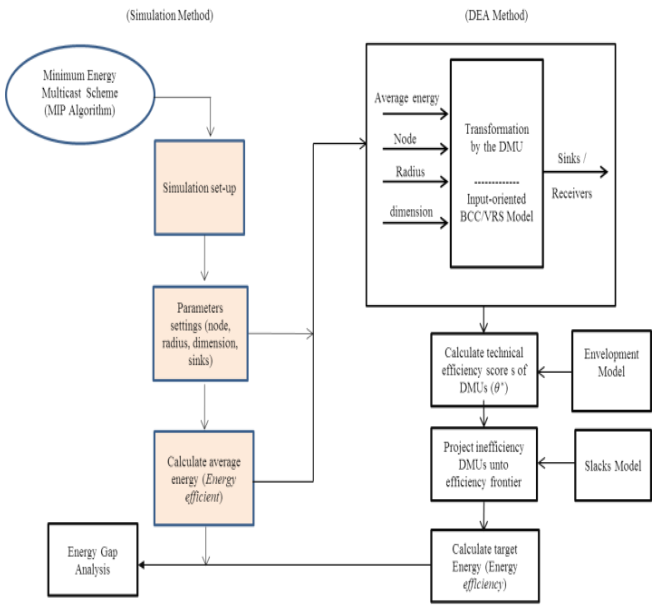



Figure 1: Existing minimum incremental energy (MIP) method

The second part of Figure 1 is the DEA Method, and the proposed method has different components. The first component converts the multiple inputs and output data by the decision making unit (DMU) transformation [15]. Then using envelopment model, the technical efficiency scores of each WSNs called DMUs are evaluated. Furthermore, with the aid of slacks model, the WSNs (DMUs) identified as inefficient are projected unto their efficient frontier. Finally, the target energy of each WSN are evaluated and compared with the average energy evaluated by the MIP algorithm to determine the energy gap. The details about the DEA components as well as the model formation are presented in the next section. The remainder of this section describes how the data set for the DEA implementation in section III are obtained. It should be noted that in order to obtain the data, the existing MIP algorithm is implemented and analyzed.

### B. Model and Approach: Source-Based Trees (SBT)

The existing MIP algorithm considered a source-based trees (SBT) approach, whereby the message is rooted at a particular WSN (the sender) and which are designed to minimize the number of transmissions needed to reach all the members of the multicast group. This type of approach is source-initiated, where multicast sessions are established. The WSN consists of *z* nodes that are randomly distributed in a certain square dimension *d*. It is assumed that any node within this region is permitted to initiate multicast sessions. Multicast request and session durations are generated randomly at the network nodes. Each multicast group consists of the source node and with at least one destination node. Also, intermediate nodes that act as relay may be needed to provide connectivity to all members of the multicast group. Therefore, the multicast tree consists of source node, all destination nodes, and all relay nodes. The nodes are equipped with certain level of energy, and it is assumed that each node can choose its energy level within the maximum $e_{max}$. Also a constant bit rate traffic model is assumed. In addition, it is assumed that bandwidth is not a problem for the transmission. Thus the focus is on transmitter energy.

Given that the received energy is varied as $r^{\alpha}$, where *r* represents the range between the transmitting node and the receiving nodes while the parameter $\alpha$ defines the characteristic of the communication medium, which its value ranges from 2 to 4. If a particular case of node *i* with the minimum transmitted energy $e_{ij}$ that enables the $i^{th}$ nodes to multicast information to $j^{th}$ node is considered, then the distance between nodes *i* and *j* represented by *r* is promotional to $r^{\alpha}$. This is normalized as follows:

$e_{ij}$ = energy needed to support link between nodes i and j

$e_{ij} = r_{ij}^{\alpha}$,

For the network to be fully connected, the maximum transmitted energy $e_{max}$ is required. It is assumed that $\alpha$ is fixed in the environment meaning that the propagation medium is uniform and there is no interference. In this scenario, omnidirectional antennas are considered so as to exploit the wireless multicast advantage (WMA). One of the algorithms designed for the implementation of this approach is called multicast incremental power (MIP) algorithm.

### C. Multicasting Incremental Power Algorithm

In this subsection, the procedure for MIP technique is summarized.

- *Step one*: Modification of *Prim's* algorithm - Find the minimum energy broadcast tree, and develop the Broadcast Incremental Power (BIP) algorithm. BIP algorithm is a modified *Prim's* algorithm making it to be a node-based with incremental function when new nodes are added.
- *Step two*: Perform pruning - The broadcast tree produced by BIP algorithm is pruned. This procedure is used to transform the broadcast tree into a multicast tree known as multicast incremental Power (MIP).
- *Step three*: Perform sweeping - Sweep to eliminate

unnecessary transmissions. This step is added to improve the performance of the multicast algorithm.

Interested readers are referred to [9], [10], [16] for the details about the MIP implementation and analysis. However, we investigate the algorithm and the simulation results of average multicast energy for 54 WSNs are presented in Table 1. This data set is considered for the DEA implementation in the next section.

Table 1: Average multicast energy calculated by MIP algorithm with their corresponding simulation parameters (here classified as data input and output) for 54 WSNs (DMUs)

| | Inputs | | | | Output |
|---|---|---|---|---|---|
| **DMU** | **Ave. energy (e)** | **Dimension (d)** | **Radius (r)** | **Node (z)** | **Sinks (g)** |
| $DMU_1$ | 7.3369 | 10 | 3 | 20 | 2 |
| $DMU_2$ | 8.19434 | 10 | 3 | 20 | 3 |
| $DMU_3$ | 8.98436 | 10 | 3 | 20 | 4 |
| $DMU_4$ | 9.0487 | 10 | 3 | 20 | 5 |
| $DMU_5$ | 9.48655 | 10 | 3 | 20 | 6 |
| $DMU_6$ | 10.4696 | 10 | 3 | 20 | 7 |
| $DMU_7$ | 9.92203 | 10 | 3 | 20 | 8 |
| $DMU_8$ | 10.7971 | 10 | 3 | 20 | 9 |
| $DMU_9$ | 10.8188 | 10 | 3 | 20 | 10 |
| $DMU_{10}$ | 7.4955 | 50 | 5 | 20 | 2 |
| $DMU_{11}$ | 8.51317 | 50 | 5 | 20 | 3 |
| $DMU_{12}$ | 9.34969 | 50 | 5 | 20 | 4 |
| $DMU_{13}$ | 9.33436 | 50 | 5 | 20 | 5 |
| $DMU_{14}$ | 10.02 | 50 | 5 | 20 | 6 |
| $DMU_{15}$ | 9.50838 | 50 | 5 | 20 | 7 |
| $DMU_{16}$ | 10.2374 | 50 | 5 | 20 | 8 |
| $DMU_{17}$ | 10.8043 | 50 | 5 | 20 | 9 |
| $DMU_{18}$ | 10.6641 | 50 | 5 | 20 | 10 |
| $DMU_{19}$ | 6.80406 | 10 | 3 | 30 | 2 |
| $DMU_{20}$ | 7.5984 | 10 | 3 | 30 | 3 |
| $DMU_{21}$ | 8.17163 | 10 | 3 | 30 | 4 |
| $DMU_{22}$ | 8.62102 | 10 | 3 | 30 | 5 |
| $DMU_{23}$ | 9.28325 | 10 | 3 | 30 | 6 |
| $DMU_{24}$ | 8.93184 | 10 | 3 | 30 | 7 |
| $DMU_{25}$ | 9.54203 | 10 | 3 | 30 | 8 |
| $DMU_{26}$ | 9.97383 | 10 | 3 | 30 | 9 |
| $DMU_{27}$ | 9.2635 | 10 | 3 | 30 | 10 |
| $DMU_{28}$ | 6.61611 | 50 | 5 | 30 | 2 |
| $DMU_{29}$ | 7.82157 | 50 | 5 | 30 | 3 |
| $DMU_{30}$ | 7.66053 | 50 | 5 | 30 | 4 |
| $DMU_{31}$ | 8.30668 | 50 | 5 | 30 | 5 |
| $DMU_{32}$ | 8.60528 | 50 | 5 | 30 | 6 |
| $DMU_{33}$ | 9.78652 | 50 | 5 | 30 | 7 |
| $DMU_{34}$ | 9.73328 | 50 | 5 | 30 | 8 |
| $DMU_{35}$ | 9.93892 | 50 | 5 | 30 | 9 |
| $DMU_{36}$ | 9.60166 | 50 | 5 | 30 | 10 |
| $DMU_{37}$ | 6.70263 | 10 | 3 | 40 | 2 |
| $DMU_{38}$ | 6.52575 | 10 | 3 | 40 | 3 |
| $DMU_{39}$ | 7.48482 | 10 | 3 | 40 | 4 |
| $DMU_{40}$ | 7.47209 | 10 | 3 | 40 | 5 |
| $DMU_{41}$ | 8.0599 | 10 | 3 | 40 | 6 |
| $DMU_{42}$ | 8.36012 | 10 | 3 | 40 | 7 |
| $DMU_{43}$ | 8.61111 | 10 | 3 | 40 | 8 |
| $DMU_{44}$ | 8.8669 | 10 | 3 | 40 | 9 |
| $DMU_{45}$ | 9.00432 | 10 | 3 | 40 | 10 |
| $DMU_{46}$ | 6.29201 | 50 | 5 | 40 | 2 |
| $DMU_{47}$ | 6.8135 | 50 | 5 | 40 | 3 |
| $DMU_{48}$ | 7.15764 | 50 | 5 | 40 | 4 |
| $DMU_{49}$ | 7.41543 | 50 | 5 | 40 | 5 |
| $DMU_{50}$ | 8.03526 | 50 | 5 | 40 | 6 |
| $DMU_{51}$ | 8.24464 | 50 | 5 | 40 | 7 |
| $DMU_{52}$ | 8.62671 | 50 | 5 | 40 | 8 |
| $DMU_{53}$ | 8.49562 | 50 | 5 | 40 | 9 |
| $DMU_{54}$ | 9.01432 | 50 | 5 | 40 | 10 |

## III. Methodology and Model Development

### A. Proposed Input-oriented BCC Models

This section derives the input-oriented BCC models. The objective of this model seeks to minimize the inputs through linear programing technique. This model keeps the current empirical level of outputs constant and attempts to minimize the inputs. Variable return to scale (VRS) with input slacks are considered in the model formulations. The resulting model is then called the *input-oriented BCC/VRS envelopment with slack*.

Using envelopment approach for efficiency measurement [17], we define performance in term of efficiency ratio considering a set of $n$ observations for the DMUs where each observation, $DMU_j$ $\{j = 1,2,\dots,n\}$ uses $n$ multiple inputs $x_{ij}$ $(i = 1,2,\dots,m)$ to produce $s$ multiple outputs $y_{rj}$ $(r = 1,2,\dots,s)$. The variable $x_{ij}$ represents the vector of inputs into $DMU_{ij}$ and $y_{rj}$ represents the corresponding vector of outputs. Then, the efficiency ratio (performance) for $DMU_j$ can be expressed as:

$$Performance = \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} = \frac{u_1 y_{1j} + u_2 y_{2j} + \cdots + u_s y_{sj}}{v_1 x_{1j} + v_2 x_{2j} + \cdots + v_m x_{mj}} \quad (1)$$

where $u_r$ $(r = 1,2,\dots,s)$ and $v_i (i = 1,2,\dots,m)$ are unknown weights. This ratio accounts for all outputs and inputs. This type of measure is called *Total productivity factor* [18]. All DEA models implement similar type of performance ratios although with their own specific characteristics. The weights assigned to each input and output is used as variables in the DEA optimisation process.

Furthermore, if we consider a particular $WSN_0$ ($DMU_0$), the objective is to

$$max \frac{\sum_{r=1}^{s} u_r y_{r0}}{\sum_{i=1}^{m} v_i x_{io}} = \frac{u_1 y_{1o} + u_2 y_{2o} + \cdots + u_s y_{so}}{v_1 x_{1o} + v_2 x_{2o} + \cdots + v_m x_{mo}} \quad (2)$$

However, this maximisation problem (2) is unbounded meaning that additional constraints ought to be established. One of the constraints is to assume a set of normalization, one for each DMU. Thus, a condition that the virtual output to virtual input ratio of every DMU must be less than or equal to unity is necessary.

### B. BCC Model in Dual Form – The Envelopment Model

An effective technique to solve the minimization problem is the Linear Programming (LP), a framework that is paramount to optimisation problems [19]. Therefore, we develop the DEA model, which is appropriate for minimisation problem using LP framework. The model is derived as follow:

$\theta^* = min\theta$

Subject to

$$\begin{aligned} &\textstyle\sum_{j=1}^{n} \lambda_j x_{ij} \le \theta x_{i0} \qquad i = 1,2,\dots,m; \\ &\textstyle\sum_{j=1}^{n} \lambda_j y_{rj} \ge y_{r0} \qquad r = 1,2,\dots,s; \\ &\textstyle\sum_{j=1}^{n} \lambda_j = 1 \\ &\lambda_j \ge 0, \qquad j = 1,2,\dots,n, \end{aligned} \quad (3)$$

where $\lambda_j$ are unknown weights with $j = 1,2,\dots,n$ corresponds to the $DMU$ numbers. $DMU_0$ is one of the $n$ DMU under evaluation, and $\theta x_{i0}$ and $y_{r0}$ are the $i^{th}$ input and $r^{th}$ output for $DMU_0$ respectively. Model (3) is based on input-orientation and we assumed the variable returns to scale (VRS). This is a form of envelopment model that is appropriate for input minimisation [20], [21]. Also, model (3) represents the first stage of our energy minimization process.

***Definition 1:*** If $\theta^* = 1$, *then the DMU under evaluation is a frontier point (efficient), that is, no other DMUs operates more efficiently than this DMU. Otherwise if* $\theta^* < 1$, *then the DMU under evaluation is inefficient, that is, this DMU can either increase its output levels or decrease its input levels.*

### C. The Slacks Model

In addition to envelopment model (3), the slacks model is formulated. The important of the slack model is to push the DMUs to their actual optimal efficiency. Slacks model is necessary in case a DMU cannot reach its efficiency frontier after proportional reductions in inputs using envelopment model (3). In order to obtain the slacks in DEA analysis [22], a second stage linear programming model is required to be solved after the dual linear programming model. As a result, the second stage linear programming model is formulated for slack values and is written as:

$max\ \sum_{i=1}^{m} s_j^- + \sum_{r=1}^{s} s_r^+$

Subject to

$$\begin{aligned} &\sum_{j=1}^{n} \lambda_j x_{ij} + s_j^- = \theta_{x_{io}}^* \qquad i = 1, 2, \cdots, \text{m}; \\ &\sum_{j=1}^{n} \lambda_j y_{rj} - s_r^+ = y_{r0} \qquad r = 1, 2, \cdots, \text{s}; \\ &\sum_{j=1}^{n} \lambda_j = 1 \\ &\lambda_j \geq 0 \qquad \text{j} = 1, 2, \cdots, \text{n}. \end{aligned} \tag{4}$$

where θ* is the DEA efficiency score resulted from initial run based on model (3), $s_j^-$ and $s_r^+$ represent input and output slacks respectively. The superscript (-) and (+) represent input reduction and output augmentation respectively.

### D. Efficient Target and Gap Calculation

The efficient target is evaluated so that inefficient DMUs are projected unto efficient frontier. The level of efficient target for inputs and outputs can be calculated using the following relations:

$$\begin{cases} \bar{x}_{io} = \theta^* x_{io} - S_i^{-*} & i = 1, 2, \dots, m \\ \bar{y}_{ro} = y_{ro} + S_r^{+*} & r = 1, 2, \dots, s \end{cases} \tag{5}$$

Numerically, efficient target can be calculated taking for instance the target for the input values, then the input values are multiplied with an optimal efficiency score (θ*), and slack amounts are subtracted from this value. The gap calculation is the different between actual input values (average energy) and the target value (target energy).

## IV. Implementation and Results of Proposed Model

### A. The Data Set

The main method of gathering the data used for implementing the model developed is through the simulation method, which its results are presented in section II. This simulation results together with their parameters are classify into inputs and outputs for the DEA solver. In the simulation approach, the output result is the average energy consumed by the multicast nodes. In this work, we assumed that the average energy (e) required by multicast nodes is known while DEA minimizes the average energy without alter the multicast nodes and other variables. So, the simulation variables and values are considered for the DEA implementation. As specified in Table 1, each of the WSN takes four inputs variables and one output variable in order to multicast a message from a source to a group of receivers.

### B. Evaluation Platform

Software tools have been specially packaged for DEA implementation and analysis. These include DEA-Solver, which is an add-on to Microsoft Excel, and specialized DEA packages like DEAP (DEA program), Warwick DEA, ON-Front and DEAOS (Data Envelopment Analysis Online Software). Some of these software like DEAP is freely available over the Internet. In addition, DEA solver is available as open source online software. This research makes use of DEAOS for the implementation. The software is also available online [23]. This software package is user friendly and adequate for research analysis. Technically, the software consists of two parts: DEA libraries (called the DEAOS solver) and the lpsolver libraries (known as linear programming solver). These two components are very important to the optimization solutions. The readers are referred to [23] for details about the DEAOS package and user's documentation.

### C. Input-oriented BCC/VRS Implementation Process

The implementation process consists of the following components: (i) *Creating a DEA Problem* – First the DEA problem must be created. Our data are organized like those in Table 1, where data were classified into input and output. Alternatively, DEA allow problem to be prepared in Excel and imported to be solved by DEA. (ii) *Configuring the variables* – DEA variables are configured into input and output data for the DMU transformation. (iii) *Configuring a DEA Problem* (set model type etc...) – The DEAOS solver provides basic parameters such as "envelopment/multiplier form", type of "returns to scale", and type of "orientation". (iv) *Saving the problem to .xls file* – DEA provide its internal Excel file, which is compatible with MS Excel. The DEA data can be saved on this file for immediate and future references. (v) *Solving the problem* – After all the necessary configurations have been done, the DEA problem is solved to provide solution according to the model selected. The solution could be viewed onscreen or export in an Excel file. (vi) *Export and Open the .xls file* – The solution file can be opened in .xls format for analysis. The solution objectives that .xls file contains are: efficiency scores, weights, slacks lambdas, peer group (efficiency reference set), and Projections. These are analyzed and summarized. Based on our problem definition and models developed, we carried out the following analysis for energy efficiency: (a) Technical efficiency/inefficiency rating, and analysis, (b) Input slacks analysis, (c) Targets (Projections) analysis, (d) Gap calculation and analysis.

### D. Results of the Input-oriented BCC/VRS model

The result obtained from the implementation of input-oriented BCC/VRS model is presented in Table 2. The analysis and discussions of the results are presented in the next section.

## V. Analysis of Results and Discussion

### A. Efficiency and Inefficiency Scores, and Analysis

In this section, input-oriented BCC/VRS model is analyzed for the technical efficiency and inefficiency of 54 WSNs. DEA makes use of model (3) to compare each DMU with all other DMUs, and then identifies those DMUs that are operating inefficiently compared with other DMUs' actual operating results. This is done by locating the best practice or relatively efficient DMUs. The amount of the inefficiency DMUs compared to the best practice DMUs is evaluated. The best practice DMUs, which are relatively efficient are assigned with efficiency rating (score) of θ = 1

while the inefficient DMUs are assigned with efficiency rating (score) of $\theta < 1$. Column 2 of Table 2 reports the results of DEA efficiency scores for all the 54 WSNs. The results show that 39 WSNs are technically efficient while 15 are inefficient. Specifically, as could be observed from Table 2, $DMU_1$ to $DMU_{28,}$ $DMU_{37}$ to $DMU_{46}$ and $DMU_{53}$ have efficiency score of $\theta = 1$ meaning that they are efficient. $DMU_{29}$ to $DMU_{36,}$ $DMU_{47}$ to $DMU_{52}$ and $DMU_{54}$ have efficiency scores of $\theta < 1$ but greater than 0 meaning that they are inefficient. However, the inefficient DMUs under input-oriented BCC/VRS model have capability to improve the performance of inefficient DMUs by reducing certain inputs (energy) proportionately. For instance, $DMU_{29}$ can improve its efficiency score by reducing certain inputs up to 8.8% (1.0 - 0.9124057). In a similar manner, $DMU_{30}$ can do so with approximately 3.9% (1.0 - 0.961053) input reduction. Also, $DMU_{54}$ is closer to an efficiency frontier, and needs only a 0.1% (1 - 0.9990051) reduction of its input resources. This analysis is followed by the slack analysis.

Table 2: Efficiency scores, input slack values, target energy and energy gap reports for input-oriented BBC model

| DMU | Efficiency Score | Ave. Energy | Target Energy | Energy gap |
|---|---|---|---|---|
| **DMU1** | **1** | **7.3369** | **7.3369** | **0** |
| DMU2 | 1 | 8.19434 | 7.77214 | 0.422203 |
| DMU3 | 1 | 8.98436 | 8.19861 | 0.78575 |
| DMU4 | 1 | 9.0487 | 8.62947 | 0.419235 |
| **DMU5** | **1** | **9.48655** | **9.48655** | **0** |
| DMU6 | 1 | 10.4696 | 9.49118 | 0.978425 |
| **DMU7** | 1 | 9.92203 | 9.92203 | 0 |
| DMU8 | 1 | 10.7971 | 10.3704 | 0.426685 |
| **DMU9** | **1** | **10.8188** | **10.8188** | **0** |
| DMU10 | 1 | 7.4955 | 7.3369 | 0.1586 |
| DMU11 | 1 | 8.51317 | 7.7528 | 0.76037 |
| DMU12 | 1 | 9.34969 | 8.1687 | 1.18099 |
| DMU13 | 1 | 9.33436 | 8.5846 | 0.74976 |
| DMU14 | 1 | 10.02 | 9.0005 | 1.0195 |
| DMU15 | 1 | 9.50838 | 9.4164 | 0.09198 |
| DMU16 | 1 | 10.2374 | 9.8323 | 0.4051 |
| DMU17 | 1 | 10.8043 | 10.2482 | 0.5561 |
| **DMU18** | **1** | **10.6641** | **10.6641** | **0** |
| **DMU19** | **1** | **6.80406** | **6.80406** | **0** |
| DMU20 | 1 | 7.5984 | 7.0868 | 0.511597 |
| DMU21 | 1 | 8.17163 | 7.81855 | 0.35308 |
| DMU22 | 1 | 8.62102 | 8.05938 | 0.561645 |
| DMU23 | 1 | 9.28325 | 8.01967 | 1.263577 |
| DMU24 | 1 | 8.93184 | 8.54103 | 0.390815 |
| DMU25 | 1 | 9.54203 | 8.64159 | 0.900443 |
| DMU26 | 1 | 9.97383 | 9.02268 | 0.951155 |
| **DMU27** | **1** | **9.2635** | **9.2635** | **0** |
| **DMU28** | **1** | **6.61611** | **6.61611** | **0** |
| DMU29 | 0.9124057 | 7.82157 | 7.13645 | 0.685125 |
| DMU30 | 0.961053 | 7.66053 | 7.36218 | 0.298355 |
| DMU31 | 0.9333465 | 8.30668 | 7.75301 | 0.553669 |
| DMU32 | 0.9381968 | 8.60528 | 8.07345 | 0.531834 |
| DMU33 | 0.8734144 | 9.78652 | 8.54769 | 1.238833 |
| DMU34 | 0.9065969 | 9.73328 | 8.82416 | 0.909119 |
| DMU35 | 0.9228243 | 9.93892 | 9.17188 | 0.767043 |
| DMU36 | 0.9755018 | 9.60166 | 9.36644 | 0.235223 |
| DMU37 | 1 | 6.70263 | 6.52575 | 0.17688 |
| **DMU38** | **1** | **6.52575** | **6.52575** | **0** |
| DMU39 | 1 | 7.48482 | 6.91686 | 0.567963 |
| DMU40 | 1 | 7.47209 | 7.23391 | 0.238177 |
| DMU41 | 1 | 8.0599 | 7.69907 | 0.360829 |
| DMU42 | 1 | 8.36012 | 7.94208 | 0.418044 |
| DMU43 | 1 | 8.61111 | 8.29616 | 0.314953 |
| DMU44 | 1 | 8.8669 | 8.65024 | 0.216661 |
| **DMU45** | **1** | **9.00432** | **9.00432** | **0** |
| **DMU46** | **1** | **6.29201** | **6.29201** | **0** |
| DMU47 | 0.9660999 | 6.8135 | 6.58252 | 0.230978 |
| DMU48 | 0.9656287 | 7.15764 | 6.91162 | 0.246017 |
| DMU49 | 0.9743615 | 7.41543 | 7.22531 | 0.19012 |
| DMU50 | 0.9459755 | 8.03526 | 7.60116 | 0.434101 |
| DMU51 | 0.9591058 | 8.24464 | 7.90748 | 0.337158 |
| DMU52 | 0.9555361 | 8.62671 | 8.24313 | 0.383577 |
| **DMU53** | **1** | **8.49562** | **8.49562** | **0** |
| DMU54 | 0.9990051 | 9.01432 | 9.00535 | 0.008969 |
| | | 468.428 | 446.198 | 22.23064 |

### B. *Input (Energy) Slacks Analysis*

The slack mathematical derivation model (4) is run after envelopment model (3). The slack analysis requires that none of the efficient DMUs have any slacks. However, inefficient DMUs have slacks values. As mentioned, slack model is needed if a DMU cannot reach the efficiency frontier after model (3) is executed. The effect of slack will be discussed in the next subsection.

### C. *Target (Projection) Analysis*

Column 3 is the record of the average energy from the simulation. The target calculation is computed by model (5). In order to calculate the target values, the input value (presented in column 3) are multiplied with an optimal efficiency score (presented in column 2) and then slack amounts are subtracted from this product. This target (projection) value is calculated and presented in column 4. This is the projected energy for multicasting in WSNs. Observe from Table 2 that the target energy for $DMU_1$ (column 4) is the same for input energy (column 3). This is because the efficiency scores $\theta = 1$ and all the input slacks are zero. In order words, $DMU_1$ is *fully efficient*. $DMU_2$ is efficient with $\theta = 1$ but it is *weakly efficient* because it has some slacks values. As it could be observed from Table 2, the projected energy (7.77214) for $DMU_2$ is different from the input energy (8.19434). However, $DMU_{29}$ belongs to the class of inefficient DMUs because the efficiency scores $\theta < 1$. This suggests that it must have some slack values and it must reduce some input values to become efficient.

### D. *Gap Calculation and Analysis*

The gap analysis, which is the difference between the MIP method (simulation model) and the empirical method (DEA model) is reported in column 5 of Table 2. This difference represents the amount of the energy saved by the individual WSN if they were to operate efficiently. Figure 2 plot the graph of energy saved by individual WSN. For example, $DMU_{23}$ has the highest energy saved which is 1.2635767 follow by $DMU_{33}$ with the value equal 1.2388326. We also calculate the total energy saved. For example, as shown at the bottom of Table 2, the total energy save is 468.428 minus 446.198. The difference, which is equal to 22.231, is equivalent to 5% - the energy saves. Furthermore, we present the pie chart of the total energy saved and the total target energy in Figure 3. This reduction in energy presented by the input-oriented BCC/VRS method is significant compare to the traditional MIP that is based on simulation method.

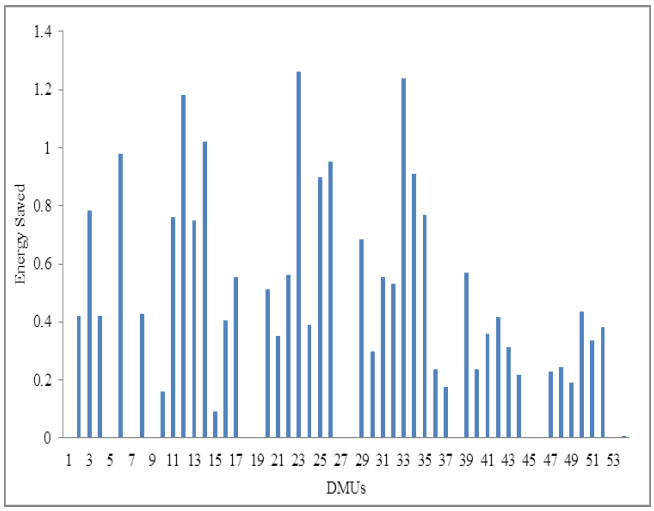



Figure 2: Energy saved by individual WSNs using input-oriented BCC/VRS DEA model

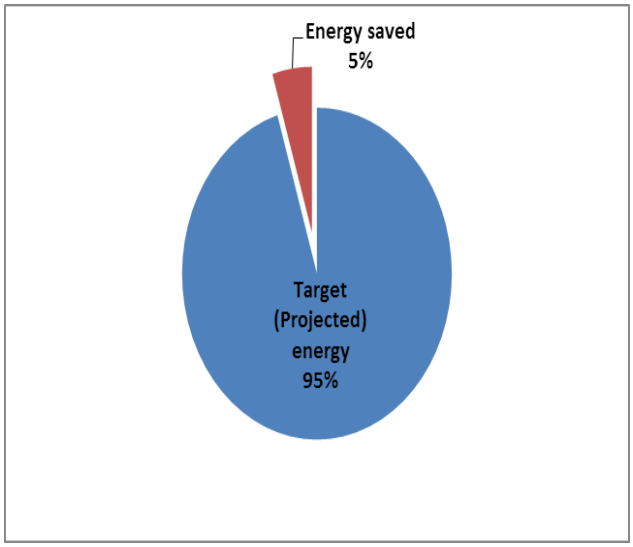



Figure 3: Percentage of the total target energy and the total energy saved using input-oriented BCC/VRS DEA model

## VI. Conclusion

This research work has been able to study the problem of minimizing energy in wireless multicast networks with application to WSNs. The MIP algorithm which has been widely considered in literature as effective method to reduce multicast energy has been explored. We found that MIP method has not efficiently minimized the multicast energy by the WSN nodes. In addition, in terms of energy efficiency, the proposed empirical model based upon BCC model outperformed the existing MIP method that is implemented using simulation method. Beside the reduction in multicast energy, the proposed model provides adequate information about WSNs. Such information includes; efficiency scores (ratings), slacks and efficiency targets for inefficient WSNs. Thus, the empirical DEA model minimizes energy consumed by the WSNs without the output values altered. This is generally not possible with the MIP based on simulation method.

## Acknowledgements

This research is supported by Telkom South Africa, Jasco / TeleSciences, and the Department of Trade and Industry / National Research Foundation / Technology and Human Resources Programme (DTI/NRF/THRIP).

## References

[1] G. Xue, "Minimum-cost QoS multicast and unicast routing in communication networks," *IEEE Trans. Commun.*, vol. 51, no. 5, pp. 817–824, May 2003.

[2] W. Chun and W. Tang, "Multicasting in Wireless Sensor Networks," *Jt. Int. Conf. Opt. Internet Next Gener. Netw.*, pp. 251–253, 2006.

[3] R. Biradar, C. Manvi, and S. Sunilkumar, "Review of multicast routing mechanisms in mobile ad hoc networks," *J. Netw. Comput. Appl.*, vol. 35, no. 1, pp. 221–239, Jan. 2012.

[4] A. Striegel and G. Manimaran, "A Survey of QoS Multicasting Issues," *IEEE Commun. Mag.*, vol. 40, no. 6, pp. 82–87, 2002.

[5] A. A. Ajibesin, N. Ventura, A. Murgu, and H. A. Chan, "Cost-efficient multicast over coded packet wireless networks using data envelopment analysis," in *2013 IEEE 10th Consumer Communications and Networking Conference (CCNC)*, 2013, pp. 546–551.

[6] M. Cagalj, J.-P. Hubaux, and C. Enz, "Minimum-energy broadcast in all-wireless networks: NP-completeness and distribution issues," in *ACM MobiCom*, 2002, pp. 172–182.

[7] S. Guo and O. Yang, "Minimum-energy multicast in wireless ad hoc networks with adaptive antennas: MILP formulations and heuristic algorithms," *IEEE Trans. Mob. Comput.*, vol. 5, no. 5, pp. 333–346, 2006.

[8] D. Yuan, J. Bauer, and D. Haugland, "Minimum-energy broadcast and multicast in wireless networks: An integer programming approach and improved heuristic algorithms," *Ad Hoc Networks*, vol. 6, no. 5, pp. 696–717, Jul. 2008.

[9] E. J. Wieselthier, G. D. Nguyen, and E. Anthony, "Algorithms for energy-efficient multicasting in static ad hoc wireless networks," *Mob. Networks Appl.*, vol. 6, no. 3, pp. 251–263, 2001.

[10] Y.-F. Wen and W. Liao, "Minimum power multicast algorithms for wireless networks with a Lagrangian relaxation approach," *Wirel. Networks*, vol. 17, no. 6, pp. 1401–1421, Jun. 2011.

[11] D. C. Wade and Z. Joe, *Modeling Performance Measurement: Applications and Implementation Issues in DEA*. Springer Science Business Media LLC, 2009.

[12] A. A. Ajibesin, N. Ventura, A. Murgu, and H. . Chan, "Data envelopment analysis: Efficient technique for measuring performance of wireless network coding protocols," in *15th International Conference on Advanced Communication Technology*, 2013, pp. 1122–1127.

[13] P. Katrina, "Understanding Cost-Effectiveness of Energy Efficiency Programs: Best Practices, Technical Methods, and Emerging Issues for Policy-Makers," *A Resour. Natl. Action Plan Energy Effic.*, 2008.

[14] S. C. Ray, *Data envelopment analysis: theory and techniques for economics and operations research*, 2st edn. Cambridge University Press, 2012.

[15] E. Chanes, A., Cooper, W. W., Rhodes, "Measuring the efficiency of decision making units," *Eur. J. Oper. Res.*, vol. 2, pp. 429–444, 1978.

[16] F. Ingelrest and D. Simplot-Ryl, "Localized Broadcast Incremental Power Protocol for Wireless Ad Hoc Networks," in *10th IEEE Symposium on Computers and Communications (ISCC'05)*, 2005, pp. 28–33.

[17] G. Kastaniotis, E. Maragos, C. Douligeris, and D. K. Despotis, "Using data envelopment analysis to evaluate the efficiency of web caching object replacement strategies," *J. Netw. Comput. Appl.*, vol. 35, no. 2, pp. 803–817, Mar. 2012.

[18] T. J. Coelli, D. S. Rao, C. J. O'Donnell, and G. E. Battese, *An Introduction to Efficiency and Productivity Analysis*, Second Edi. New York, USA: Springer Science + Business Media, 2005.

[19] K. Cullinane, D.-W. Song, and T. Wang, "The Application of Mathematical Programming Approaches to Estimating Container Port Production Efficiency," *J. Product. Anal.*, vol. 24, no. 1, pp. 73–92, Sep. 2005.

[20] L. M. Seiford, "Data envelopment analysis: The evolution of the state of the art (1978-1995)," *J. Product. Anal.*, vol. 7, no. 2–3, pp. 99–137, Jul. 1996.

[21] Z. Joe, *Qualitative Models for Performance Evaluation and Benchmarking: Data Envelopment Analysis with Spread sheets*. Springer Science Business Media LLC, 2009.

[22] K. Tone, "A slacks-based measure of efficiency in data envelopment analysis," *Eur. J. Oper. Res.*, vol. 130, no. 3, pp. 498–509, 2001.

[23] "Data Envelopment Analysis," 2014. [Online]. Available: http://www.deaos.com/. [Accessed: 20-May-2014].

## Biography

**Adeyemi A. Ajibesin** received the MSc degree in Electrical Engineering, and BSc degree honors with distinction in Computer Science from University of Cape Town, and Olabisi Onabanjo University respectively. He is presently studying towards his PhD degree at the University of Cape Town. His research interests include Computer networking, telecommunication and wireless technologies.

# On Rayleigh Approximation of the Multipath PLC Channel: Broadband through the PLC Channel

A.M Nyete, T.J.O. Afullo, and I. Davidson
Discipline of Electrical, Electronic and Computer Engineering
University of KwaZulu-Natal
Durban, South Africa
Email: 212536330@stu.ukzn.ac.za , afullot@ukzn.ac.za, davidson@ukzn.ac.za

**Abstract: The powerline carrier communication (PLC) channel has received a lot interest from researchers in the recent past due to its potential in bridging the digital divide between rural and urban communities. The power grid presents a ready medium for communications albeit with limitations just like any other technology. PLC technology is actually one of the least expensive ways of providing internet access to people in remote/rural areas. In this paper, we present a novel approach in characterizing the fading characteristics of the PLC channel. We develop a fading model based on the none line-of-sight (NLOS) and multipath characteristics of the channel. More specifically, the concept of Rayleigh fading phenomena is utilized to study and propose a multipath model of the PLC channel. The validity of the model proposed is then tested by comparing error rate performance characteristics for different number of branching nodes in an orthogonal frequency division multiplexing (OFDM) based binary phase shift keying (BPSK) PLC system.**

**Index Terms: Powerline carrier communication, NLOS, multipath, Rayleigh, broadband, OFDM, BPSK.**

## I. INTRODUCTION

The power grid offers a ready medium for communication purposes. As such, it is the most expansive network that is in existence. It is a ubiquitous network that provides a connectivity point from every socket in every room in every building connected to the grid. Even though the use of the power grid as a communication channel has been in existence for a long time; mainly for control and monitoring functions of the power grid elements, the potential of such a channel as a possible avenue in the delivery of internet and as part of home networking solutions was ignored for a long time. However, in the recent past, driven by the advancement of digital error control techniques, digital modulation techniques, digital equalization as well as interleaving and decoding functions of the digital signal processing era, the PLC channel has been explored as a broadband channel. PLC is attractive as a means of providing broadband connectivity to rural communities where deployment of other technologies may not be viable due to economic aspects in terms of initial set up capital requirements and the purchasing power of the people. PLC is one of the cheapest technologies since the infrastructure is already in place. There is no need for laying of new cables; only terminal equipment like modems are required [1-7].

PLC technologies are divided into two broad categories; narrowband and broadband. Narrowband PLC (NB-PLC) operates in the frequency range between 3-500 KHz. The achievable data rates are in the range of a few kilobits per second. Its application is in the automation of services like home equipment automation, automatic metering reading (AMR), ground lights control at airport runaways as well as street light control, among a host of other uses especially for smart grid applications. Broadband PLC (BB-PLC) operates in the frequency range between 1-300MHz and the achievable data rates can be as high as 300 megabits per second. Broadband PLC is mainly applied in the networking of home equipment and provision of broadband internet [1, 2, 7].

The PLC channel is however horrible for communication purposes. This is because of the time-varying, frequency-selective nature of the channel. The channel characteristics vary from hour to hour, and is different for different geographical locations (urban, rural, suburban, industrial, residential etc) as well. More specifically, the channel properties are determined by the number and length of the electrical branches, the cable type and diameter, the network topology, end to end distance between the transmission and reception side, voltage levels (low, medium or high voltage) and whether the power cable is overhead or underground. Thus the characteristics of the PLC channel vary across the three main voltage levels in a power grid, viz a viz the low voltage (LV) access network, medium voltage (MV) distribution network and high voltage (HV) transmission network. Attenuation (fading) is one of the major channel characteristics and it varies with the power grid type; with the LV network experiencing the highest attenuation levels and HV network the least, multipath due to various reflections in the branching nodes, noise from various loads connected to the network and electromagnetic interference from broadcast stations [2, 6-9].

Driven by the desire to provide broadband communications through the power grid, there is need to accurately predict the characteristics of the channel in terms of amplitude, phase, attenuation and noise. Various models have been fronted by the PLC research community towards this goal. One of the earliest models was that by Hensen and Schulz [10, 11], which showed that the attenuation varies proportionally to the frequency. Philipps [12, 13], in his pioneering work on the PLC channel proposed models based on the concatenation of

several series resonant RLC circuit as well as the multipath behavior exhibited by the channel. He employed an evolutionary strategy in the development of the models. On the other hand, Zimmerman and Dostert [14, 15] extended the work done by Philipps by including the attenuation aspect of the PLC channel in the multipath model. This model is one of the widely used models in PLC research.

## II. THE MULTIPATH PLC CHANNEL MODEL

Zimmerman and Dostert proposed a PLC channel model that presents the channel as a multipath environment. This model was developed from channel measurements. This model is based on the fact that the power network is made up of multiple branches which are terminated in either matched or mismatched loads. Thus signal reflections are bound to occur at the branching nodes due to the impedance mismatch. Thus the received signal at the receiver is made of different versions of the transmitted signal that are delayed and attenuated in different proportions. Hence, the received signal is a vector addition of the different signal components. The model is summarized by the following expression [14]:

$$H(f) = \sum_{i=1}^{N} g_i \cdot e^{-(a_0 + a_1 f^k)\cdot d_i} \cdot e^{-j2\pi f \tau_i} \qquad (1)$$

Where $N$ is the number of dominant paths considered in the propagation; $g_i$ is the weighting factor, which a product of the different reflection and transmission factors; $a_0$ , $a_1$ and exponent $k$ are the parameters that define the frequency-dependent attenuation; $d_i$ is the path length, $\tau_i$ is the path delay given by the following expression [14]:

$$\tau_i = \frac{d_i}{v_p} = \frac{d_i\sqrt{\varepsilon_r}}{c_0} \qquad (2)$$

Where $\varepsilon_r$ is the insulating material's dielectric constant, $c_0$ is the speed of light , $d_i$ is the length of a path and $v_p$ is the propagation speed. Thus we can see from (1) that the model is characterized by three different components: the weighting factor, the attenuation portion and the delay portion. The factor $e^{-(a_0 + a_1 f^k)\cdot d_i}$ determines the amount of attenuation that takes place during signal transmission in the PLC channel. The factor $e^{-j2\pi f \tau_i}$ is the delay portion. The transmission and reflection coefficients are always less than one and so it goes without saying that the net product of all the transmission and reflection coefficients is also less than one, viz a viz:

$$|g_i| \leq 1 \qquad (3)$$

The attenuation factor is obtained from the complex propagation constant by using transmission line analogy, that is:

$$\gamma = k_1\sqrt{f} + k_2 f + jk_3 f \qquad (4)$$

Where $k_1\sqrt{f} + k_2 f$ is the attenuation constant and $k_3 f$ is the phase constant. The constants $k_1$, $k_2$ and $k_3$ summarize the geometrical and material properties of the network. Thus it can be seen from (4) that the attenuation increases with frequency. The weighting and the delay factors are obtained when the frequency response of the PLC channel is converted into time domain. The weighting factor is inversely proportional to the delay factor. This is due to the reduction in signal power as the signal travels through points of discontinuity.

## III. RAYLEIGH MULTIPATH FADING PLC CHANNEL MODEL

Transmission through a PLC channel is in general not a line-of-sight (LOS) case. This is primarily so because the signal that is sent at the transmitter will pass through multiple paths (different network branches), suffering different degrees of reflections and transmissions at the branching nodes due to impedance mismatch. Thus, this forth and back reflections of the signal produces echoes of the main signal. Some of these echoes add constructively while others do so destructively. Based on the NLOS nature of a typical power network, and the number and length of the branches, the PLC channels can be classified into two main categories:

Good PLC channels-These are PLC channels which have a small number of branches and the branches have large electrical length. This channel has shallow notches in the frequency response characteristics. The PLC channels for networks near rural and suburban areas belong to this class. Thus this channel experiences low attenuation.

Bad PLC channels-These are PLC channels which have a large number of electrical branches whose electrical length is small. These channels have deep notches in the frequency characteristics. The PLC channel in urban and densely populated areas belongs to this class. Thus high attenuation is the main characteristic of this channel.

A graphical representation of the multipath scenario in terms of the transmitted and received signals, based on the intuitive visualization that an impulse transmitted at the transmitter will be received as a train of impulses, is shown in Figure 1.

Let the transmitted bandpass signal be :

$$x(t) = \Re\{x_b(t)e^{j2\pi f_c t}\} \qquad (5)$$

Where $x_b(t)$ is the baseband signal, $f_c$ is the carrier frequency and $t$ is time.

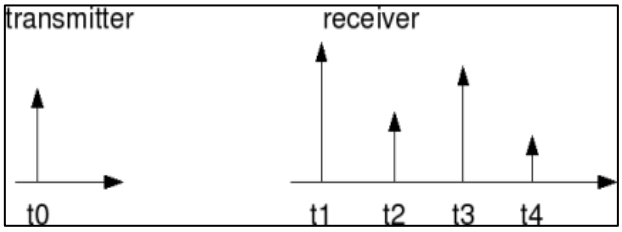



Figure 1: Impulse response of a multipath channel.

As we can see from Figure 1, the signal that is transmitted gets to the receiver via multiple paths where the $n^{th}$ path has an attenuation factor $\alpha_n(t)$ and a delay that is given by $\tau_n(t)$. Thus, the signal that is received can be represented as:

$$r(t) = \sum_n \alpha_n(t) x[t - \tau_n(t)] \tag{6}$$

If we then substitute Equation (5) into Equation (6) for the baseband signal, we obtain the following:

$$r(t) = \Re\left\{\sum_n \alpha_n(t) x_b[t - \tau_n(t)]\, e^{j2\pi f_c[t-\tau_n(t)]}\right\} \tag{7}$$

And, the received signal baseband equivalent is:

$$r_b(t) = \sum_n \alpha_n(t)\, e^{-j2\pi f_c \tau_n(t)} x_b[t - \tau_n(t)] \tag{8}$$

Equation (8) can also be written as:

$$r_b(t) = \sum_n \alpha_n(t)\, e^{-j\theta_n(t)} x_b[t - \tau_n(t)] \tag{9}$$

Where $\theta_n(t) = 2\pi f_c \tau_n(t)$ is the phase of the $n^{th}$ path.

Thus, we obtain the impulse response of the channel as:

$$h_b(t) = \sum_n \alpha_n \delta\big(t - \tau_n(t)\big)\, e^{-j\theta_n(t)} \tag{10}$$

Equations (5) to (10) summarize the multipath scenario experienced in many typical communication channels, both wired and wireless.

If the delay $\tau_n(t)$ changes by $\frac{1}{f_c}$, the phase change on each path can be $2\pi$ radians, if a Rayleigh fading model is adopted. Since the distance between the branching nodes is much larger than the wavelength of the carrier frequency, the assumption that the phase is uniformly distributed between 0 and $2\pi$ radians and, that the phases of each path are independent would be reasonable.

If the number of paths is large, each path can be modeled as a circularly symmetric complex Gaussian variable with time as the variable through the application of the Central Limit Theorem; and this underscores the basis of the Rayleigh fading multipath model for the PLC channel proposed.

A circularly symmetric Gaussian random variable is of the form:

$$Z = X + jY \tag{11}$$

Where the real and imaginary parts are zero mean independent and identically distributed (iid) Gaussian random variables. Also, for the circularly symmetric Gaussian random variable $Z$:

$$E\left[e^{j\theta} Z\right] = e^{j\theta} E[Z] \tag{12}$$

The variance is used to completely specify the statistics of the circularly symmetric Gaussian random variable:

$$\sigma^2 = E[Z^2] \tag{13}$$

Then, the circularly symmetric Gaussian random variable $Z$ which has a probability density,

$$p(z) = \frac{z}{\sigma^2} e^{\frac{-z^2}{2\sigma^2}}, \qquad z \geq 0 \tag{14}$$

is known as the Rayleigh random variable. Even though the receiver and transmitter are fixed in PLC channels, the channel characteristics vary with time of day, topology of the network, cables types and diameter, number of branching nodes as well as the number and length of the branches, among other factors. This dynamic nature of the channel renders the Rayleigh fading approximation of the PLC channel in a multipath environment valid.

As such, how often the channel is changing determines the "Doppler effect" suffered. Figures 2 and 3 show fast and slowly varying simulated typical channel response envelopes for an approximate 20MHz channel based on the proposed Rayleigh multipath PLC channel. From these figures, we observe that for a channel that is highly varying, there are several deep notches in the response characteristic, which can be traced back to the fact such a PLC channel will have many points of reflection or scattering (branching nodes) and that would also mean that there are more electrical loads connected to such a network and vice versa.

Finally, we emphasize here that the model proposed is valid for a channel that has a large number of branching nodes (reflectors) or branches; and this is actually a scenario that stands out in many practical PLC networks, especially on the low voltage side of the power grid. The PLC channel approximation that is proposed here is validated in the next section by assessing how it varies in terms of the error rate performance characteristics for different number of branching nodes.

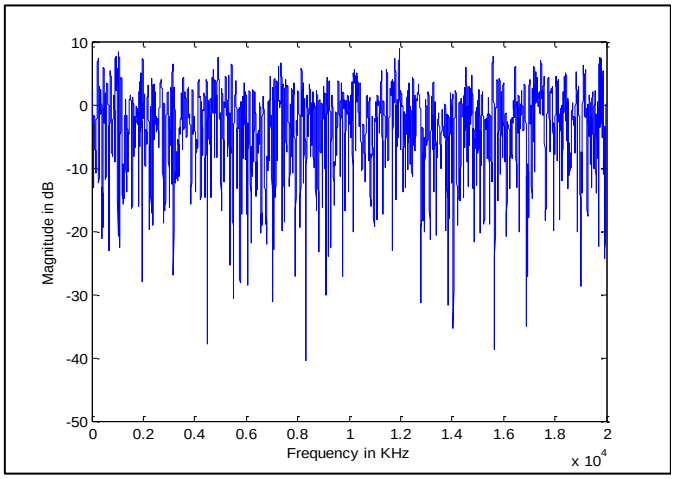



Figure 2: A highly-varying Rayleigh Multipath PLC Channel

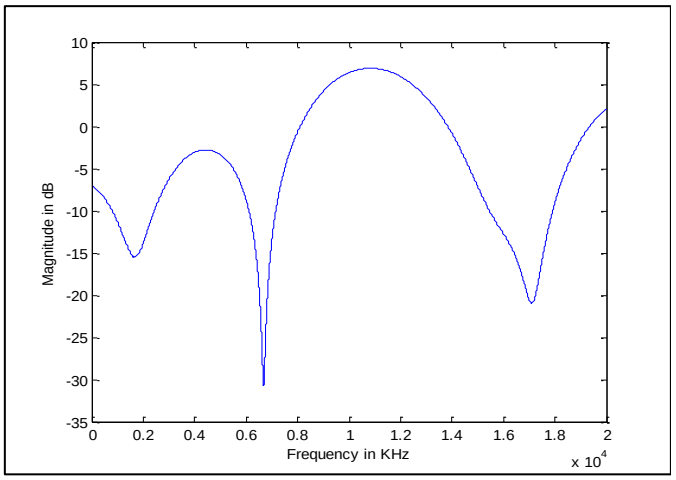



Figure 3: A slowly-varying Rayleigh Multipath PLC Channel

## IV. PERFORMANCE EVALUATION OF THE PROPOSED MODEL

As explained in section III, the channel is modeled by considering the number of branching nodes that comprise the PLC network. If we consider the network to be comprised of *n* branching nodes, with each real and imaginary parts of each node being a Gaussian random variable that is independent, then the impulse response is defined by:

$$h(t) = \frac{1}{\sqrt{n}}[h_1(t-t_1) + h_2(t-t_2)] + \cdots + h_n(t-t_n) \quad (15)$$

where $h_1(t-t_1)$, $h_2(t-t_2)$ are the first and second branching nodes channel coefficients respectively and so on, and $\frac{1}{\sqrt{n}}$ is normalization factor of the average channel power over multiple channel to 1.

The channel is highly varying in time, which therefore means that the symbol that is transmitted gets multiplied by a randomly varying complex number. Since this complex number is modeling a frequency selective Multipath PLC channel, the real and imaginary parts are Gaussian distributed having zero mean and variance 0.5. Equalization then needs to be performed at the receiver by dividing the received symbol by the *a priori* complex number. Assuming that the number of branching nodes is lower than the cyclic prefix duration, meaning that there is no intersymbol interference, then the error rate characteristics for an OFDM based BPSK PLC system should be the same as that experienced in a BPSK PLC system for the same channel conditions.

Even though the channel consists of several branching nodes; that is, frequency selective, in the presence of OFDM Modulation, each subcarrier experiences independent flat fading. In this validation, the following OFDM parameters are used: FFT size=64, number of subcarriers=52, FFT sampling frequency=20 MHz, subcarrier spacing=312.5 kHz, cyclic prefix duration=0.8 µs, data symbol duration=3.2 µs, and the total symbol duration=4 µs. The number of branching nodes considered is 5, 10 and 15.

## V. RESULTS AND DISCUSSION

The performance of the proposed model is investigated in terms of the BER and Eb/No characteristics. Figures 4, 5 and 6 show these characteristics for 5, 10 and 15 branching nodes respectively. From these curves, we see that the simulated BER and the theoretical BER are very close. Again, we notice that there is no much difference in the performance of the proposed model as we vary the number of branching nodes. For example, at a BER value of $10^{-3}$, all the three graphs have an Eb/No value of approximately 24 dB. Also, at a BER value of $10^{-4}$, for 5 branching nodes, the Eb/No value is 33 dB; for 10 branching nodes, the Eb/No value is 34.5 dB; for 15 branching nodes, the Eb/No value is 34 dB. Therefore, we conclude that as long as the value of the cyclic prefix duration used is greater than the number of branching nodes, the performance of the proposed model is satisfactory.

Lastly, we note that the location of the transmitter or receiver plays an important role in the Eb/No values, that is, a receiver that is close to a noise source would have a very poor Eb/No value and vice versa. However, other digital signal processing techniques can be implemented to improve the performance of the system, for example channel coding.

## VI. CONCLUSION

In this paper, a PLC multipath model based on Rayleigh fading approximation, anchored on the fact the PLC channel is a NLOS channel with no dominant path between the transmitter and the receiver has been proposed.

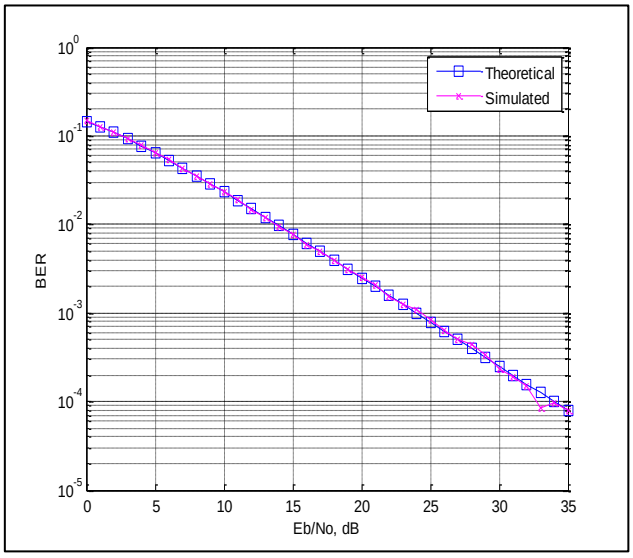



Figure 4: BER vs Eb/No characteristics for 5 branching nodes

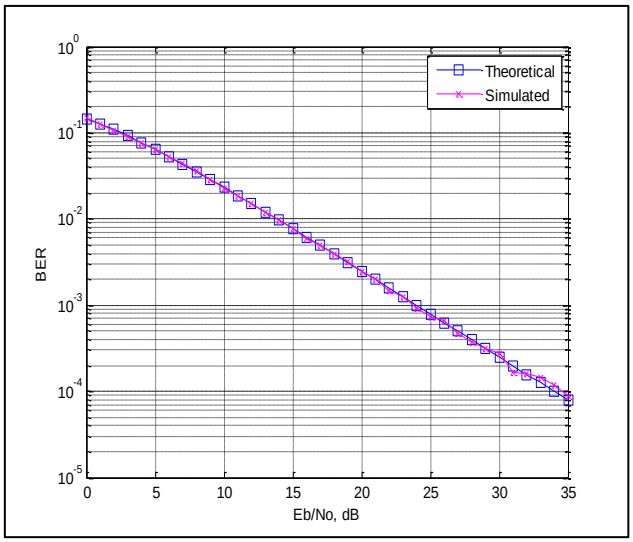



Figure 5: BER vs Eb/No characteristics for 10 branching nodes

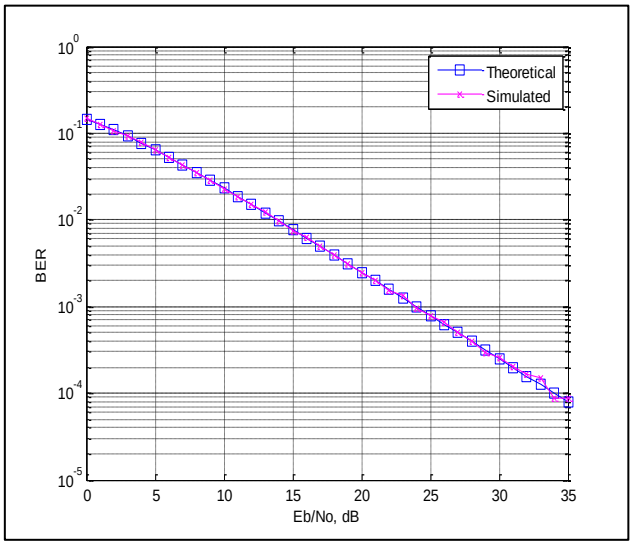



Figure 6: BER vs Eb/No characteristics for 15 branching nodes

This model has been evaluated by considering different number of branching nodes (points of scattering or reflections), whose value is less than the cyclic prefix duration used for the simulation. The simulation results obtained are in good agreement with the theoretical results. Thus, we conclude that this model approximates the PLC channel in a very satisfactory manner. This model is geared towards the realization of universal broadband provision that is powered by the ubiquitous power network (PLC channel).

## ACKNOWLEDGEMENT

The authors are grateful for the support received from ESKOM through the EPPEI program.

## REFERENCES

[1] L.T Berger, A. Schwager and J.J. Escudero-Garzas, "Power Line Communications for Smart Grid Applications," *Journal of Electrical and Computer Engineering*, Vol. 2013, Article 712376, 16 pages.

[2] H.C. Ferreira, H.M Grove, O. Hooijen and A.J. Han Vinck, "Power Line Communications: An Overview," *IEEE AFRICON 1996*, 24-27 September 1996, Stellenbosch, pp. 558-563.

[3] S. Galli, A. Scaglione and K. Dostert, "Broadband is Power: Internet Access through the Power Line Network," *IEEE Communications Magazine*, pp. 82-83, 2003.

[4] F. Zwane and T.J.O Afullo, "An Alternative Approach in Power Line Communication Channel Modelling," *PIER C*, Vol. 47, pp. 85-93, 2014.

[5] C. T. Mulangu, T. J. Afullo and N. M. Ijumba, "Modelling of Broadband Powerline Communication Channels", *SAIEE*, Vol. 102 (4), pp. 107-112, December 2011.

[6] A.G. Lazaropoulos, "Broadband Transmission Characteristics of Overhead High-Voltage Power Line Communication Channels," *PIER B*, Vol. 36, pp. 373-398, 2012.

[7] A.M. Nyete, T.J.O. Afullo and I. Davidson, "Performance Evaluation of an OFDM-based BPSK PLC System in an Impulsive Noise Environment," *To appear in the 35th PIERS Conference Proceedings, 25-28, August 2014, Guangzhou, China (Accepted).*

[8] E. Biglieri, "Coding and Modulation for a Horrible Channel," *IEEE Communications Magazine*, May 2003, pp. 92-98.

[9] C. T. Mulangu, T. J.O. Afullo and N. M. Ijumba, "Estimation of Specific Attenuation due to Scattering Points for Broadband PLC Channels", *PIERS, Malaysia, March, 2012*, ISSN: 1559-9450, pp. 92-96.

[10] C. Hensen and W. Schulz, "Time Dependence of the Channel Characteristics of Low Voltage Power-lines and

its Effects on Hardware Implementation," *AEU International Journal of Electronic Communication*, Vol. 54, No. 1, pp. 23-32, Feb 2000.

[11] N. Andreadou and F. Pavlidou, "Modeling the Noise on the OFDM Power-Line Communication System," IEEE Transactions on Power Delivery, Vol. 25, No. 1, January 2010.

[12] H. Philipps, "Modeling of Power Line Communications Channels", *Proceedings of the 3rd International Symposium on Power Line Communications and Applications*, 1999, Lancaster, U.K. pp. 14-21.

[13] H. Philipps, "Performance measurements of powerline channels at high frequencies," *in Proc. 1998 Int. Symp. Powerline Communications and its Applications, Tokyo, Japan, March 1998, pp. 229-237.*

[14] M. Zimmerman and K. Dostert, "A Multi-Path Signal Propagation Model for the Power Line Channel in the High Frequency Range," *Proceedings of the 3rd International Symposium on Power Line Communications and Applications*, 1999, Lancaster, U.K. pp. 45-51.

[15] Zimmermann M. and K. Dostert, "A Multipath Model for the Power Line Channel", *IEEE Transactions on Communications*, Vol. 50, No. 4, pp.553-559, April 2002.

**Abraham M. Nyete** obtained his BSc. degree in Electrical and Electronic Engineering from the University of Nairobi, Kenya in 2007 and the MSc. in Electronic Engineering from the University of KwaZulu-Natal, Durban, South Africa, in 2014. He is currently a postgraduate student in the Department of Electrical, Electronic and Computer Engineering at the University of KwaZulu-Natal. His research interests include radio wave propagation, wireless communications and power line communications.

**Thomas J.O. Afullo** obtained his Bachelors degree in Electrical and Electronic Engineering (Hon) from the University of Nairobi in the year 1979, the MSEE from the University of West Virginia, USA in year 1983 and the licence in technology and PhD in Electrical Engineering from Vrije Universiteit (VUB), Belgium in the year 1989. He is currently a Professor and Discipline Academic Leader of Electrical, Electronic and Computer Engineering at the University of KwaZulu-Natal, Durban, South Africa.

**Inno Davidson** holds the BSc. Eng and MSc. Eng degrees from the University of Ilorin, Nigeria and PhD in Engineering from the University of Cape Town, South Africa. He is currently the Director/Research Coordinator at the Eskom CoE and EPPEI SC in HVDC Engineering at the University of KwaZulu-Natal.

# Development of an improved routing metric based on IBETX Metric for Wireless Ad-hoc Networks.

M. S. Kabiwa[1], K. Djouani[1, 2] and A. Kurien[1]
Department of Electrical Engineering / FSATI
Tshwane University of Technology[1], P. O. Box 3211, Pretoria 0001
Tel: +27 12 382 5911, Fax: +27 12 382 5114
and LISSI Lab, Paris, France
University of Paris-East Creteil[2]
Email: maximekabiwa@gmail.com; djouanik@tut.ac.za and kurienAM@tut.ac.za

**Abstract— In networking, a routing metric provides quantifiable values that can be used to judge the cost or the efficiency of a specific route. This paper presented an enhanced metric based on the IBETX routing metric. It tackles the issue of interflow interference by considering the interference arising at the physical layer. Simulation results over a discrete event simulator show that at high traffic rate, the proposed metric gives 6% higher throughput than the IBETX metric and 11% higher than ETX metric. It also succeeds to reduce the normalised routing load by 12% compared to the IBETX metric and 11% to the ETX metric. In terms of average end-to-end delay, it produces 6% less delay than the original IBETX metric and 8% less delay than the ETX metric.**

**Index Terms—Wireless ad-hoc networks, routing metric, interference, ETX, IBETX**

## I. INTRODUCTION

Ad-hoc wireless networks consist of mobile battery-powered devices that attempt to communicate without exploiting any fixed infrastructure and pre-determined organization of available links. Each node must discover which peers are available for direct connection and establish communication paths along the network. Messages are relayed or routed over multiple wireless hops to reach their destination. In order to have acceptable performance from underlying ad-hoc wireless networks, the routing protocol plays a key role. It is the method to direct communication through the network. The network routing process involves two steps: the first considers the assigning of cost metrics to links. The second provides routing information in the network [1]. Optimal paths are determined based on routing metrics; they play an important role in characterizing the links as they obtain the information that will be used to make routing decisions.

For decades, most routing protocols used the strategy of finding the shortest path between the source and the destination in the network. This encouraged the use of minimum hop-count for route selection in wireless ad-hoc routing protocols. There are other deep-seated reasons which supported such a choice. The first one was mobility; mobility has always been a major research area in Mobile Ad-hoc Networks (MANETs).The selection of the shortest path between a source-destination pair reduced the probability of route breakage due to mobility of an intermediate node. The second one was energy consumption. In MANETs; the longer the route from the source to the destination, the more hops data has to cross which consequently impacts on the energy consumption. In static wireless ad-hoc or mesh networks in which nodes have low mobility or are stationary, energy consumption and mobility are not valid concerns. In fact, selecting the shortest path between a pair of nodes is a good choice only if every wireless link in the network has the same characteristics. However, this is not the case in reality; there can be huge differences among links (in terms of link latency, link interference and link capacity or link loss ratio) and hence, using the minimum hop count for quality based applications may not be a good choice. These critical disadvantages of the minimum hop count metric motivated the need of better performing link or cost metrics that can take into consideration multiple factors influencing wireless networks. In this paper, it is proposed an enhanced metric based on the Interference and Bandwidth adjusted Expected Transmission count (IBETX) [2] routing metric.

The remainder of this paper is organised as follows. In Section 2 a brief review of different interference models is presented. An enhanced routing metric based on the IBETX metric is proposed in Section 3. In Section 4, the simulation environment, results and their analysis are presented. Conclusions as well as recommendations for future work are drawn in Section 5.

## II. RELATED WORK

The performance of wireless networks is highly dependent on the amount of interference experienced by the wireless links. Wireless networks suffer much from interference due to the absence of dedicated bandwidth of wireless links and the shared nature of the wireless medium. Interferences degrade wireless networks by their strong contribution to the generation of wireless losses. They also increase the contention level when an interfering signal is above the carrier-sense threshold, causing frame emission to be delayed or lost until the channel is clear. This poor feature of wireless networks necessitates a prudent modelling of interference to be used in several design areas like channel

assignment, routing etc. Understanding and managing interference is essential to the performance of wireless networks. There are several interference models that have been proposed in the literature and used in transmission scheduling studies. The range-based or protocol interference model [3] is based on transmission and interference ranges. The transmission range is the maximum range where a radio signal can be properly received and the interference range is defined as the area where a sending node can disturb the transmission from a third node. The challenge is relevant to the complexity of measuring such ranges in real world, especially in ad-hoc wireless networks where the topology is unpredictable. This model is also very strict since it was designed to guarantee that the links do not interfere with each other through the particular channels assigned for each one. However, the limited number of available channels in 802.11 IEEE standard physical specifications do not allow one channel to be assigned to each wireless link. For these reasons, the range-based model cannot be used in the routing metric to capture interference. The logical Interference [4] is based on the Carrier Sensing Multiple Access with Collision Avoidance (CSMA-CA) from Media Access Control (MAC) which requires the station to wait until the channel is free before starting the transmission, once that the shared channel maybe occupied by transmissions from other nodes that are using the same channel within the interference range. This model is less restrictive than the range-based model. But It may not be suitable to meet high traffic demand of wireless ad-hoc networks. The reason is that CSMA-CA is a very conservative mechanism due to the combination of carrier sensing and collisions avoidance techniques, several nodes in the network are silenced when a certain communication is ongoing. Both logical and range- based interference model are computed before the actual data transmission [5], they may not be an accurate indicator of the current channel interference and will result in performance deterioration in wireless networks. In [2] and [4] the authors proposed respectively the Interference and Bandwidth adjusted ETX (IBETX) and Expected Link Performance (ELP) routing metrics that tackle the issue of inter-flow interference. These metrics tackle the interference by considering the logical interference model and do not explicitly or neglect the interference arising at the physical layer.

## III. PROPOSED MODEL

### A. Interference Modelling

Based on the limitations of the previous models, a physical interference model is considered to capture the interference experienced by links in the network; this model is simpler and less restrictive compared to the protocol and logical interference models, since it relies solely on the interfering signal strength values, such as the Signal to Interference plus Noise Ratio (SINR) and the Signal-to-Noise Ratio (SNR). Furthermore, the physical model is more realistic as it does physically model the interferences by accounting for the physical quality of wave superposition where interfering signals superpose to produce the resulting signal and interference at the receiving side does not only count from neighbouring nodes, but also from all transmitting nodes in the same contention domain [6]. In contrast to the protocol and logical interference models that capture interference which occurs before transmission, the physical model has the immense advantage of displaying the actual transmission of the packet when the interfering signal may cause failed transmissions. The question is whether actual capacity improvement can be achieved by considering the physical interference model. Under the physical interference model [7], successful reception of a packet sent by a node *i* to node *j* depends on the signal quality perceived by the receiver *j* that is given by the Signal to Interference plus Noise Ratio (SINR) value and this value must be higher than a predefined threshold, with this default threshold called capturing threshold. The interference experienced by a packet for a node *i* for a link $l = (i, j)$ is defined as follows:

$$IR_{(i,j)} = \frac{SINR_l(i)}{SNR_l(i)} \tag{1}$$

where

$$SNR_l(i) = \frac{P_l(j)}{Noise_i} \tag{2}$$

$$SINR_l(i) = \frac{P_i(j)}{Noise_i + \sum_{k \in InterferenceSet[(i)-(j)]} \theta(k) P_i(k)} \tag{3}$$

Here $InterferenceSet[(i)-(j)]$ is the set of nodes that can interfere with node *i*, $P_i(k)$ is the signal power of a packet from node *k* at node *i*, and $\theta(k)$ is the normalized rate at which node *k* generates traffic over a period of time, it also represents the fraction of the transmitter *k*'s received signal power that is projected onto the signal space of user *i*. In addition, $\theta(k)$ depends on the modulation schemes, spreading codes and data rates of the user. $P_i(j)$ is the signal strength of a packet from node *i* at node *j* and $Noise_i$ is the ambient or background noise at *i*. Considering a bidirectional communication link $l = (i, j)$ for a *DATA/ACK* like communication, the interference experienced by the link *l* is defined as:

$$Link\_Interference_l = \min\left(IR_l(i), IR_l(j)\right) \tag{4}$$

To estimate the link interference, the link SINR is considered in contrast of the local SINR of the listen mechanism as used in the MIND metric [5] since it works locally at a node without considering the link. The two end nodes may have an asymmetric view of the channel which may introduce some inaccuracy. For a wireless link $(i, j)$, $SINR(i, j)$ denotes the SINR measured at node *j* for data received from *i*. Considering a link $(i, j)$ , the receiver *j* maintains the SINR of the frames received from node *i* for a sliding window of time. Node *j* periodically unicasts small packets (probes) to the node *i* which contain the average SINR of the frames received from node *i*. This technique will be more accurate because the SINR of frames is captured in both directions. It should be noted that the

proposed model does not fully capture sender-side interference which results in back off and increases the expected transmission time. SINR and SNR are widely considered as a good indicator of link as they provide measurements from physical layer. However, the variations in the physical layer are so fast (in order of microseconds) compared to the routing layer. These different time scale variations suggest that each layer should attempt to compensate for variation at the physical layer first.

B. Probability of Success of a link

This part of the metric aims at estimating the expected delivery ratio or the probability of success of the link. A node may need to retransmit a packet several times (at the link layer) due to repeated losses. This is an indication of poor quality and represents an inefficient use of resources. In this technique, periodically on a window of time, the ratio of probes successfully delivered from the current node to its neighbour called forward delivery ratio $d_{forward}$ which represents the delivery probability of data packets and the ratio of probes received from neighbours called the backward delivery ratio $d_{backward}$ which is the successful transmission probability of Acknowledgment packet measured by the data sender. For a link $l$ formed between node $i$ and node $j$ the probability of success is given by [2]:

$$\Pr obability\_of\_Success_l = d_{forward}^{(l)} \times d_{backward}^{(l)} \quad (5)$$

The routing metric should select links with higher delivery ratio or lower packet loss since minimising the number of transmission does not only optimise the overall throughput of the network, it does also minimise the total consumed energy if a constant transmission power level is considered [8].

C. Bandwidth Estimation

It is an important characteristic that the routing metric must consider the transmission rate of the wireless link at each hop since a link with high transmission rate takes a small amount of channel time while a link with low transmission rate or bandwidth takes longer time which means that the node will occupy the channel for a longer period of time and disturb other nodes that are transmitting in the neighbourhood. The aim of this component should help the routing metric to favour wireless links with higher transmission rates in order to achieve better performance. For a wireless link $(i, j)$ the bandwidth is given:

$$Link\_Capacity_{(i,j)} = Bandwidth_{(i,j)} \quad (6)$$

The proposed routing metric is defined for a wireless link $l$ as follows:

$$IBETX_{improved} = \frac{\Pr obability\_of\_Success_l}{Link\_Capacity_l} \times Link\_Interference_l \quad (7)$$

Then end-to-end path metric for a path $p$ is given by:

$$IBETX_{improved_P} = \sum_{l \in P} IBETX_{improved_l} \quad (8)$$

## IV. SIMULATION ENVIRONMENT AND RESULTS

This section provides the details concerning the simulation environment. The wireless network consists of 50 static nodes randomly placed in a square flat area of $1000m$ by $1000m$. In order to increase interference among contending nodes in the same transmission domain, the number of connections is chosen to be 30. The effective carrier sensing range is of $447m$. Each node maintains send buffer of 50 packets with drop-tail mode. All packets both data and routing are sent by the routing layer and are queued at the interface queue until the MAC layer can transmit them. The traffic sources transmit Constant Bit Rate (CBR) with User Datagram protocol (UDP) at the transport layer. The bandwidth provided to all the wireless links is 6 Mbps. Each simulation runs for 10 different topologies with identical traffic model for 900 s each. All the simulations have been done on an Intel(R) Core(TM) i5-2100 with 3.10 GHz processor and 4.00 GB Random Access Memory (RAM) over the operating system Ubuntu Pangolin 12.04.

Table 1: Traffic and topology parameters

| Parameters | Values |
|---|---|
| Network area | $1000m \times 1000m$ |
| Number of nodes | 50 |
| Random topology | 10 |
| Number of conncetions | 30 |
| Packet rate | 1, 2, 4, 6, 8, 10 (packet/s) |
| Packet size | 512 Bytes |
| Traffic type | CBR |
| Data rate | 6 Mbps |
| Simulation time | 900 s |

The original IBETX metric was implemented over a proactive routing protocol which is the Destination-Sequenced Distance Vector (DSDV) [9] routing protocol. It is a non-adaptive routing protocol where routes are preco-computed. It has the advantages to guarantee loop-free paths and higher efficiency in route discovery (low latency). However, network resources are unnecessarily consumed when the network is stable and congestion control is worst [10]. The packet delivery ratio dramatically decreases with network size [10]. Based on the previous drawbacks the proposed metric along with others metrics is implemented over the AODV [11] routing protocol. At the Media Access Control (MAC) layer, the simulator uses the Distributed Coordination Function (DFC) compliant with the enhanced IEEE 802.11a standard (802.11Ext) [12]. 802.11Ext standard introduces two new modules: Mac802.11Ext and WirelessPhyExt. These extensions are based on Mac802.11 and WirelessPhy, but did a major modification to the original code, aiming at a significantly higher level of simulation accuracy. There are different modulation schemes that can be used to modulate the data namely Binary Phase

Shift Keying (BPSK), Quadrature Phase Shift Keying (QPSK) and 64 Quadrature Amplitude Modulation (64-QAM). Using higher modulation schemes (QPSK or 64-QAM) will achieve very high data rate but this comes at a cost very high SINR, Complexity of the receiver and higher energy consumption. Based on these drawbacks, the BPSK modulation scheme for the raw data transmitted and received is used at the physical layer. The following table summarises the simulation parameters used at the physical and MAC layers in the performance evaluation.

Table 2: parameters at the physical and MAC layers

| Parameters | Values |
|---|---|
| MAC protocol | IEEE 802.11a Ext |
| Physical layer | WirelessPhyExt |
| Frequency | 5.2 GHz |
| Antenna | Omni-directional |
| Propagation model | Two-ray ground model |
| RTS threshold | 2346 |
| Transmission power | 0.001W |
| Noise floor | $2.512\times10^{-13}$ |
| SINR data capture | 10dB |
| Modulation scheme | BPSK |
| Routing protocol | AODV |

In this paper, three metrics are used to study their effects on the overall network performance. These metrics are normalised routing load, packet average end-to-end delay and network throughput. To examine the performance of the routing protocol with distinct routing metrics under various sending rate, the packet rate is varied from 1 to 10 packets per second. 50 nodes are randomly placed to form a static network. CBR traffic is randomly generated by 30 source-destination pairs with packet size of 512 bytes. The value is chosen due to smaller payload sizes penalize protocol if append source route to each data packet. All the simulations were conducted on a discrete event simulator (NS-2.34).

### *A. Throughput*

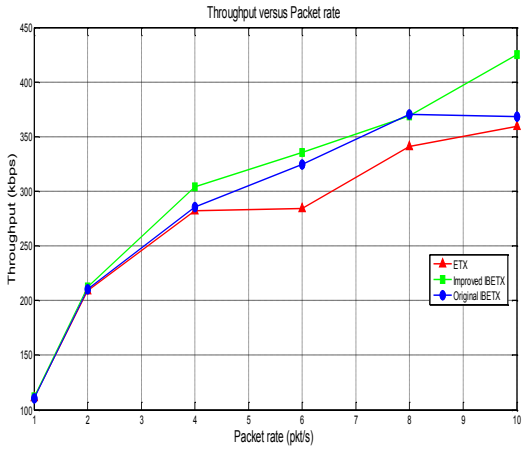



Figure 1: Throughput achieved for varying packet rate

In Figure 1, at low packet rate (below 2 packets per second) all the routing metrics have similar throughput. This is simply due to the fact that there is minimal interference in the network due to low sending rate and the choice of any route yields equally favourable results. As the sending rate increases, both the improved IBETX and the original IBETX routing metrics exhibit higher throughput compared to the ETX [13] metric. This is due to the fact on top of measuring the probability of success for data packet using probes (HELLO packets in this work) as the ETX metric, The improved metric and the original tackle the issue of bandwidth sharing mechanism in wireless networks (the links with lower bit rate degrade the performance of the faster links). Taking the bandwidth of all links in the same contention domain into account gives accurate information about the link status as compared to only considering the probability of success. As, HELLO packets are smaller in size as compared to the data packets, so, the idea of measuring the link quality by calculating the bandwidth of the link along with the probability of success are not sufficient. Routing metrics must incorporate the interference. The interference phenomenon takes places at the physical layer of the receiver terminal, as an interfering (undesired) signal disturbing the reception of a given desired signal. The original IBETX implements a logical interference model that rightly predicts the medium congestion and collision. This model is very conservative due to the fact that several nodes in the network are silenced when certain communication is ongoing and it occurs before transmission of the actual data packet. At high sending rate, it can be observed that the improved routing metric achieves 6% more throughput than the original IBETX and 11% more than the ETX metric in overall.

### *B. Normalised Routing Load*

Normalised routing load is an important performance metric for comparing these routing metrics, as it measures the scalability of a protocol, the degree to which it will function in congested or low bandwidth environments, and its efficiency in terms of consuming node battery power. Protocols that make used of routing metrics that send large number of routing packets can also increase the probability of collision and may delay data packets in the network interface transmission queues.

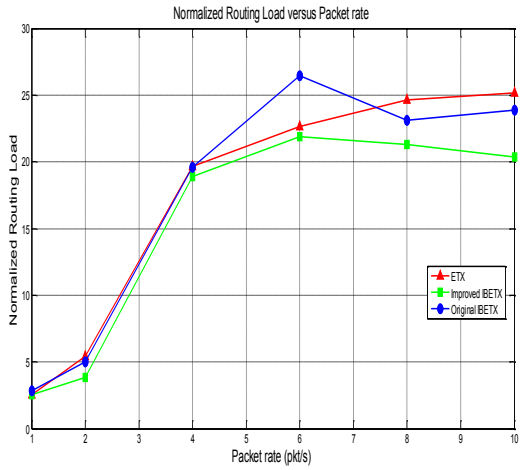



Figure 2: Normalised routing Load generated by varying packet rate

In Figure 2 as the sending rate increases the normalised routing load increases for all routing metrics. However, it is a bit low for the enhanced routing metric compared to the original IBETX and ETX routing metrics. Globally, the enhanced routing offers a normalised routing load reduction of 12% compared to the original IBETX routing metric and 11% to the ETX metric.

*C. Average end-to-end delay*

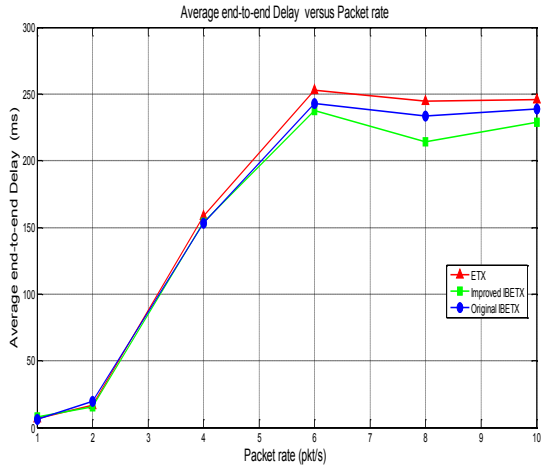



Figure 3: Average end-to-end delay generated for varying packet rate

Considering a stressful situation (above 4 packets per second), the enhanced routing metric offers a reduction of 5% in delay compared to the original IBETX routing metric and 8% to the ETX metric. It can also be observed that for all the routing metric

## V. CONCLUSION

In this paper, it has been proposed an enhanced routing metric based on the IBETX metric. As IBETX, the proposed metric is a hybrid and threefold metric. Firstly, it tackles the interference in wireless networks by probing the physical layer. Capturing interference at the physical is simpler, more realistic and less restrictive compared to the one at the MAC layer. It relies solely on the interfering signal strength such as SINR and SNR. These two parameters have the huge advantage of displaying the actual transmission of the packet. Secondly, it provides the transmission rate information to all nodes in the same contention domain by considering the bandwidth sharing mechanism of 802.11. And Finally, It bypasses congested regions in the network using expected delivery ratio.

The enhanced metric has been compared through extensive simulations over a discrete event simulator (NS-2) with two existing routing metrics (IBETX and ETX) and the results show performance improvement in terms of throughput , average end-to-end delay and normalised routing load in a moderate density and high data traffic rate network.

In this paper, it has been only considered the interference viewed by the receiver side; in the future it will be investigate ways to incorporate sender side interference. For all the simulations, a medium scale network for checking the proposed metric was considered. It will be very interesting to see the performance of the proposed metric in highly dense network.

## VI. REFERENCES


[1] B. Awerbuch, D. Holmer, and H. Rubens, "High Throughput Route Selection in Multi-rate Ad Hoc Wireless Mesh Networks", in wireless on demand network systems. Vol. 2928 Heidelberg: Springer Berlin, 2003, pp253-270

[2] N. Javaid, "Analysis and Design of Quality Link Metrics for Routing Protocols in Wireless Networks", PhD Thesis, University Paris-East, December 2010

[3] P. Gupta and P. Kumar, "The Capacity of Wireless Networks", In IEEE Transactions on Information Theory, 2000, pp388-404

[4] U. Ashraf, S. Abdellatif and G. Juanole,"An Interference and Link Quality Aware Routing Metric for Wireless Mesh Networks", In IEEE 68th Vehicular Technology Conference, 2008

[5] C. M. Borges, D. Pereira and M.Curado and E. Monteiro "Routing Metric for Interference and Channel Diversity in Multi-Radio Wireless Mesh networks", In Proceedings of the 8th International Conference on Ad-Hoc, mobile and Wireless Networks, 2009, pp43-49

[6] S. Jinzhao, C. Jing and W. Wei,"Gray Physical Interference Model based Link Scheduling Algorithms", Science China Information Sciences, Vol 55, 2012, pp1337-1350

[7] K. Jain, J. Padhye and V. Padmanabhan and L. Qiu, "Impact of Interference on Multihop Wireless Network performance", In Wireless Networks, 2005, pp471-487

[8] C. Koksal and H. Balakrsihnan,"Quality Aware Routing Metrics for Time Varying Wireless Mesh Networks", Selected areas in Communications IEEE Journal, Vol 24, 2006, pp1984-1994

[9] C. E. Perkins and P. Bhagwat, "Highly Dynamic Destination-sequenced Distance –Vector Routing (DSDV) for mobile computers," SIGCOMM computer communications. Rev., vol. 24, pp. 234-244, 1994

[10] M. S. Kabiwa, K. Djouani and A. Kurien, "Performance Evaluation of IBETX Routing Metric over DSDV Routing Protocol in Wireless Ad hoc Networks", 8th International Symposium on Intelligent Systems Techniques for Ad hoc and Wireless Sensor Networks (IST-AWSN 2013), 2013, pp1108-1115

[11] C. Perkins, et al. IETF RFC3561, "Ad hoc On-Demand Distance Vector (AODV) routing". Available online: http://www.ietf.org/rfc/rfc3561.txt

[12] Qi Chen et al "Overhaul of IEEE 802.11 Modeling and Simulation in NS-2" MSWIM'07 Proceedings of the 10th ACM Symposium on Modeling, Analysis and Simulations of Wireless and Mobile Systems, 2007, pp159-168.

[13] D. S. J de Couto, "High-throughput routing for multi-hop wireless networks", PhD Dissertation, MIT, 2004.



**Maxime Kabiwa** received a Bachelor of Science honours in Physics from University of Douala (Cameroon) in 2007 and a Bachelor of Technology in Electrical Engineering in 2010 From Tshwane University of Technology. He is presently a final year Master student at the same institution. His research interests include Wireless Networks, Signal processing, theoretical and mathematical Physics, and Optimisation.

# Enhanced Backoff Mechanism for the Traditional Carrier Sense Multiple Access with Collision Avoidance in a IEEE 802.11p VANET

IB Kam[1], K Djouani[1, 2], AM Kurien[1]
[1]F'SATI/Dept. of Electrical Engineering, Tshwane University of Technology, Pretoria, 0001, South Africa
[2]University of Paris-Est. Creteil, LISSI Lab, Paris, France
Email: barnelisme@gmail.com [1] ; {djouanik[1], kurienAM[1],}@tut.ac.za

***Abstract*—A Vehicular ad-hoc network (VANET) is characterized by a high mobility and the speed of its nodes. The traditional Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) may not efficiently handle the requirements of the proposed DSRC IEEE 802.11p as initially designed in terms of channel access delay and throughput. This paper proposes an enhancement to the backoff mechanism of the traditional CSMA/CA to achieve higher throughput. The method proposed considers the fact that vehicles on the road may have the same speed at exactly the same time but will always have different locations at specific points in time. It is therefore possible to derive a unique spacing parameter based on the speed and location of each node to space out the random backoff periods that are independently generated by each mobile node. The aim of this backoff modification is to decrease the probability of two nodes generating the same random backoff timer. By reducing the number of collisions that occur during the channel access process, a higher throughput could be achieved. Modifications are proposed at the Medium Access Control (MAC) layer which manages the channel access mechanism. Comparisons are done in terms of throughput, packet delivery fraction, drop MAC retry exceed count and end-to-end delay. The results obtained show that the proposed Speed and Location Based CSMA/CA (SL-CSMA/CA) performs better when compared to the traditional CSMA/CA in a vehicular ad-hoc environment.**

**Keywords—VANET; CSMA/CA; V2V, V2I**

## I. INTRODUCTION

Wireless Technologies have grown rapidly over the past years. It has become an important area of interest in the research society. Their use in different sectors of industry aims to improve the quality of life. The Transport industry domain has been recently benefiting from the adoption of wireless technologies. Intelligent Transportation Systems (ITS) were created to improve the quality of the transport industry. The improvement can be considered at different levels. One of the important aspects relates to safety. By providing a means of communication between transportation vehicles such as cars, ships or airplanes, disasters may be avoided or at least reduced. Vehicular ad-hoc networks (VANETs) have been created to allow cars on the road to share information in a Vehicle to Vehicle (V2V) communication system or in a vehicle to Infrastructure (V2I) communication system. In a vehicular environment, VANETs provide communication support to the ITS. The goal is to assist in accident avoidance, traffic management and enable intra-vehicles internet. The IEEE 802.11p amendment proposed to provide a Wireless Access in Vehicular Environment (WAVE) uses the Dedicated Short Range Communication (DSRC) band at 5.9 GHz [1]. The Federal Communications Commission (FCC) has allocated 75 MHz of spectrum at 5.9 GHz that will be used by IEEE 802.11p [2]. The amendment is mainly considered at the Medium Access Control (MAC) layer and physical layer (PHY). The standard must be able to handle a good Quality of Service (QoS) in an environment where mobile nodes move at high speeds and where the topology of the network may change rapidly [3].

DSRC IEEE 802.11p considers a 10 MHz OFDM (Orthogonal Frequency Division Multiplexing) channel spacing in contrast to the 20 MHz used by the IEEE 802.11a in order to double all OFDM symbols used in the IEEE 802.11a [4]. The MAC access mechanism that is originally proposed is the traditional Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) for V2V communication mode. The amendment in the IEEE 802.11a done for DSRC IEEE 802.11p at the MAC layer considers the creation of an environment that enables an efficient communication group setup with minimum overhead when compared to the typical IEEE 802.11 MAC layer. Due to the rapidly and ever changing topology of VANETs the MAC layer has to be simplified to allow fast and efficient communication set up. For this reason, authentication and association procedures are not required for a mobile node to join a network [5]. Nodes may join a BSS (Basic Service Set) by just responding to an advertisement message. This arrangement allows for the efficient transmission and reception of data [5].

VANETs are characterised by high velocities and mobility of its nodes. On the road, vehicles can join and leave the network at different speeds and approach from different

A work partially funded by the South African National Research Foundation.

directions creating an ever changing network topology. Individual vehicles can behave in a way that may affect the entire network in a chain reaction. For example, considering a scenario on the highway, when a car suddenly brakes from a high speed to a speed that is close to zero, cars that are following that specific vehicle will be forced to either rapidly change lanes or also decelerate. Furthermore, a car that was in communication with a neighbouring car may decide to accelerate and this could cause a communications break between them. An efficient network management system must be able to support the wireless communication set up for such environments.

The two types of communication in VANETs imply different network management systems. Vehicle to Infrastructure (V2I) establishes a communication between the On Board Unit (OBU) of a vehicle and a Road Side Unit (RSU) or Infrastructure as seen in [12]. In V2I, the RSU is in charge of the network synchronisation and time slot allocation. RSU may implement Time Division Multiple Access (TDMA), Frequency Division Multiple Access (FDMA) or other access techniques to synchronise the network access. In a V2V set up, the communication is directly established between the two OBUs of vehicles wishing to communicate. In contrast to the V2I set up, a V2V communication does not use a central synchroniser Infrastructure with network access and management being done by nodes. Therefore, the network needs to be self-organised.

Contention based techniques are used in ad-hoc networks where no central processing infrastructure is used. Contention based MAC access techniques make nodes wishing to transmit packets over the physical medium to contend for access. Thus the transmission is done in a first-come first-served basis. A Distributed Coordination Function (DCF) for sharing access to the physical medium based on the CSMA/CA protocol is defined by the IEEE 802.11 standard for WLAN. Important frames are used for collision avoidance. The goal of the CSMA/CA protocol is to maximize throughput by reducing the collisions due to the contending nodes that share the same channel [6]. In CSMA/CA, each node works independently and generates its own backoff timer [7]. If more than one node generates the same random number or if their backoff timer reach zero at the same time, they will transmit at the same time and a collision will occur. This paper proposes an enhancement to the backoff mechanism of the traditional CSMA/CA to achieve higher throughput. The rest of the paper is organised as follow: Section II presents an overview of related works on MAC access techniques. Section III presents the modelling of the proposed method. In section IV, results are discussed and the paper is concluded in section V.

## II. RELATED WORK

Fairness, channel access delay and synchronisation are major challenges in IEEE 802.11p WAVE. The use of RSU makes it much easier to manage the network but suffers from environmental and financial constraints. If a large number of RSUs have to be deployed along the roads, the large financial implication needs to be considered in the placing of the RSUs.

In [7], taking into account the space used to cover the RSU, the bandwidth is divided into overlapping clusters. In each cluster, the channel is divided into time slots. A cluster's ID is assigned to each cluster. After entering the area covered by the RSU, a vehicle that has received a beacon message from the Infrastructure will work out its relative position to the RSU and request to register under the RSU's control. The ID of the cluster in which the vehicle is located determines the priority or channel allocation method. The authors in [7] further highlight that the cluster size affects the throughput drastically and that the protocol may not scale to high vehicular traffic and would not work in areas where there are no RSUs.

A position based MAC mechanism is proposed in [8]. Contention based and contention free mechanisms are used in combination for channel access: Contention Based Phase (CBP) and Contention Free Phase (CFP). The traditional CSMA/CA function is used by each node in the CBP before accessing the network. Once a node has successfully gained access to the network, it enters its CFP. The CFP is done by the aid of a RSU that remembers and registers a successful node in the network. From this point, the channel access becomes deterministic. The RSU unit also determines when nodes must start contending for the channel in the CBP and when to start their CFP for those who get access after the CBP. The frame period of the CBP and CFP is referred to as a SF (Superframe) [8]. Through heartbeat messages, each vehicle also sends its location to the RSU so as to enable the RSU to prioritise channel access based on the location of vehicles compared to road intersections and areas registered as potentially dangerous or of high priority. In this method, the traditional CSMA/CA is used with all its drawbacks. Furthermore, financial constraints do not allow the seamless deployment of RSU along the roads [8].

The proliferation of RSU may be avoided by the use of self-organised networks. A self-organised network would make it possible for the nodes to manage the network themselves. Contention based mechanisms like CSMA/CA allow nodes to contend for the network access and avoid RSUs [13]. However, they require and demand very efficient MAC access algorithms to create a fair network with acceptable channel access delays.

In [9], it is highlighted that the Automatic Identification System (AIS) as used in the shipping industry is a good technique to be applied in WAVE. The AIS has ships as mobile nodes. Every ship broadcasts its location, speed, size and direction to other ships to avoid accidents. The time is divided into time slots and nodes (ships) use the Self-Organising Time Division Multiple Access (STDMA) algorithm to select their slots. When the network is heavily loaded, nodes wishing to access the network steal slots from the furthest located node based on the information received on the regular broadcast messages to avoid interference with neighbouring nodes. To decrease the possibility of having two nodes with the same slot to move closer to each other, the occupation of time slots is done for a limited period of time before it can be released [9]. In the simulations conducted in this paper, nodes are not allowed to overtake each other. All the nodes move at the same speed and in two directions only.

In this paper, an enhancement of the back-off algorithm of the traditional CSMA/CA is introduced to reduce the collisions that occur at channel access when nodes contend for the channel. A spacing parameter is derived from the speed and location of each node to generate a random back-off delay that is different from those that are generated by close neighbours.

## III. SPEED AND LOCATION BASED CSMA/CA (SL-CSMA/CA)

The objective of the SL-CSMA/CA protocol is to enhance the MAC layer CSMA/CA strategies and to increase throughput and packet delivery fraction. This is achieved by reducing the probability of MAC collisions. In the traditional CSMA/CA as originally designed, nodes transmit their packet when their back-off timer reach zero. The probability of two nodes generating the same back-off number exists even though it could be low [10]. The method considered in this work takes advantage of the unique properties of each vehicle on the road that is contending for channels at a specific point in time Ti. The fact that no two car can have exactly the same speed and location at exactly the same time shows that at a point in time $T_i$ every node has a unique property that is different from its neighbours.

The scenario of two nodes selecting the random number in the Contention Window (CW) can be compared to the scenario of two fair dices that are flipped at the same time. These are two independent events that can generate the same outcome. A collision occurs when the outcomes are the same. This proposed technique assumes that every mobile node is equipped with a GPS (Global Positioning System) that can display the location and speed of the node.

Assuming that two mobile nodes A and B are contending for channel access in a WAVE, let P be the probability that A and B choose the same random back-off timer in the CW. It is assumed that mobile nodes A and B are respectively traveling at speeds of $v_A$ and $v_B$ and are respectively located at positions $A_{xy}$ and $B_{xy}$ at time $T_i$. If A generates a random number $r_A$ and B generates $r_B$ at $T_i$. Then, there is a number $0 \leq x \leq 1$ such that

$$P\big(r_A(T_i) = r_B(T_i)\big) = x \tag{1}$$

The speed based channel access algorithm tries to find a number $\hat{x} \cong 0$ such that

$$P\big(\widehat{r_A}(T_i) = \widehat{r_B}(T_i)\big) = \hat{x}, \tag{2}$$

where

$$\widehat{r_A}(T_i) = r_A(T_i) + \delta_A(T_i) \tag{3}$$

and

$$\widehat{r_B} = r_B(T_i) + \delta_B(T_i). \tag{4}$$

$r_A, r_B \in [0, W]$. All random numbers are chosen from the current CW. $\delta$ is the Spacing parameter derived from the speed and location of each mobile. It transforms equation (1) to equation (2). The method is based on the fact that in a network where no accident has occurred, $P\left(A_{xy}(T_i) = B_{xy}(T_i)\right) = 0$. This is the probability that the two nodes be at exactly the same position at the same time. This method assumes that every vehicle is equipped with a GPS with good precision in a way that close neighbours do not make the false assumption that they are at the same position.

In the traditional CSMA/CA, every node chooses a random back-off timer $r$ such that $0 < r < W - 1$, where $W$ is the maximum number in the contention window. $W$ doubles when a collision occurs on the medium. The developed SL-CSMA/CA considers the spacing parameter $\delta$ such that

$$\delta = |W - tan \propto -v| \tag{5}$$

$\tan \propto$ is the positive trigonometric tangent of the angle formed by the mobile's coordinates on a Cartesian map with respect to the origin. Values of $\frac{\pi}{2}$ must be discarded for all $\propto$ to avoid infinite values of the tangent. In a real life scenario, the GPS coordinates would have to be transformed to Cartesian coordinates. The position of nodes are simulated in a two dimensional map in the NS2.34 simulator. Distances are expressed in meters and all values are kept within the CW range. In a case where cars have almost the same velocity and are all confined in the same

side of a road, the accuracy of the GPS will be of high importance in differentiating the positions of the vehicles.

$\boldsymbol{v}$ is the normalised speed of the node. It is obtained by dividing the current node's speed by lowest speed allowed on a highway which is sixty kilometres per hours (60km/h) in the case of South Africa where the study is currently conducted.

$$v = \frac{v_{node}}{v_{min}} \quad (6)$$

$\boldsymbol{W}$ is the maximum value in the current window size that doubles when a collision occurs.

The spacing parameter $\delta$ at time $T_i$, is then added to the random number $r$ at time $T_i$ such that

$$\hat{r}(T_i) = r(T_i) + \delta(T_i) \quad (7)$$

The purpose method of this method is to reduce the probability of having equal random numbers uniformly chosen in the CW. In this method, the W is dynamically adjusted depending on the location and speed of each vehicle. This increases the dispersion of numbers chosen for the backoff period. The randomness may still lead to equal backoff value but with reduced probability of occurrence.

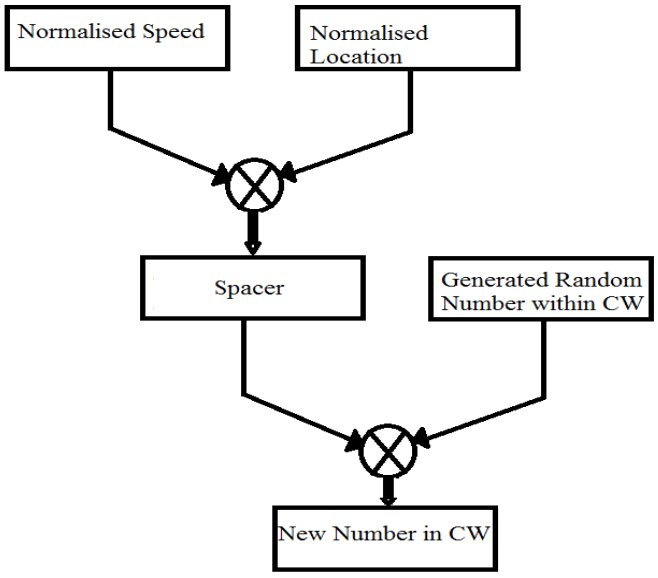



**Figure 1** – Overall Process in Generating a new Random Number

The performance of the traditional CSMA/CA is compared with SL-CSMA/CA in terms of throughput, packet delivery fraction, Drop MAC Retry Exceed Count and end-to-end delay. The results of the comparison are presented in the next section.

## IV. Results and Discussion

### A. Simulation Setup

The traditional CSMA/CA is first implemented as the MAC access mechanism in a wireless and mobile node scenario. Its back-off mechanism is then updated to fit the design of the SL-CSMA/CA. The simulations were conducted in the network simulator NS2.34. The operational frequency of 5.9 GHz was used with a 10 MHz channel spacing as per the IEEE 802.11p standard. The sensitivity of the wireless interface was set to -85 dBm. Traffic was generated between mobile nodes that are moving in different directions at different speeds. Line of sight between communicating nodes is assumed to represent a highway scenario. The range of speeds is from 60 km/h to 90 km/h in a topology area of 5000 x 5000 m. The propagation model used was the Two-Ray ground model. Small TCP packets of size 512 bytes are used in all connections for simplicity and to keep all data transmissions identical. Further works consider increasing packets sizes and analysing the deterioration of the performance. AODV is used as routing protocol. The transmission range is set to 500 m. All simulations were run for a period of 200 for node numbers of 10, 20, 30, 40, 50, 60, 70, 80, 90 and 100. A total of 10 simulations were run for the traditional CSMA/CA as well as for the proposed SL-CSMA/CA. The described simulation set up was to simulate a South African highway scenario. Nodes that overtake each other can move in different directions causing communication breaks. This explains the trends in the graphs which are not uniform.

### B. Throughput

The graph below presents the results obtained and highlights the performance achieved in terms of throughput when simulating the traditional CSMA/CA and the developed SL-CSMA/CA. As the number of nodes increases by steps of 10, the throughput of the SL-CSMA/CA grows faster than the traditional CSMA/CA. The throughput's mean average of the Traditional CSMA/CA lags that of the proposed SL-CSMA/CA by 7.3%.

The maximum throughput achieved in the simulations was higher than 2 Mbps because this is not an exclusive data packet throughput. Headers, preambles, link layer acknowledgements and interframe spaces were not excluded in the calculation of throughput.

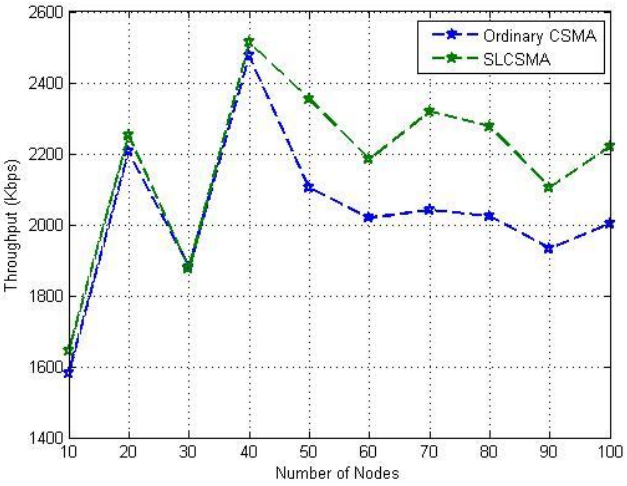



**Figure 2** – *Throughput of SL-CSMA/CA versus CSMA/CA*

### *C. Packet Delivery Fraction*

Packet Delivery Fraction (PDF) is the ratio of the number of received packets to the number of sent packets at the MAC layer and it does not have a unit and is expressed as a percentage. As the number of nodes increases, results show that both PDF lines drop below 75. SL-CSMA/CA still outperformed the CSMA/CA in terms of PDF. The sudden drops as observed at 30 nodes point are due to the mobility design that is used where communication breaks may occur. Nodes in communication may suddenly experience a break in communication requiring a new setup due to the different speeds and directions taken by each node.

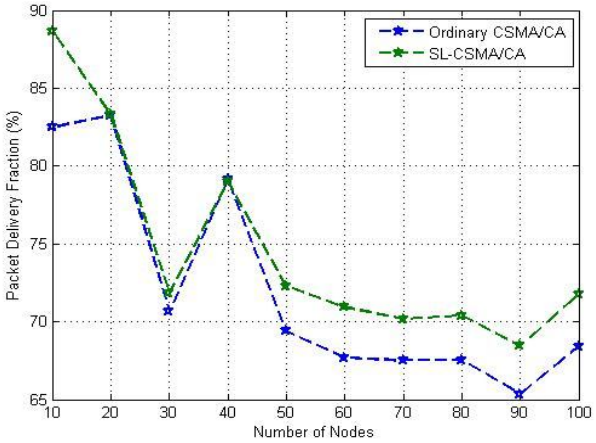



**Figure 3** – *Packet delivery Ratio of SL-CSMA/CA versus CSMA/CA*

### *D. Drop MAC Retry Exceeded Count*

The Drop MAC Retry Exceeded Count represents the number of tries and retries that fail causing a drop of packet after the counter has reached its maximum allowed value when attempting to access the channel. A lower number of try and retry failures are experienced in the developed SL-CSMA/CA than in the traditional CSMA/CA. A high level of interference would cause a high number of Drop MAC Retry Exceed Count. Therefore, a smaller number of attempt failures may imply that there is less interference in the MAC access process. The data represented are for the entire simulation time and for all the nodes in the topography. In IEEE 802.11p, the accepted MAC Retry Exceeded Count is 7 [11].

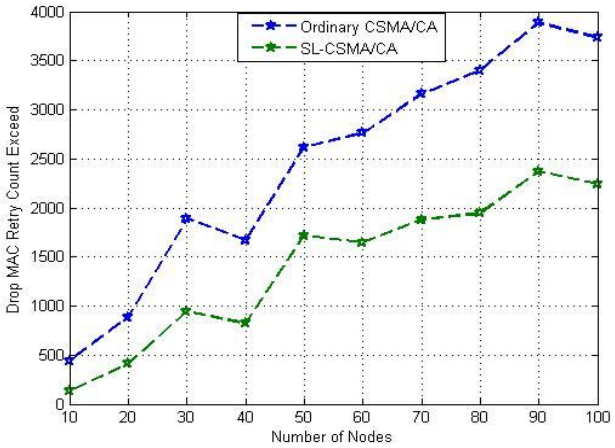



**Figure 4** – *Drop MAC Retry Exceeded Count of SL-CSMA versus CSMA/CA*

### *E. End-to-End Delay*

The graph below presents the performance achieved in terms of end-to-end delay when simulating the traditional CSMA/CA and the developed SL-CSMA/CA. From the plotted curves, it can be observed that not much was gained in terms of end-to-end delay when comparing the two methods. The two methods performed relatively similarly in terms of end-to-end delay.

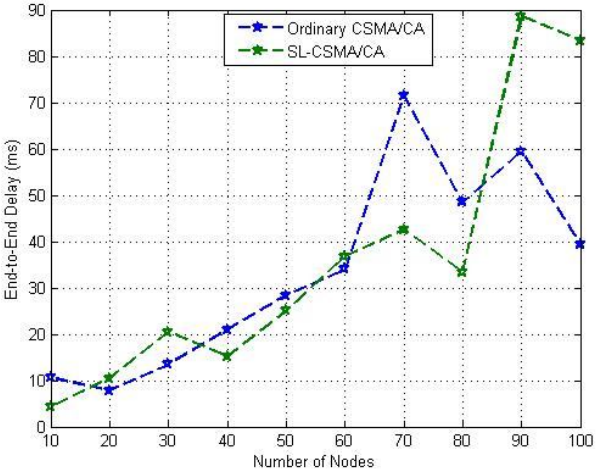



**Figure 5** – End-to-end *Delay of SL-CSMA/CA versus CSMA\CA*

## V. CONCLUSION

This paper presented the development of an enhancement of the back-off mechanism of the traditional CSMA/CA. Due to the fact that the nodes generate their random number independently from the CW, there is a probability that more than one node generates the same number and that their counter reaches zero at the same time causing a transmission collision. The work presented enhances the back-off mechanism in such a way that the probability of choosing the same random number by close neighbours is decreased and brought closer to zero. A spacing parameter is derived from the speed and location of each node and combined to the generated random back-off to create new random back-off timers. Simulation results have shown the performance of the developed SL-CSMA/CA over the CSMA/CA in terms of throughput, packet delivery fraction and Drop MAC exceed count. However, it was shown that end-to-end delay did not improve with the SL-CSMA.CA. Further work will consider the reduction of delays at channel access.

## REFERENCES


[1] Q. Xu, D. Jiang, R, Senguptat, D. Chrysler, "Design and Analysis of Highway Safety Communication Protocol in 5.9 GHz Dedicated Short Range Communication Spectrum." Vehicular Technology Conference, VTC 2003-Spring. The 57th IEEE Semi-annual. vol.4, 2003 pp 2451 - 2455.

[2] R. Saeed, M. Abakar, A. Hassan, and O. Khalifa, "Design and evaluation of lightweight IEEE802.11p-based TDMA MAC method for road sideto-vehicle communications," in Computer and Communication Engineering (ICCCE), 2010 International Conference on, May 2010, pp.1–5.

[3] IEEE Computer Society, "IEEE Standard for Information technology--Telecommunications and information exchange between systems--Local and metropolitan area networks--Specific requirements, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, Amendment 6: Wireless Access in Vehicular Environments", July 2010

[4] D. Jiang, L. Delgrossi "IEEE 802.11p: Towards an International Standard for Wireless Access in Vehicular Environments" Vehicular Technology Conference 2008 pp 2036 - 2040

[5] Bo Li, MS. Mirhashemi, X. Laurent, Jinzi Gao, "Wireless Access for Vehicular Environments," Internet: http://www.mehrpouyan.info/Projects/Group%205.pdf, [Apr. 14, 2014].

[6] X. Wang and K. Kar, "Throughput Modelling and Fairness IssuesIn CSMNCA Based Ad-Hoc," NetworksINFOCOM 2005. 24th Annual Joint Conference of the IEEE Computerand Communications Societies. Proceedings IEEE (Volume:1 ),March 2005vol. 1 ISSN :0743-166X pp:23 - 34

[7] R. Singh , "RSU Centric Channel Allocation in Vehicular Ad-Hoc Networks," Wireless Communication and Sensor Networks (WCSN), 2010 Sixth International Conference. Allahabad 15-19 Dec. 2010.pp 1 – 6.

[8] A. Böhm, M. Jonsson, "Real-Time Communication Support for Cooperative, Infrastructure-Based Traffic Safety Applications," International Journal of Vehicular Technology, vol. 2011, Article ID 541903, 2011. doi:10.1155/2011/541903, pp. 17

[9] K. Bilstrup, E. Uhlemann, EG. Strom, and U. Bilstrup, "On the Ability of the 802.11p MAC Method and STDMA to support Real-Time Vehicle to Vehicle Communication," EURASIP Journal on Wireless Communications and Networking, Vol. 2009, No. Article ID 902414. Mar 2009 .

[10] M. Bertocco, G. Gamba, A. Sona"Is CSMA/CA really efficient against interferencein a Wireless Control System? An experimental answer,"Emerging Technologies and Factory Automation Conference, 2008, E-ISBN :978-1-4244-1506-9 pp 885-892

[11] W. Alasmary, W. Zhuang, "Mobility impact in IEEE 802.11p infrastructureless vehicular networks", Recent Advances in Analysis and Deployment of IEEE 802.11e and IEEE 802.11p Protocol Families Vol. 10, Issue 2, March 2012, pp 222–230

[12] R. Saeed, M. Abakar, A. Hassan, and O. Khalifa, "Design and evaluation of lightweight IEEE802.11p-based TDMA MAC method for road sideto-vehicle communications," in Computer and Communication Engineering (ICCCE), 2010 International Conference on, May 2010, pp.1–5.

[13] Y. LIU, Y. WANG, S. CHEN2, X. LI and Z. RAO, " A Hybrid MAC Mechanism for Multiple Load Intelligent Vehicle Transportation Network," International Journal on Smart Sensing and Intelligent Systems Dec 2011 Vol. 4 Issue 4, pp 662



**About the Author**

Ifer Barbana Kam obtained the Bachelor of Technology in Electrical Engineering light current at the Tshwane University of Technology in 2010. He is a Master's candidate in the field of Telecommunication the Tshwane University of Technology. His fields of interests are Vehicular Ad-hoc Networks and MAC protocols in Wireless Networks.

# Capacity Performance Analysis in MIMO Vehicular Networks

F Nyongesa, K Djouani and Y Hamam
Department of Electrical Engineering
Tshwane University of Technology
Private Bag X680, Pretoria 0001
Tel: +27 12 382 4809 or +254 722825139
Email: fcnyongesa@yahoo.com, djouani@gmail.com, HamamA@tut.ac.za

Abstract- **Vehicle-to-vehicle (V2V) and Vehicle-to infrastructure (V2I) communication supporting Intelligent Transportation Systems (ITS) provides a paradigm shift in road transport and promises safety, comfort and convenience for drivers and passengers. V2V communication is a wireless technology implemented in IEEE 802.11p/WAVE protocols supporting a medium access control (MAC) and physical layer that are constrained to provide low latency and high data rate transmission for time-critical safety emergency messaging and bandwidth-hungry multimedia streaming for infotainment. The standard incorporates multiple-input multiple-output (MIMO) technology to leverage the shortcomings of the single antenna systems but the extent to which MIMO achieves capacity performance in the setup remains an open subject. This paper presents capacity performance analysis results of 2x2 and 4x4 MIMO systems with reference to the single-antenna systems and draws a generalization of MIMO performance in Vehicular networks.**

Index Terms— **ITS, MIMO, vehicle-to-vehicle, channel capacity, space-time coding**.

## I. INTRODUCTION

Vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication implemented in vehicular ad hoc networks (VANETs) has recently attracted interest of researchers concerned with Intelligent Transportation System (ITS) [1]. ITS is concerned with safety, security and efficiency of road transport systems and many countries have already implemented some ITS applications such as traffic management, traveller information, emergency management, statutory electronic payment and security-safety oriented applications. When fully realized V2V communication will significantly reduce collisions among vehicles and save lives.

Implemented in wireless communication technology, V2V communication utilizes IEEE 802.11p standard for its medium access control (MAC) and physical layer (PHY) functions in the lower protocol stack and IEEE 1609x (WAVE- Wireless Access in Vehicular Environment) for upper layer functions for data and management planes [2]. The real-time constraints imposed on the vehicular wireless channel of low latency and high throughput sets a requirement for a robust wireless link that experiences low BER with large bandwidth capacity. A technique which promises high data rates and reliability in VANET is multiple-input multiple-output (MIMO) based on IEEE 802.11p/WAVE protocols. Operating at 5.850-5.925 GHz, WAVE systems adopt orthogonal frequency division multiplexing (OFDM) and achieve data rates of 3 – 27 Mbps and MIMO technology provides further improved transmission quality and achieves higher data rates in multipath environments without the need for extra spectral resources or energy [3]. However, the channel suffers from impairments such as pathloss, multipath and shadowing, factors that degrade both the signal quality and the data rate. In multipath, each component experiences different attenuation, phase shift, angle of arrival, Doppler shift and time delay that add up to degrade the channel signal-to-noise ratio (SNR). Multipath has been extensively studied such as in [4] and [5] whereas Doppler shift has been investigated in [6] and [7]. Consequently, for MIMO technology to deliver optimum results intervention techniques are mandatory not only to enhance system capacity but also system reliability.

Spatial multiplexing and diversity schemes have been deployed to leverage channel capacity and reliability under multipath fading environment. In [8] space-time block coding (STBC) was introduced as a new scheme employing multiple antennas in Rayleigh fading channels and tremendously increased the multiplexing and diversity gain in MIMO systems while maintaining receiver simplicity. In [9] space-time block coding (STBC) is deployed in MIMO system to leverage diversity that significantly enhances the BER performance. It is further demonstrated that the diversity order and choice of constellation scheme have an effect on the obtained signal quality where better BER performance was obtained in high diversity order and low signal constellation configuration. In the review presented in [10] the Bell-Lab Layered Space-Time (BLAST) and Space-Time Coding (STC) systems are analyzed for multiplexing and diversity gain in MIMO systems. The authors observed that whereas BLAST system significantly increased the system capacity STBC achieved both capacity and diversity gain provided the signal was properly pre-coded. They viewed the STC schemes to a large extent as a tradeoff between the conflicting three goals of maintaining a simple decoding, maximizing error performance, and maximizing the information rate. Although a number of research activities have been conducted on multiplexing gain in MIMO systems as evidenced by the preceding literature,

none has produced a comparative performance based on the order of the MIMO channel.

This paper undertakes the capacity performance analysis of multi-antenna systems based on the order of the MIMO channel. In particular, the 2×2 and 4×4 MIMO performance results are compared and used to generalize capacity performance in MIMO systems with reference to the single-input single-output (SISO) channel.

The rest of the paper is arranged as follows. Section II gives the MIMO background, Section III is the system model, Section IV is simulation results and Section V is the conclusion.

## II. MIMO BACKGROUND

The quality of a wireless link can be described by transmission rate, transmission range and reliability. Conventionally, the rate can be increased by reducing the range and reliability. By contrast, the range can be extended by reducing the rate and reliability, while the reliability may be improved by reducing the range and the rate. With MIMO-based systems, the three parameters can be simultaneously improved [11]. Initial tests of MIMO systems have shown that an increase in capacity, coverage and reliability are simultaneously achievable with the aid of MIMO techniques [12]. In [13] two signal processing techniques, interference alignment (IA) and interference cancellation (IC), are applied to overcome the capacity limitation dependence on the number of antennas. Using this method, the authors proved that for a MIMO system with a large number of antennas, the throughput on the uplink can be double whereas on the downlink it can be almost double the conventional MIMO capacity. This is made possible by combining the two methods into a single one, interference alignment and cancellation (IAC), showing that combination increases the throughput in scenarios where neither interference alignment (IA) nor interference cancellation (IC) applies separately. In [14] a more powerful technique targeting the network layer, signal alignment (SA), is combined with physical layer network coding (PNC) technique to exploit precoding space at the transmitter and spatial diversity of MIMO system, leveraging higher transmission rates to outperform existing techniques including MIMO or PNC alone, interference alignment (IA) and interference alignment and cancellation (IAC). A transmit beamforming method based on maximum-norm combining (MNC) scheme for MIMO systems in [15] exploits the rotation transformation for the complex vector orthogonalization to achieve higher system capacity and improved quality of service (QoS) as a result of spatial multiplexing and diversity gain at minimal decoder complexity.

In a single-input single-output (SISO) vehicular wireless channel, the signal propagates from the transmitter to the receiver via several propagation paths. Multiple-input multiple-output (MIMO) systems not only enable multipath propagation but also offer substantial spectral efficiency and reliability advantages for the same power and bandwidth resources [12]. A MIMO system with NT transmit antennas and NR receive antennas exploits multipath to achieve diversity gain (NT.NR) in addition to spatial multiplexing gain (min(NT,NR)) to support network reliability and capacity, respectively [16]. As the preceding review indicates, MIMO achieves enhanced network performance in vehicular communications through:

### *A. Diversity Gain*

Diversity, at the transmitter or receiver, is a technique that exploits multipath fading in wireless fading channel. In transmit diversity, same information symbols are sent over multiple independently fading sub-channels, in time, frequency or space. At the transmitter, this is achieved by beamforming if the channel state information (CSI) is available or by space-time coding (STC) if CSI is not available. For NT transmit antennas and NR receive antennas, the maximum diversity gain is NT.NR, asymptotically achievable if the MIMO channel is full rank and the transmitted signal is suitably constructed [17]. In receive diversity, the signals are received by NR receive antennas and signal processing algorithms at the receiver separate received signals and recover transmitted data with higher probability of success. These algorithms include zero forcing (ZF), minimum mean square error (MMSE) and maximum ratio combining [10]. Receive diversity suffers a setback of increased receiver implementation complexity that results in increased physical size and power consumption [18].

### *B. Spatial Multiplexing Gain*

If the individual transmit-receive paths fade independently, the channel matrix is well conditioned with high probability, in which case multiple parallel spatial channels are created. By transmitting independent information streams in parallel through the spatial channels, the data rate is increased [19] in a scheme called spatial multiplexing. Spatial multiplexing exploits the spatial dimension to increase the link capacity for no additional power or bandwidth expenditure. The spatial multiplexing gain, that is asymptotically achievable, is given by min(NT,NR) if the MIMO channel is full rank and a spatial multiplexing scheme is employed. It has been shown that in the high-SNR regime, the capacity of a channel with NT transmit and NR receive antennas, and independent and identically distributed (i.i.d.) Rayleigh-faded gains between each antenna pair, is proportional to min(NT,NR) [20].

### *C. Array Gain*

Array gain can be made available at the transmitter and/or receiver and results in an increase in the average signal-to-noise ratio (SNR) due to coherently combining signals from different antennas, even in the absence of multipath fading. Since this gain requires channel state information (CSI), it can be easily attained at the receivers where CSI is typically available. For NR antennas, this gain makes the average SNR at the output of the combiner NR times greater than the average SNR at any single antenna element and has a significant contribution to the overall performance of the antenna system [21].

## III. CAPACITY IN MIMO SYSTEMS

High data rates can be realized by means of spatial multiplexing, where independent information streams are transmitted in parallel over different transmit antennas in a

MIMO system.

*A. MIMO Capacity Model*

The input-output relationship in a MIMO system can be described by

$$y = Hx + n \tag{1}$$

Where $x = \left[ x_1 x_2 ... x_{N_T} \right]^T$ column vector of transmitted signal, $y = \left[ y_1 y_2 ... y_{N_R} \right]^T$ column vector of received signal, $n = \left[ n_1 n_2 ... n_{N_R} \right]^T$ received noise (column) vector and

$$H = \begin{bmatrix} h_{11} & h_{12} & ... & h_{1N_T} \\ h_{21} & h_{22} & ... & h_{2N_T} \\ ... & ... & ... & ... \\ h_{N_R 1} & h_{N_R 2} & ... & h_{N_R N_T} \end{bmatrix} \tag{2}$$

is the $\left( N_R \times N_T \right)$ MIMO channel matrix with $h_{ij}$ representing the complex gain of the channel between the *j-th* transmit antenna and the *i-th* receive antenna. The channel gains are assumed independent and identically distributed (i.i.d.) and quasi-stationary, i.e., the gains are static over an OFDM symbol-time but may vary over the next slot with zero mean and unity variance, a model adopted for low velocity vehicular environment.

Further, it is assumed that the noise sample $n_i,\ i = 1, 2, ..., N_R$ is a circularly symmetric complex Gaussian (CSCG) random variable with zero mean and variance $\sigma^2$, denoted as $n_i \sim \mathbb{N}_C \left( 0, \sigma^2 \right)$. That is, $\Re\{n_i\} \sim \mathbb{N}\left( 0, \frac{\sigma^2}{2} \right)$, $\Im\{n_i\} \sim \mathbb{N}\left( 0, \frac{\sigma^2}{2} \right)$, and they are independent. In general, the channel gains may be correlated. Assuming $H$ is known at the receiver but not at the transmitter, the energy constraint for the average transmit power, $P$ $\left( i.e., tr\left( E\left\{ xx^H \right\} \right) \le P \right)$, the ergodic capacity of the MIMO channel is given by [22]

$$\begin{aligned} C &= E\left\{ \log\det\left( I_{N_R} + \frac{1}{N_T}\frac{P}{\sigma^2} HH^H \right) \right\} \\ &= E\left\{ \log\det\left( I_{N_T} + \frac{1}{N_T}\frac{P}{\sigma^2} H^H H \right) \right\} \quad bps/Hz \end{aligned} \tag{3}$$

where the expectation is taken with respect to the distribution of the random channel matrix $H$.

If the average signal-to-noise ratio (SNR) at each receiver branch is defined as $\rho = \frac{P}{\sigma^2}$, then the capacity (full rank) can be expressed as

$$C = \sum_{k=1}^{p} E\left\{ \log\left( 1 + \frac{\rho}{N_T}\lambda_k \right) \right\} \tag{4}$$

Where $\rho = \min\left\{ N_T, N_R \right\}$ and $\lambda_1, ..., \lambda_p$ are the eigen values of the matrix $HH^H$ or $H^H H$.

For a SISO system, the channel capacity is given by [23]

$$C = E\left\{ \log\left( 1 + \rho \left| h_{11} \right|^2 \right) \right\} \tag{5}$$

Comparing (4) with (5), it is observed that the capacity of a MIMO system is equivalent to the sum of p parallel SISO channels, each one with an equivalent SNR equal to $\lambda_i$.

*B. SIMO Capacity Model*

When the CSI is not known at the transmitter and $p = N_T$ in (4), then the channel capacity can be expressed as

$$C = N_T \log\left( 1 + \frac{\rho}{N_T}\lambda \right) \tag{6}$$

For the case of a SIMO channel, $N_T = 1$, $r = N_{\min} \triangleq \min\left( N_T, N_R \right) = 1$, $h \in \mathbb{C}^{N_R X 1}$ and $\lambda_1 = \left\| h \right\|_F^2$. Consequently, the channel capacity is given by

$$\begin{aligned} C_{SIMO} &= \log\left( 1 + \rho \left\| h \right\|_F^2 \right) \\ &= \log\left( 1 + \frac{P}{\sigma^2} \right) = \log\left( 1 + \frac{E_x}{N_0} \left\| h \right\|_F^2 \right) \end{aligned} \tag{7}$$

If $\left| h_i \right|^2 = 1,\ i = 1, 2, ..., N_R$, and consequently $\left\| h \right\|_F^2 = N_R$, the capacity is given as

$$C_{SIMO} = \log\left( 1 + \frac{E_x}{N_0} N_R \right) \tag{8}$$

From (8), it can be seen that the channel capacity in a SIMO system increases logarithmically as the number of antennas increases. Only single data stream can be transmitted and the availability of CSI at the transmitter side does not improve the channel capacity at all.

*C. MISO Capacity model*

In the MISO channel model, the channel model is given as $h = \mathbb{C}^{1 \times N_T}$, thus $r = 1$ and $\lambda_1 = \left\| h \right\|_F^2$. When CSI is not available at the transmitter, the channel capacity is given as

$$C_{MISO} = \log\left( 1 + \frac{E_x}{N_T N_0} \left\| h \right\|_F^2 \right) \tag{9}$$

If $\left| h_i \right|^2 = 1,\ i = 1, 2, .., N_T$ and $\left\| h \right\|_F^2 = N_T$, Equation (9) reduces to

$$C_{MISO} = \log\left(1 + \frac{E_x}{N_0}\right) \quad (10)$$

From (10), the capacity is the same as that of SISO channels (Equation (5)). The benefit of multiple antennas cannot be ignored even though the capacity is the same as that of single transmit antenna system. Although the maximum achievable transmission speeds of the two systems are the same, the MISO system offers capability of utilizing the multiple antennas in diversity schemes to improve reliability [24].

## IV. SIMULATION RESULTS

Simulations were performed in Matlab and the results compared with theoretical models described in Section III.

### A. *Ergodic Capacity of MIMO Systems*

Figure 1 presents results of 2×2 and 4×4 MIMO systems with a reference to 1×1 SISO channel. It can be seen that MIMO outperforms SISO and the capacity performance increases linearly with the number of antennas or MIMO order, where 4×4 posts better results than the 2×2 system as verified in Table 1. This is in agreement with the theoretical model where Equation (4) provides higher capacity for increased eigenvalues which, on the other hand, depend on channel matrix described by the MIMO order. Higher MIMO orders yield larger-dimensioned channel matrices with more eigenvalues supporting higher capacities.

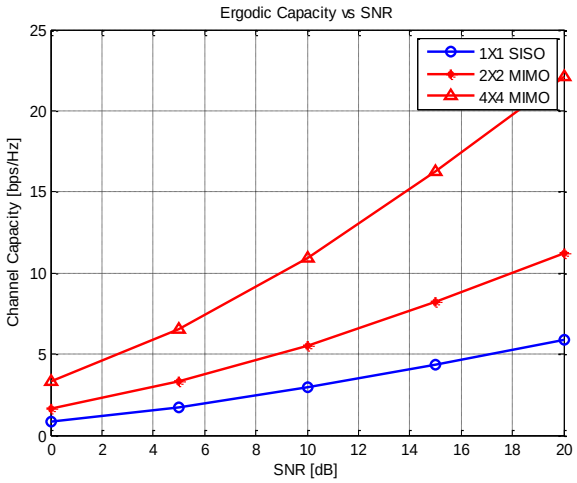



Fig.1: Ergodic Capacity of 2×2 and 4×4 MIMO Systems

The increase with the number of antennas follows a logarithmic function in agreement with (4). These results are consistent with adoption of MIMO techniques in Vehicular networks which require high bandwidth to support time-critical safety applications alongside bandwidth-hungry multimedia streaming associated with infotainment services provided in ITS [25].

Table 1: Performance Comparison

| CONFIG \ SNR dB | 4 | 8 | 12 | 16 | 20 |
|---|---|---|---|---|---|
| 1x1 SISO | 2.0 | 2.5 | 3.5 | 4.5 | 6.5 |
| 2x2 MIMO | 3.0 | 5.5 | 6.5 | 9.0 | 11.5 |
| 4x4 MIMO | 4.0 | 11.0 | 13.0 | 18.0 | 22.0 |

### B. *SIMO and MISO Capacity Systems*

In Fig.2, SIMO and MISO systems capacity performance results are given referenced to SISO system. In the MISO channel with two transmit antennas and one receive antenna the performance is observed to coincide with that of the SISO channel. This is consistent with the theoretical model where Equation (10) presents a single data stream in the way as Equation (5). However, the advantage of MISO over SISO lies in the multi-antenna configuration that is utilized in the implementation of diversity performance. The SIMO system on the hand posts better performance and this is explained from Equation (8) where it is seen to depend logarithmically to number of receive antennas. Above all, when both the number of transmit and receive antennas increase, capacity performance is drastically improved. This can be observed by comparing Figures 1 and 2.

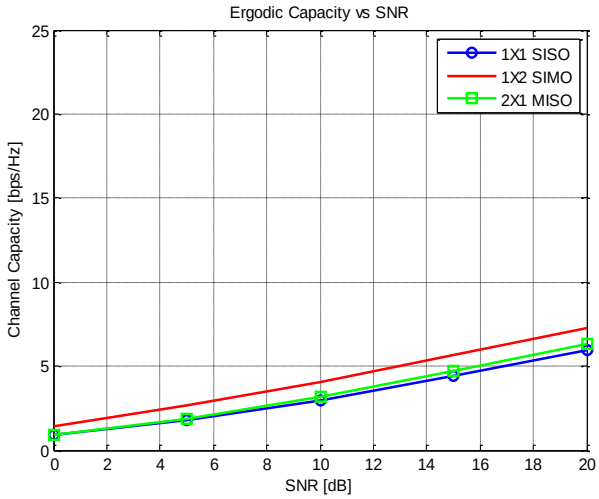



Fig.2: Ergodic Capacity SIMO and MISO Systems

## V. CONCLUSION

MIMO systems carry a promise for capacity performance in current and future wireless networks that support critical bandwidth-demanding services like vehicular networks. It has been demonstrated that MIMO systems outperform systems depending on conventional single antennas configurations to yield high data rate required in certain demanding applications, like vehicular networks. Moreover, the actual capacity realized increases linearly as the number

of transmit and receive antennas is increased, where 4x4 system is double the 2x2 system performance, providing a window for scalabilty. The large area provided by the vehicular surface provides further advantages of managing multi-antenna construction as a guarantee for physical system implementation in vehicular networks.

REFERENCE


[1] US DOT, "Connected Vehicle Research Program: Vehicle-to-Vehicle Safety Application Research Plan", NHTSA, 2011.

[2] John B. Kenney, "Dedicated Short-Range Communications (DSRC) Standards in the United States", Proceedings of the IEEE 2011.

[3] Arrate Alonso, Alexander Paier, Thomas Zemen, Nicolai Czink, Frederik Tufvesson, "Capacity Ealuation of Measured Vehicle-to-Vehicle Radio Channels at 5.2 GHz", Proc. of the IEEE Int. Conf. On Communications (ICC'10), Capetown, May 2010.

[4] Christoph F. Mecklenbrauker, Andreas F. Molisch, Johan Karedal, Fredrik Tufvesson, Alexander Paier, Laura Bernado, Thomas Zemen, Oliver Klemp, and Nicolai Czink, "Vehicular Channel Characterization and Its Implications for Wireless System Design and Performance", Proceedings of the IEEE, Vol. 99, No.7, pp. 1189 – 1212, July 2011.

[5] Chieg-Xiang Wang and Xiang Cheng, "Vehicle-to-Vehicle Channel Modeling and Measurements: Recent Advances and Future Challenges", IEEE Communications Magazine, pp. 96 – 103, Nov 2009.

[6] Khaldoun Albarazi, Utayba Mohammad, and Nizar Al-holou, "Doppler Shift Impact on Vehicular Ad-hoc Networks", Canadian Journal on Multimedia and Wireless Networks, 2011.

[7] Samarendra Nath Sur and Rabindranath Bera, "Doppler Shift Impact On The MIMO OFDM System In Vehicular Channel Condition", I. J. Info. Technology and Computer Science, 2012.

[8] Vahid Tarokh, Hamid Jafarkhani, and A. R. Calderbank, "Space-Time Block Codes from Orthogonal Designs", IEEE Trans. on Info Theory, Vol. 45, No. 5, pp. 1456 – 1467, July 1999.

[9] Luis Miguel Cortes-Pena, "MIMO Space-Time Block Coding (STBC): Simulations and Results", Design Project: Personal and Mobile Communications, Georgia Tech (ECE 6604), April, 2009.

[10] Inaki Berenguer and Xiaodong Wang, "Space-Time Coding and Signal Processing for MIMO Communications", Journal of Science and Technology, Nov. 2003, Vol.18, pp. 689-702.

[11] Ahmad Baheej Al-Khalil, Ali Al-Sherbaz, and Scott Turner, "Enhancing the Physical Layer in V2V Communication Using OFDM-MIMO Techniques", PGNet: htt://www.cms.uk/pgnet2013/proceedings/papers/1569763289.pdf

[12] Helmut Bolcskei, "MIMO-OFDM Wireless Systems: Basics, Perspectives, and Challenges", IEEE Wireless Communications, 2006.

[13] Shyamnath Gollakota, Samuel David Perli, and Dina Katabi, "Interference Alignment and Cancellation", ACM SIGCOMM, 2009.

[14] Ruiting Zhou, Zongpeng Li, Chuan Wu, and Carey Williamson, "Physical Layer Network Coding with Signal Alignment for MIMO Wireless Networks", Proc. of the 8th IEEE Int. Conf. On Mobile Ad-hoc and Sensor Systems (MASS), 2011.

[15] Heuncchul Lee, Seokhwan Park, and Inkyu Lee, "Transmit beamforming Method Based on Maximum-Norm Combining for MIMO Systems", IEEE Trans. On Wireless Communications, Vol. 8, No. 4, pp. 2067 – 2075, Apr 2009.

[16] Swarun Kumar, Diego Cifentes, Shyamnath Gollakota, and Dina Katabi, "Bringing Cross-Layer MIMO to Today's Wireless LANs", ACM SIGCOMM, 2013.

[17] Gordon L. Stuber, John R.Barry, Steve W. Mcclaughlin, Ye (Geoffrey) Li, Marry Ann Ingram, and Thomas G. Pratt, "Broadband MIMO-OFDM Wireless Communications", Invited Paper, Proc. of the IEEE, Vol. 92, No. 2, Feb 2004.

[18] Marco Zoffoli, Jerry D. Gibson and Marco Chiani, "Source Coding Diversity and Multiplexing Strategies for a 2X2 MIMO System", Info. Theory and Applications Workshop, University of California, San Diego, LA Jolla, CA, Jan 27 – Feb 1, 2008.

[19] Lizhong Zheng, "Diversity and Multiplexing: A Fundamental Tradeoff in Multiple-Antenna Channels", IEEE Transactions on Information Theory, Vol. 49, No.5, May 2003.

[20] G. J. Foschini and M. J. Gans, "On Limits of Wireless Communications in a Fading Environment When Using Multiple Antennas", Wireless Personal Communications 6, pp. 311-335, 1998.

[21] Hung Tuan Nguyen, Jorgen Bach Andersen and Cert Frolund Pedersen, "On the performance of Link Adaptation Techniques in MIMO Systems", Wireless Personal Communication, 2006.

[22] Venugopal V. Veeravalli, Yingbin Liang, and Akbar M. Syeed, "Correlated MIMO Wireless Channels: Capacity, Optimal Signaling, and

Asymptotics", IEEE Trans on Info Theory, Vol. 51, No.6, pp. 2058 – 2072, Jun 2005.

[23] Bengt Holter, "On the Capacity of the MIMO Channel – A Tutorial Introduction", Dept. Telecommunications, Norwegian University of Science and Technology, Stavenger, Norway, Tech.Rep., 2001.

[24] Angel Lozano, and Nihar Jindal, "Transmit Diversity vs. Spatial Multiplexing in Modern MIMO Systems, IEEE Trans. On Wireless Communications, Vol.9, No.1, pp.186 – 197, Jan 2010.

[25] Sherali Zeadally, Ray Hunt, Yuh-Shyan Chen, "Vehicular ad hoc Networks (VANETS): status, results, and challenges", Springer Science + Business Media, LLC 2010.

**F Nyongesa** received his BSc in 1983 from University of Nairobi, Kenya, and MSc in 1990 from University of Bradford, UK. Currently he is studying towards his Doctorate at Tshwane University of Technology, South Africa. His research Interests include Digital Signal Processing, Microelectronics and Wireless Communications.

# Stock Position Tracking and Theft Detection System

Solomon Petrus Le Roux, *Member, IEEE*,16084454@sun.ac.za and
Riaan Wolhuter, *Senior Member, IEEE*,wolhuter@sun.ac.za
Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa

***Abstract*—It is estimated that South Africa's red meat industry loses more than R300 million per year due to livestock theft. A Stock Position Tracking and Theft Detection System was designed, developed and tested to serve as a possible solution to the problem. This paper describes the problem identification, system requirements, hardware design, communication protocol design and testing of a prototype system, which consists of a base station and multiple tags. The tags periodically determine their GPS locations and communicate this data to the base station over a $433Mhz$ RF link. The base station communicates the collected data to a server, for processing and analysis. Tests showed that the communication system had a total coverage area of 309.78 hectares, with $993m$ line of sight communication. In a live field test, the base station was placed in a typical environment, the tags were fitted onto sheep and their movement successfully monitored. We conclude that the system can be implemented as a viable solution to detect stock theft in South Africa.**

***Index Terms*—Livestock, Animal tracking, Theft detection, Anti-poaching.**

## I. Introduction

South Africa is a country which has extensive agricultural activities and possibilities. However farmers face severe difficulties due to livestock theft. The Stock Position Tracking and Theft Detection System (SPTTDS) focusses on providing a solution to minimize this problem. This paper considers the formulation, design, implementation and testing of a practical solution to the problem. The proposed solution consists of small, animal borne tracking devices (known as tags) together with a central data collection point (known as a base station). The tags communicate their location information to the base station. The base station in turn, communicates the collected data to a server enabling the stock owner to detect any abnormal positioning or behaviour in real time.

The paper is organized as follows. The problems associated with livestock theft in South Africa are identified in Section II. This is followed by criteria for possible solutions in Section III, followed by a proposed engineering design in Section IV. The hardware design of the different subsystems are discussed in Section V. The communications protocol defined for such a system is particularly important and is covered in Section VI. Practical tests and evaluation are described in Section VII. Section VIII concludes the paper.

## II. Background

A large number of people in South Africa are dependent upon some form of agriculture to make a living. These range from large commercial farmers to individuals who produce just enough food to sustain their families. South African citizens are also dependent upon farmers to produce food for the country. Although most farmers do their best in maintaining big herds of livestock, the increasing criminal activity drives farmers to find other alternative ways of making a living. Crime is no longer restricted to just occasional theft of sheep, but large numbers of stock are stolen via organized criminal networks. This section of the paper considers statistics associated with livestock theft in South Africa.

The Parliamentary Monitoring Group [1] addressed the problem and states that among various challenges that are faced by South African farmers, stock theft is one of the biggest challenges. While it is a priority crime in all provinces except Gauteng, it is much more serious in regions that are bordering other countries (cross-border stock theft) e.g. some areas of the Eastern Cape, Free State, KwaZulu-Natal and Limpopo. Stock theft is not new and some even consider it to be as old as farming itself. However, cross-border stock theft is said to have intensified in the 1990's and became more widespread, organized and violent. While farmers initially had to deal with petty theft of three to four sheep, they must now deal with syndicates who steal entire truck loads of livestock at a time. Stock theft is equally affecting both the commercial and emerging sectors and crime statistics have shown that stock theft has increased in the country.

The Parliamentary Monitoring Group further stated that, "stock theft has an influence on price increases and threatens the sustainability of the livestock industry as some farmers are leaving the industry due to stock theft. This is a serious concern as South Africa is already importing meat and other livestock products to meet the local demand" [1]. According to Mr Aggrey Mahanjana, Managing Director of the National Emergent Red Meat Producers Organisation (NERPO), "the meat industry loses approximately more than R300 million per year due to stock theft" [1]. The stolen livestock is regularly smuggled across borders to Lesotho, Swaziland and Mozambique. According to former Deputy-Minister of Agriculture, Forestry and Fisheries (AFF), Dr Pieter Mulder, "the country lost about 34 000 cattle (worth about R255 million), 60 000 sheep (R71 million) and 28 000 goats (R40 million) to stock theft in the 2008/09 financial year, amounting to a total of R366 million" [1]. Based upon this information alone, it is apparent that livestock theft is a major problem in South Africa.

## III. Solution Criteria

Farmers have implemented various techniques to limit stock theft, with little success. Some deploy night guards which keep watch over the livestock. Others use cattle dogs (Anatolian shepherd dogs) which live with the livestock and ward off intruders. Some communities make use of controlled access points on farm roads and others utilise a community watch system, where members of the community patrol the areas at night. Although these all appear to be good solutions, stock theft continues to increase and in some areas farmers are left with no option

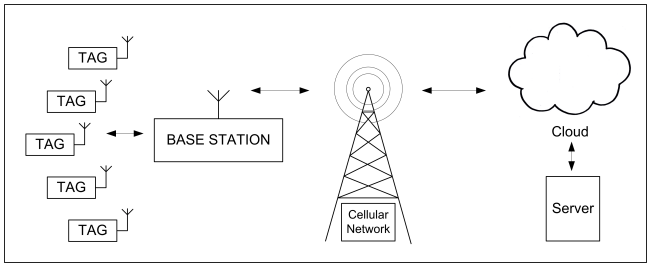



Fig. 1. The Overall Communication System

but to reduce or completely eliminate their herds.

Recently some farmers have turned towards technology for a possible solution. Many see the monitoring of the position of livestock as a possible control measure. This will give the farmer the advantage of knowing the locality of his stock at all times. If the stock moves outside a predefined boundary, an alarm of some sort can be triggered.

During personal communication with farmers, it seems that a general solution could have the structure illustrated in Figure 1. The livestock are fitted with electronic devices known as tags. The tags must be able to communicate their GPS (Global Positioning System) locations to a central data collection point known as a base station. The base station must be able to transmit the collected data over a wireless medium for example a cellular network, satellite constellation or high powered RF (Radio frequency) link to a server, where the data can be processed and analyzed.

The size of the tags is important since each tag must be small enough to fit within a protective casing. The casing will be fitted to the livestock using a collar. The tags must be light and small enough not to harm the animal. This needs to be taken into consideration in choice of system components, to realise the smallest possible footprint. The size of the base station is not as important, as it will be a stand-alone unit located in the veld. The tags need to be battery powered and should operate for a long time, before the battery is depleted.

## IV. Proposed Solution

The design approach for the base station and tags will differ, to meet their very different respective constraints and purposes. The tags communicating with the base station and the latter with the server, form the SPTTDS, as described in the following sections.

### A. Base Station Description and Block Diagram

The base station block diagram is presented in Figure 2. It consists of the following components:

- A power source and power regulator that supply and regulate power to all the components,
- A Microcontroller Unit (MCU) to control all the peripheral devices and process data,
- A RF communication module and RF antenna to receive the tag's transmitted data and to transmit an acknowledge after reception
- A GSM module and antenna to communicate the collected data to a server

The purpose of the base station is to collect the GPS coordinates from the various tags and to communicate the collected data to a server. The GPS coordinates are

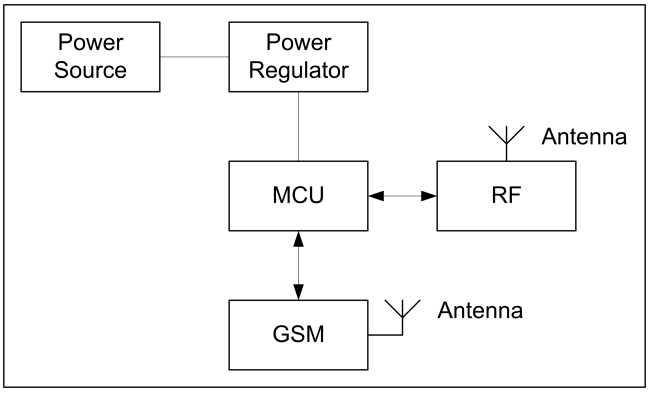



Fig. 2. Base Station - Block Diagram

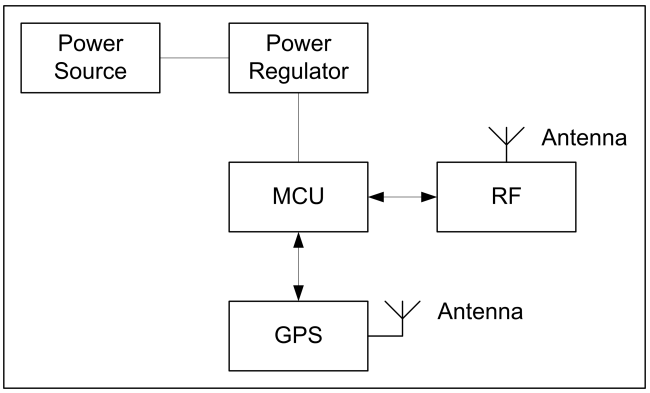



Fig. 3. Tag - Block Diagram

received using the RF transceiver. The collected data is transmitted to the server over a wired or wireless medium. For practical purposes a wireless medium, communicating over a cellular network using a GSM (Global System for Mobile Communications) module is assumed. It is assumed that the base station can be positioned to have good cellular reception. However provision must be made to route the collected data over another communication medium (e.g. data cabling for direct communication to a computer or over wireless communication media), in the case where cellular reception is problematic. The communication system makes use of the RF communication medium between the base station and tags to eliminate the uncertainty of cellular network coverage of the individual tags. The base station and tags must communicate with each other while they remain in the reception radius of the RF transceivers ($\pm 993m$ Line of Sight). The base station receives its power from a rechargeable Lithium Ion battery. It is assumed that a sufficient power source (renewable energy source or Eskom) is connected to the base station to charge the battery and allow for regular transmissions to the server.

### B. Tag Description and Block Diagram

The tag block diagram is depicted in Figure 3. It consists of the following components:

- A power source and power regulator that supply and regulate power to all the components
- A MCU to control all the peripheral devices and process data
- A GPS module and antenna that receive and communicate GPS data to the MCU
- A RF communication module and RF antenna that transmit the GPS data to the base station

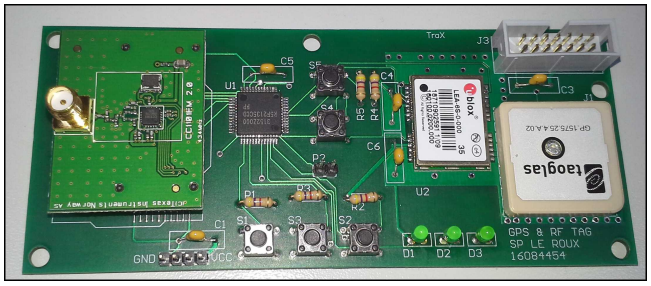



Fig. 4. Assembled Tag

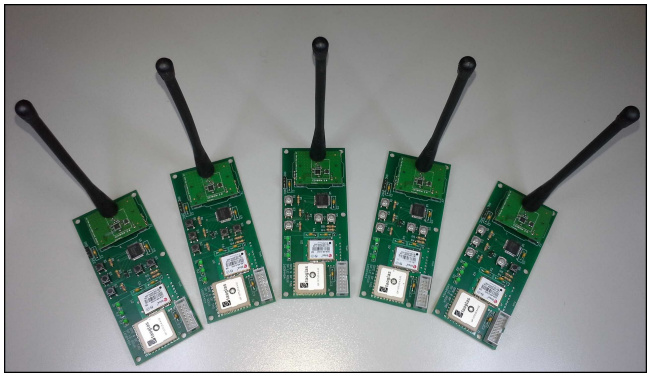


Fig. 5. Five Assembled Tags

The purpose of the tag is to determine, process and transmit its GPS coordinate using a RF transceiver. The transmitted data is intended to be received by the RF transceiver on the base station. The tag receives its power from a rechargeable Lithium Ion battery. In order to achieve the maximum operational time per charging cycle, an effective transmission protocol should control the tag's operation. The physical dimensions of the tag must be small enough to be mounted within a casing, that can be fitted on a livestock unit using a collar.

## V. Hardware Design

After PCB design and manufacture, the base station and the tags were assembled, using normal lab equipment. An assembled tag (with dimensions of 13 x 5.5 x $1cm$) is shown in Figure 4. Figure 5 depicts all the assembled tags. Figure 6 shows the assembled base station (with dimensions of 21 x 9.5 x $1.3cm$). Six power regulators were assembled. The power regulator circuit (with dimensions of 6.5 x 5 x $2cm$) can be seen in Figure 7.

Three operational frequencies from the ISM band, i.e. 433MHz, 915MHz and 2.4GHz, were considered for the

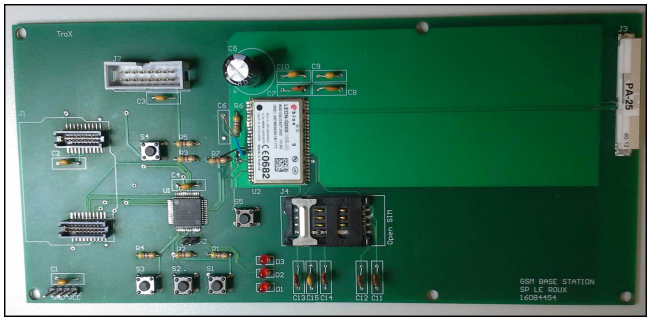



Fig. 6. Assembled Base Station

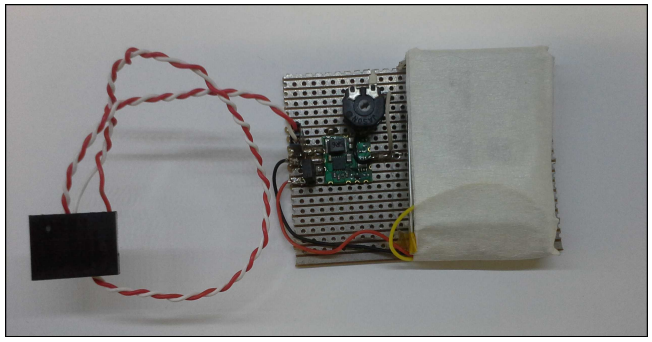


Fig. 7. Power Regulator

SPTTDS. Free Space Loss [2] in RF propagation is, of course, dependent upon transmitter-receiver distance ($d$) and frequency ($f$). However, a tradeoff had to be achieved considering availability of components, cost and size limitations vs. the coverage required. $433MHz$ was chosen as the operational frequency, as antenna size is still manageable and components readily available. It also ensures the maximum communication distance among the three options.

## VI. Communication Protocol Design

The communication protocol defines the tags-to-base station- and base station-to-server communication. Before designing the protocol, possible appropriate wireless communication technologies for the SPTTDS were investigated. This included options such as static routing, dynamic routing, centralized routing, distributed routing, ad-hoc- and star networking topologies.

### A. Investigation

A few of the most important routing topologies can be briefly summarised as follows:

1) Dynamic routing: "In dynamic routing each node continuously learns the state of the network by communicating with its neighbors. Based on the information collected, each node can compute the best path to desired destinations. One disadvantage in dynamic routing is the added complexity in the node" [3].
2) Distributed routing: "In distributed routing nodes cooperate by means of message exchanges and preform their own routing computations" [3].
3) Ad-hoc routing: Ad-hoc network topologies make use of both dynamic and distributed routing. In an ad-hoc network, data can be routed to the base station from one node to the other. Ad-hoc networks could provide good network coverage, as the network coverage is not restricted to the coverage provided by the base station. The coverage expands as new nodes join the network, assuming that these nodes remain in the coverage area provided by other nodes. Forwarding and routing addresses between nodes, are stored in routing tables. The investigation showed that an ad-hoc network would provide a good communication solution, if a power efficient ad-hoc routing protocol could be implemented. This has been the subject of extensive research and many solutions have been published.
4) Deterministic protocols: Typical examples are Round Robin and Token Ring strategies. These are simple, robust, but inefficient at low traffic loads.

5) Contention type Carrier Sense protocols (CSMA): These are very common and a good choice for low traffic loads, providing low latencies.
6) Both the Deterministic and CSMA groups in their pure form present examples of a static routing approach. There is only one pathway between client and server. "In centralized routing a network control center computes all paths and then uploads this information to the nodes in the network" while "in static routing paths are pre-computed based on the network typology, link capabilities, and other information" [3].

Although ad-hoc routing could provide an effective network protocol, it was decided to use a simple CSMA based Star Network typology with static and centralized routing. The decision was mainly based upon the lack of complexity and potential reliability of the strategy. The investigation showed that static routing would be a good choice, based on the following SPTTDS network characteristics:

- Static network paths (communication between base station and tags)
- Relatively small network size (one base station and five tags at this stage, but would be capable of handling many more without unacceptable latency)
- Traffic load relatively low, due to the minimal information from each tag
- During abnormal stock behaviour, reliable alarm forwarding is a prime requirement. Traffic will also quickly peak and presents a typical case for choice of a simple strategy with little overhead regarding routing tables and route optimisation.

### B. Implementation

Further to the investigations set out above, a simple Star Network typology was implemented in the SPTTDS. In this typology the tags check for channel availability and upon gaining access, transmit their data to the base station, which forwards it to a server. The protocol functions as follows: After a system reset the tags:

- Start in power save mode
- Sleep for a programmable duration of time
- Wake-up from power save mode
- Perform hardware configurations
- Determine their GPS location
- Transmit the GPS data to the base station with CCA (Clear Channel Assessment) after a carrier sense and backoff routine for when the channel is occupied
- Wait for an acknowledgement from the base station (with time out for no acknowledge)
- Return to power save mode

The protocol implemented on the base station functions as follows: After a system reset the base station:

- Performs hardware configurations
- Receives a data packet from a tag
- Acknowledges the packet
- Sends the received data to a server

## VII. Measurements and Results

This section presents the overall performance evaluation of the SPTTDS. A number of tests were performed with the assembled base station and tags, as follows:

- A GPS Accuracy Test to determine the accuracy of the coordinates obtained from the tags.
- A Line of Sight Test to determine the maximum communication distance between a single tag and the base station.
- A Clear Channel Assessment Test to illustrate the working of carrier sense with collision detection.
- An Acknowledge Test proved the correct functioning of a communication acknowledge.
- A Movement Simulation Test to simulate moving tags in the Communication System.
- Finally a Live Test where the base station was placed in the veld and tags were fitted onto sheep and monitored.

Most of these tests were performed on a farm called Rooivlei in Carnarvon, Northern Cape. The owner of the farm, Mnr. Fanie Dippenaar, provided the livestock and staff members to assist with the tests. The animals were treated with care and in accordance to the ethical regulations provided by Stellenbosch University. The necessary ethics clearance was obtained to perform tests on live animals. No animals were harmed during the tests.

### A. GPS Accuracy Test

A GPS Accuracy Test was performed to determine the accuracy of the coordinates obtained from the tags. A tag was placed on an arbitrary location and five sets of coordinates were transmitted to the base station. A reference GPS receiver was placed on the same location and five coordinates were collected. The data was compared and the tag proved to be accurate with an GPS accuracy of $< 3m$.

### B. Line of Sight Test

A Line of Sight Test was performed to determine the maximum communication distance between a single tag and the base station. The base station was placed on a high geographic location with open ground around the vantage point. A tag was fitted onto a $2m$ high walking stick. The tag was kept at about $2.5m$ off the ground and the GPS coordinates were transmitted as the tag moved further away from the base station. The obtained coordinates were displayed using Google Earth and can be seen in Figure 8. The figure further shows that the maximum communication distance was $993m$, which corresponds to the theoretical distance of $\pm 1000m$ line of sight. The distance was confirmed with a Garmin eTrex GPS. The bottom graph in the figure shows the elevation of the terrain along the red line. Note that the furthest coordinate measured ($993m$), was transmitted over a slight ground elevation at $760m$. From the obtained distance, the coverage by a single base station was estimated to be $309.78$ hectares [2]. Although the system can communicate over fairly long distances, it is recommended to operate in an $100m$ margin from the absolute maximum communication distance.

### C. Clear Channel Assessment Test

A Clear Channel Assessment Test was performed to demonstrate the functioning of the carrier sense with collision detection. If a tag wakes up in transmit mode and senses a clear channel, it immediately transmits the data to the base station. However if the channel is not clear the tag attempts to transmit the data five times. An exponential back-off waiting time is introduced after every transmission attempt. If the data was not transmitted after five attempts the tag enters sleep mode. To test the effectiveness of this measurement, tags were placed at arbitrary fixed locations.

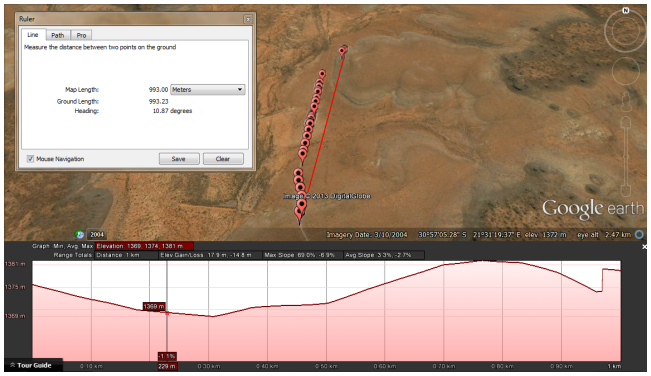



Fig. 8. Line of Sight Results

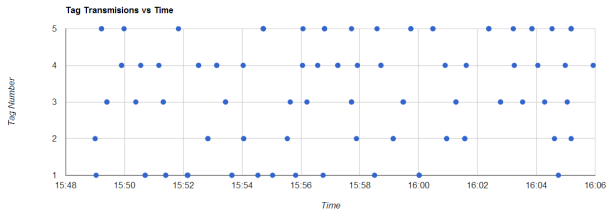



Fig. 9. Clear Channel Assessment Results

Taking a typical transmit time with acknowledge from the base station (10 seconds) into consideration, the tags were configured to transmit every 6 seconds. This scenario will cause high network traffic and frequent collisions. Figure 9 shows the number of data transmissions for every tag over time. If one considers a specific tag (for example Tag 2), the dynamic transmission waiting time if a collision has been encountered, can be seen.

### *D. Acknowledge Test*

An Acknowledge Test was performed to demonstrate that a tag receives an acknowledgement from the base station after data reception by the base station. If a tag wakes-up in transmit mode, it immediately transmits the data to the base station (assuming a clear channel). The tag then enters receive mode where it waits for an acknowledgement from the base station. If an acknowledgement has not been received within a specific time period, the same data is re-transmitted. The tag attempts five re-transmissions. If no acknowledgement has been received within these five attempts, the transmission failed and the tag enters sleep mode. If the tag received an acknowledgement within the five attempts, the transmission is completed and the tag enters sleep mode. After testing the system it was shown [2] that the Acknowledge Test yielded good results.

### *E. Movement Simulation Test*

A Movement Simulation Test was performed to demonstrate the functioning of the SPTTDS. The tags were manually moved around while they communicated with the base station. This data is illustrated in Figure 10, where a different colored flag represents a different tag. Note how the tags moved about the geographic location.

### *F. Live Test*

A Live Test was performed to evaluate the system performance in a practical environment. The five tags were fitted to five sheep, using collars. The tags were configured to transmit approximately every 2.5 minutes. The tags were

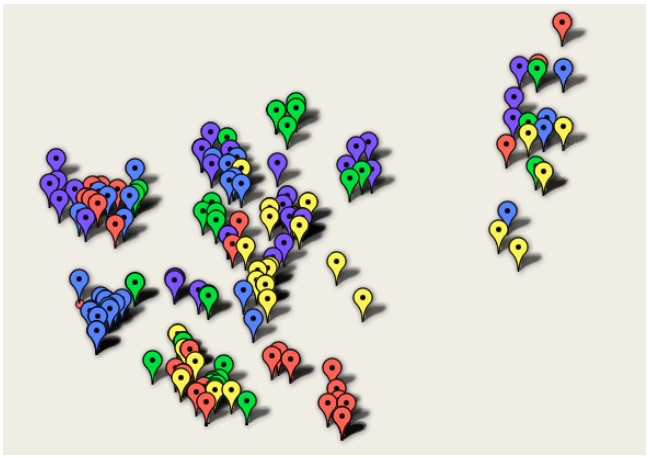


Fig. 10. Movement Simulation Results

Fig. 11. Tags on Sheep

fitted around the necks of the sheep, as can be seen in Figures 11 and 12. The five collared sheep were placed back into the herd and monitored for two hours. The SPTTDS performed as planned and some valuable lessons were learned from practice. An important observation was that a practical application would require the tags to be very durable, as some of the antennas broke off during the test.

Despite the loss of antennas, the test proved to be successful and sufficient data were collected to successfully

Fig. 12. Fitting the Tag

Fig. 13. Live Test Coordinates

demonstrate the functioning of the SPTTDS. Figure 13 shows the coordinates obtained from the live test. As in earlier testing, a different colored flag represents a different tag. The movement of the sheep in the veld during the two hour period, is clearly visible.

## VIII. Summary and Conclusion

The SPTTDS consisted of a base station and five tags. The tags periodically acquired their GPS locations and communicated this location data to the base station over a RF link. The RF link operated in the ISM band with an operational frequency of 433Mhz. The base station transmitted the collected data to a server via GSM.

The functionality of the SPTTDS was validated by various tests, including a Live Test in which stock position was monitored in real time and in a realistic farming environment. GPS locations were plotted on a map to demonstrate and visualize the correct functioning of the system. The approximate line of sight tag-to-base station communication distance was found to be $993m$, resulting in a total coverage of $309.78$ hectares. In future work, we will consider extending the coverage of the system by introducing multiple base stations interconnected by ad-hoc routing techniques.

In conclusion, the Stock Position Tracking and Theft Detection System functioned correctly and we believe it can contribute towards a possible solutions for livestock theft detection in South Africa.

## References


[1] T. P. M. Group. (2010, May) An overview of stock theft in south africa. [Online]. Available: http://www.pmg.org.za/report/20100526-overview-stock-theft-south-africa

[2] S. P. le Roux, "Stock position trancking and theft prevention system," 2013.

[3] W. Leon-Garcia, *Communication Networks*, 2nd ed. McGraw-Hill, 2004.



**Solomon Petrus Le Roux** received his Bachelor of Engineering (BEng), Electrical and Electronics Engineering in 2013 form Stellenbosch University. He is currently studying towards his Master's Degree in Telecommunications at Stellenbosch University. His research interests are Advanced Wireless Tracking and Anti-Poaching Systems for Endangered Wildlife Species.

# A Link Quality Aware Rumor Based Protocol for Wireless Sensor Networks

N'Guettia W. Kouassi[(1)], Karim Djouani[(1,2)], and Anish Kurien[(1)]
Department of Electrical Engineering
[(1)]French South African Technology Institute (FSATI), Tshwane University of Technology
Private Bag X680 Pretoria 0001, South Africa
Tel: +27 12 3824809, Fax: +27 12 3825294
[(2)] LISSI Lab. /Université Paris-Est Créteil, France (UPEC)
Email: kouawilly@live.com, djouani@univ.paris12.fr, kurienam@tut.ac.za

***Abstract*---- Wireless Sensor Networks (WSNs) are employed nowadays in many environments and fields. Today more and more prototypes for routing in WSNs are being seen aiming to either reduce energy or to enhance the QoS. The Relative Coordinate Rumor Routing (RCRR) protocol was recently proposed to overcome some of the drawbacks of the Rumor Routing (RR) which is a hybrid protocol for random walk route seeking. Compared to the RR, the RCRR algorithm is based on a straight line routing scheme. This protocol integrates a topological localisation method useful for determining the nodes' positions. It propagates the information on a hop count approach without any knowledge of the link availability. The RCRR performs very well in terms of energy saving but at the expense of throughput which is lower than that of RR. In this paper, a modified version called Link State C-Rumor (LSCRUMOR) that takes into account the availability of the link is presented. The Expected Transmission Count (ETX) metric is used to decide which direction to relay the packet according to the best link. The simulation of our method conducted under the network simulator NS2 shows an improvement of the throughput, the Packet Delivery Ratio and the average End-2-End delay, compared to the native RR and RCRR.**

**Keywords—Rumor Routing, RCRR, Energy saving, ETX, Routing Algorithms, Link Quality, Throughput.**

## I. INTRODUCTION

One of the most explored fields of telecommunications of this era, which has been seeing extensive ongoing research all around the world is the domain of Wireless Sensor Networks (WSNs). The latter days have been very significant for the development of this area. WSNs consist of sensor devices scattered in an area to monitor various types of surrounding conditions. They are used in several applications including military, medical, and structural monitoring. The sensors work together to forward information to a data analytic centre [1]. In such environments, the extension of the lifetime of the network, the energy consumption, the storage capacity, or the routing protocol employed, are among others the constraints that constitute limiting factors. These motivated a lot of ongoing research around the world.

Routing protocols play a vital role in WSNs as they determine the energy cost incurred as the data is being relayed through. To the extent of reducing this energy consumption, several contributions were proposed such as the SPIN cited in [2] [3]. More protocols are discussed in that paper by the authors.

The early routing algorithms for wireless networks were topology-based. The forwarding decisions were based on the availability of the links between nodes. One of the well known routing protocols is the Rumor Routing (RR) protocols [4] which combines both proactive and reactive routing methods to perform the transmission of data. This hybrid protocol based on a random walk method has served as reference model for the development of other protocol such as the Relative Coordinates Rumor Routing (RCRR) [5].

The RCRR was developed to overcome the drawbacks of the RR, such as unnecessary generation of routes, location information, and wastage of energy. The proposed algorithm successfully achieved the extension of the lifetime of the network by 43%. However, the proposed method lacked some crucial link quality awareness without which the throughput observed was poor. In fact, the RCRR is based on the minimum-hop-count metric, and if it is assumed that a link could be either working or not, this method should integrate a link metric for the awareness before processing the data. The minimum-hop-count tends to choose routes that have less capacity instead of the best available links in the network.

This study investigates the Expected Transmission Count (ETX) [6] as an alternative to the hop count metric used by both the RR and the RCRR. Currently, the ETX is one of the most favoured metric because of its good accuracy in determining the link with the best quality and its ability to improve the network throughput for long routes. The metric was implemented on the RCRR and with various simulations run under NS2 to evaluate the performance of the model. The results showed a significant improvement of the performance of the modified RCRR.

The remainder of this paper is organised as follows: Section

2 presents the state of the art; Section 3 presents the proposed metric and the model; Section 4 deals with the different simulations conducted under different scenarios with a comparative analysis of the original RCRR with the LSCRUMOR. Finally, Section 5 concludes the paper and provides recommendations and future works.

## II. BACKGROUND AND RELATED WORK

### A. Routing in WSNs

This section presents a short state-of-the-art of the most commonly used routing protocols in WSNs. This cannot be done without presenting a classification of these protocols as routing protocols can be classified in different ways. In general the protocols can be divided into three categories depending on the network structure as described in [3] [7], i.e. flat-based routing, hierarchical-based routing, and location-based. In flat networks, each sensor node collaborates with the others to perform the sensing tasks. As it is not feasible to assign one global identifier to each node in such case, the design of a data centric routing technique was needed. Early works on data centric routing are presented in [8] under the names Sensor Protocol for Information via Negotiation (SPIN) and Directed Diffusion (DD) [9].

Some variants of the DD were proposed as well. These are the Gradient-Based Routing (GBR) [10] which has been improved for network coding and a competitive algorithm of it was presented in [11]. The RR is another variant as well which main concept is to route the queries to the notified nodes rather than flooding the entire network to retrieve information. The protocol offers better tradeoffs between setup overhead and delivery reliability. The *agents* that RR delegates to travel the network perform a random walk search until they find the path leading to the event. A modified version has been proposed in [12] in which it is demonstrated that a highly efficient data centric model of routing is more likely to improve the longevity of the network. The authors made use of the existing features of the RR to present their model that handles the nodes failures more efficiently. Another variant of the RR, which exhibits better energy consumption, is the RCRR [5]. In [13] [14], the authors evaluate the performance of this protocol under the network simulator NS2, and presented a modified model that overcame the scalability problem encountered with the original RCRR. It was seen that the RCRR outperformed the RR in the different scenarios applied.

The concept of intersecting the query path and the event path was presented in [15] as a routing scheme that builds both the query path and the event path without any help of the geographic information. In [16], the author proposes a virtual coordinates based routing (VCR) method that adopts a new criterion based on virtual coordinates converted from the absolute coordinates. This technique shows a good trade-off between energy consumption and end-to-end delivery latency

### B. Link Quality Metrics in WSNs

In [6] the authors present the expected transmission count, metric as a new route metric for finding high-throughput paths in multi-hop wireless networks. The ETX is define as the expected total number of packet transmissions (including retransmissions) required to successfully deliver a packet along that path. Practically the paths with the minimum ETX have the highest throughput. The ETX metric incorporates the effects of link loss ratios, asymmetry in the loss ratios between the two directions of each link, and interference among the successive links of a path. The Expected Transmission Time (ETT) [17], which was proposed as an improved version over the ETX, takes into account the differences in link transmission rates. The ETT of a link is the expected MAC layer duration to successfully transmit a packet over the link. The weight of a path is the summation of the ETT's of the links on the path. ETT perform much better than the ETX in environments with varying data rates.

A variant of the ETT called the Weighted Cumulative ETT (WCETT) is proposed in [17] and is used seamlessly in multi-channel environments. It reduces the intra-flow interference by reducing the number of nodes transmitting on the same channel. But some drawbacks were found and presented in [18]. WCETT does not explicitly consider the effects of inter-flow interference; this causes the protocol to route towards a dense area.

In [19] a metric called AirTime, that utilised by the IEEE 802.11s wireless mesh standard, is elaborated to contrast the WCETT failure highlighted above. A variant of the ETX called the Interference and Bandwidth adjusted Expected Transmission count (IBETX) metric [20] has been proposed as well. The proposed solution is to bypass the regions which are congested in the network by calculating the expected link delivery. The technique calculates the interference of a link and overcomes the drawbacks of the native ETX. Though there is a plethora of link metrics that have been developed, only the ones related to this study are discussed.

### C. Location methods in WSNs

The location information can be very useful when utilised in WSNs. By determining the nodes' positions, events within the network can be more accurately localised, and the choice of the route can be directed more usefully. Many localisation algorithms that were proposed for sensor networks provide per-node location information. According to the mechanisms used to estimate the location, the localisation protocols can be divided into two categories: *range-based* and *range-free*. The range-based protocols are defined as protocols that use absolute point-to-point distance estimates (range) or angle estimates to determine the location. The range-free protocols on the other hand do not make any assumptions about the availability of such information. In WSNs, the measurement techniques are used to calculate the distance between nodes by measuring the signal strength.

### D. *Overview of the RCRR scheme*

The RCRR has inherited the characters of the RR but does not forward the information in a random walk fashion but rather make uses of available information on node position to determine the next hop. The nodes choose the next hop according to the horizontal and vertical directions formed by four beacons placed at the cardinal ends of the network area. During a phase carried out when the nodes are deployed and called "*initialisation*" the nodes cooperates with the beacons to get their coordinates. The 4 beacons which know their own coordinates send broadcast packets to update each node in the whole network with a coordinates set as (x, y).
The localisation employed here is rather a virtual localisation based on the nearer hop coordination.

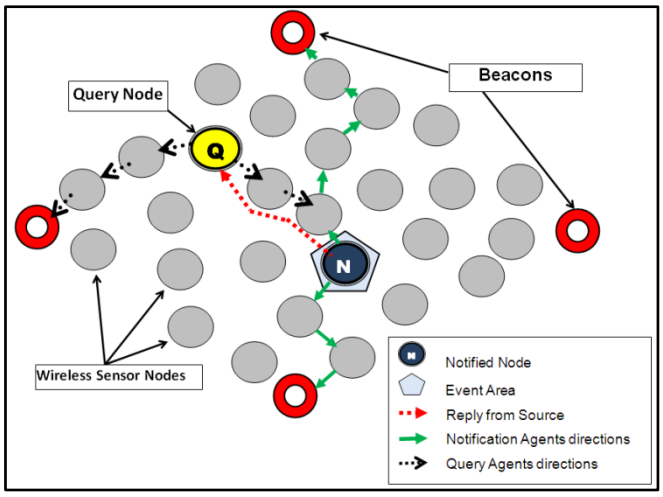



Figure 1: RCRR General Routing Scheme [14]

The beacons that are placed strategically along the vertical and the horizontal axis have knowledge of their own coordinates, while the other sensors do not. the sensor node which detects the event in the network assigns two notification agents to travel, one according to the increasing coordinates and the other one according to the decreasing coordinates (towards vertical beacons). The same process is carried out with the node that makes a request but this time the agents are assigned to travel towards the horizontal beacons. Every agent is assigned to travel in a different direction. The intersection of the line formed by those agents is then use to reply to the query node.

## III. PROPOSED APPROACH

### A. *Presentation of the ETX Metric*

ETX was developed to determine the highest throughput route. The ETX metric for a link as described by the author is the expected number of data transmissions required to send a packet over the link, including retransmissions. The ETX of an entire route is then computed by taking the sum of the ETX of each link. The protocol employed here will select the routes with the minimum ETX value. This value is calculated using the forward and reverse delivery ratios of that link. The forward delivery ratio ($d_f$) is the estimated probability of successful delivery of a packet at the recipient; the reverse delivery ratio ($d_r$) is the probability that the Acknowledgement packet (*ACK*) is successfully received at the transmitting node, given that the data packet was received successfully. As the design of the ETX metric does not depend on a particular routing protocol, its feasibility is investigated with the RCRR. ETX is able to find the path that has the maximum throughput in spite of the loss in link quality that may occur. Every node keeps track of the probes it receives during the last *w* seconds; this is then used to calculate the delivery ratio from the sender at any time *t* as follows:

$$d(t) = \frac{count(t-w, t)}{w/\tau} \quad (1)$$

where the value *Count* (*t*−*w*, *t*) is the number of probes received during the time window *w*, and $w/\tau$ reflects the number of probes that are supposed to be received.

The ETX of a single link is computed as follows:

$$ETX\ (link) = \frac{1}{d_f * d_r} \quad (2)$$

The ETX of a route "*R*" is the sum of all the ETX of a link "*l*" and is computed as follow:

$$ETX(R) = \sum_{l \in R} \frac{1}{d_f^l * d_r^l} \quad (3)$$

### B. *The Link State Coordinates Rumor (LSCRUMOR)*

As the RCRR is oriented by its coordinate's scheme, the proposed version named LSCRUMOR will also make use of the position information and calculate the ETX of every link surrounding the query node and the notified node, prior to establishing the route. Since the RCRR delegates two agents to travel: one towards increasing coordinates and another towards decreasing coordinates, the ETX will be calculated as so. The probe window is set to 10s within which the packets are sent at a rate of 1 packet per second. The following picture is an overview of the algorithm.

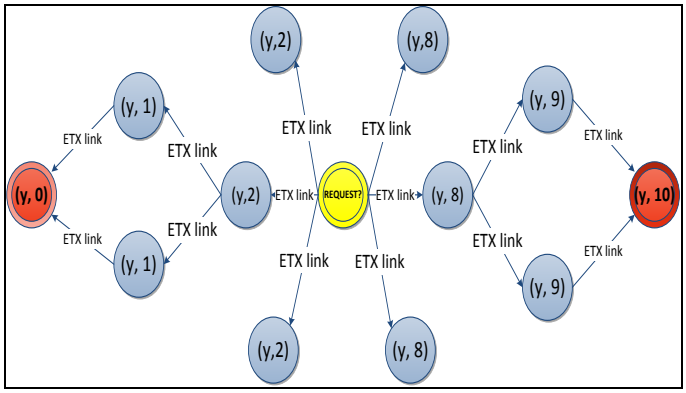



Figure 2: LSCRUMOR checks the different links at the Query node

Figure 2 above depicts the checking process where LSCRUMOR computes the *ETX (link)* of every node starting by the Query node. The coordinates *(y, 0)* and *(y, 10)* are the horizontal positions of the beacons. The remaining nodes in the network obtain their set of coordinates from the beacon and process them to their neighbours for an automatic update. It is represented in the picture by the virtual values *(y, 1)*, *(y, 2)*, etc. The value of "*y*" is the vertical coordinate of the node and the numbers "*0, 1, 2 ...*" are the horizontal ones. The yellow node noted as *request* is the querying node. It assigns two agents to check the energy remaining at the node and the availability of the link between the neighbour node and itself; one of the agents is sent toward the left node and the other one toward the right node. When there is no energy left at the node, the algorithm exits and continue with another. To avoid sending the two agents in the same direction, LSCRUMOR take advantage of the knowledge of the position of every sensor. If a node is at *(y, 2)* for instance, the link between his concurrent nodes will not be checked to avoid a return to the source. After this process the choice of the best route is determine by summing up the *ETX (link)* and look for the best *ETX (Route)*.

### C. About the implementation of ETX on NS2

In ns2 the implementation of ETX is protocol dependant. Meaning that the way it is integrated might differ from one routing protocol to another to meet the requirement. In our work the main modification where done as follow:

- *lscrumor.h*: the header file containing the probe management classes and the ETX functions
- *lscrumor.cc:* this file contain the algorithm procedures and the different functions used to compute the forward/reverse delivery ratio, and the ETX values. The forward delivery ratio is calculated as:

$$forward\ delivery = Probe\_count/Probe\_window$$

The reverse delivery ratio is:

$$reverse\ delivery = \frac{ProbeNeighbours\_[neighbour]}{Probe\_window}$$

Where the *ProbeNeighbours_[neighbour]* is the total number of neighbours present in the vicinity of the probe packet. The ETX value is then computed as indicated in section II.A.

- *lscrumor_pkt.cc*: where the probe packet is define.

The structure of the packet is as follow:

```
struct hdr_lscrumor_probe{
   Packet Type
   Source IP Address
   Broadcast ID
   # Neighbours from which probes have been received
   Addresses of neighbours
   Number of the probes that are received
   Time at which the probes packets are sent
   Size = sizeof(hdr_lscrumor_probe)
   Return size
}
```

The following picture is the selected path after all the computations are done.

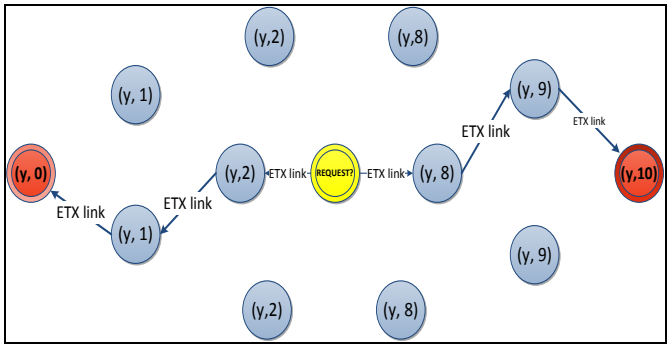



Figure 3: Data Processing Through Selected Path

Figure 3 above is the path followed by the data after the choice of the best link. Figure 4 below repeats the same process employed at the query node. However, this time it is vertically as the sensing node notifies about the detected event. Once the vertical and the horizontal line cross their way, the original RCRR algorithm is employed to create a route for the reply.

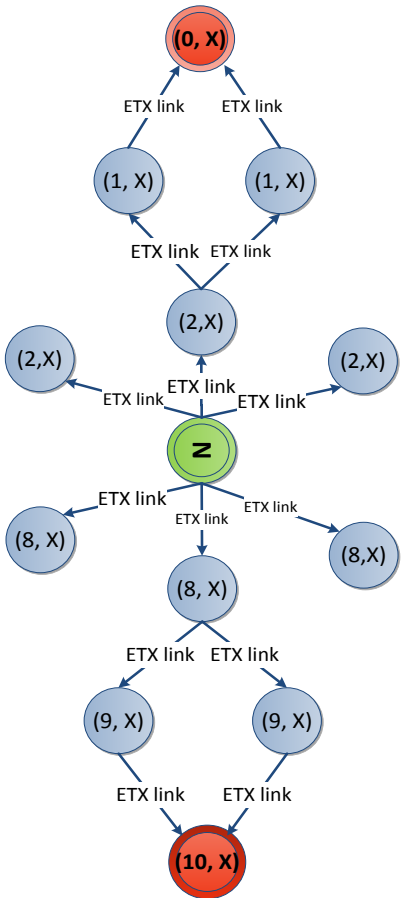



Figure 4: LSCRUMOR checks the links at the notified node

## IV. SIMULATION AND RESULTS

### A. *Parameter settings*

TABLE1: NS-2 SIMULATION SCENARIO

| Description | Parameters |
|---|---|
| Number of nodes | 30 -100 |
| Time to Live | 15 hops |
| Traffic type | CBR |
| Transport type | UDP |
| Simulation time | 785 seconds |
| Event details | Synchronous/Asynchronous |
| Simulation area | 2500*1900 m |
| Data flow start | 6.5 seconds |
| Number of Sources | 3 (with 3 sinks) |
| Packet size | 100 bytes |
| TX power | 600 mW |
| RX power | 100 mW |
| Mac value | S-Mac |
| # simulation/ area | 10 |

The S-MAC protocol presented in [3] is used as medium access control protocol. S-MAC employs four new techniques to reduce energy consumption and to support self-configuration:

- It enables a low-duty-cycle operation of nodes in a multi-hop network to cause nodes periodic listening and sleeping. It also forms many virtual clusters based on common sleep schedules;
- S-MAC also adopts a similar contention schemes mode to implement the basic MAC layer requirements;
- It can avoid overhearing unnecessary traffic so as to save energy by exchanging listening for sleeping;
- It supports efficient transmissions of long messages without massive delays.

For energy saving perspective, the S-Mac is chosen.

### B. *QoS Evaluation (Throughput and End-2-End delay)*

The figure below shows the comparison between the RR and the RCRR in terms of throughput. The test was run over different network sizes which ranged from 30 to 100 nodes.

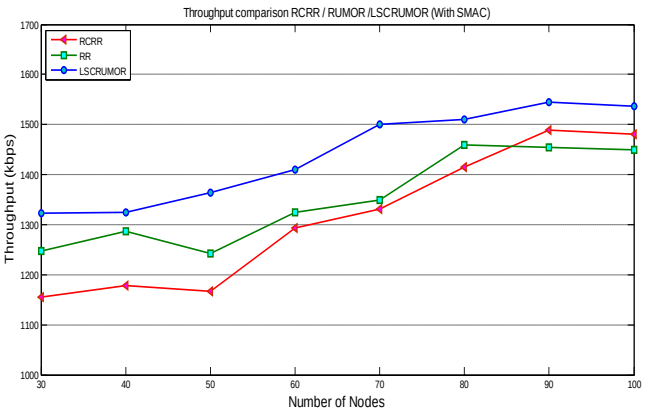



Figure 5: Throughput comparison RR /RCRR/LSCRUMOR

Figure 5 is a comparison of the three considered protocols. The improvement of the LSCRUMOR (blue line) compared to the native RCRR (red line) and the RR (green line) is depicted throughout this graph. The RCRR protocol tends to behave better in dense environments with a higher number of nodes. As the network increases, our protocol performs better due to the choice of non lossy links introduced by the ETX and the minimisation of the number of transmission and retransmissions. The next graph is a comparison of the average end-to-end delay.

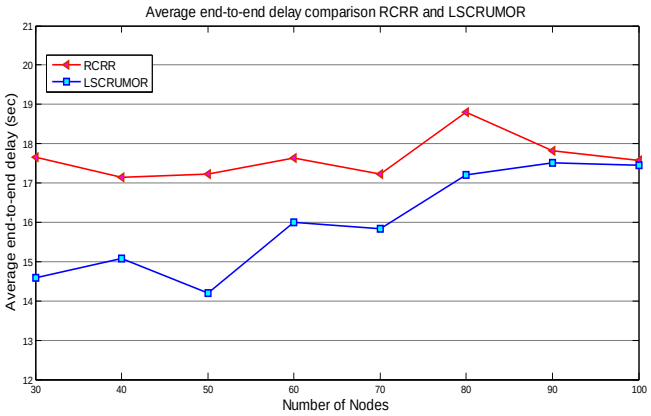



Figure 6: Average End-to-End delay comparison RCRR versus LSCRUMOR

The Figure 6 above shows the improvement regarding the average end-to-end delay when we simulate the LSCRUMOR protocol. With a smaller network area, LSCRUMOR dispatches data packets more efficiently compared to the RCRR. The high delay observed with the RCRR is due to the waiting time for retransmission as it does not implement the knowledge of the availability of the link before processing, and the probable collisions that may occur.

## V. CONCLUSIONS

The Relative Coordinate Rumor Routing is an energy efficient routing protocol. The implementation of the link awareness is an added value to exhibit higher throughput. The Link State Coordinate Rumor is a new version that improves the throughput of the RCRR by more than 10% and allows a better average end-to-end delay. The overhead of ETX does not affect the decision of route selection under low-data rate. However with the ETX, as the network grows there is more contention and more data flows. The probing scheme incurs some overhead and could alter the energy cost. We would suggest that future works investigate the integration of a link metric based on the signal strength.

.

## VI. ACKNOWLEDGEMENTS

The authors would like to thank F'SATI at Tshwane University of Technology and the TELKOM Centre of Excellence at F'SATI, TUT for making this research possible. This work is based on the research supported in part by the National Research Foundation of South Africa (Grant reference number (UID) 80050).

## VII. REFERENCES


[1] R. Kay and F. Mattern, "The Design Space of Wireless Sensor Networks", in *IEEE Wireless Communications*, page(s): (6): 54–61, 2004.

[2] Al-Karaki, Jamal N., and Ahmed E. Kamal. "Routing techniques in wireless sensor networks: a survey." *Wireless Communications, IEEE* 11.6 (2004): 6-28.

[3] Akkaya, Kemal, and Mohamed Younis. "A survey on routing protocols for wireless sensor networks." *Ad hoc networks* 3.3 (2005): 325-349.

[4] D. Braginsky, David, and Deborah Estrin. "Rumor routing algorthim for sensor networks." *Proceedings of the 1st ACM international workshop on Wireless sensor networks and applications*. ACM, 2002.

[5] G. Huanan "Relative coordinates rumor routing in wireless sensor network", MTech Dissertation, Tshwane University of Technology, South Africa 2010

[6] De Couto, Douglas SJ, et al. "A high-throughput path metric for multi-hop wireless routing." *Wireless Networks* 11.4 (2005): 419-434.

[7] Aquino-Santos, Raúl, et al. "Performance analysis of routing strategies for wireless networks." *Revista Facultad de Ingeniería Universidad de Antioquia* 52 (2010): 185-195.

[8] Ye, Fan, et al. "A scalable solution to minimum cost forwarding in large sensor networks." *Computer Communications and Networks, 2001. Proceedings. Tenth International Conference on*. IEEE, 2001.

[9] C. Intanagonwiwat, R. Govindan, D. Estrin and J. Heidemann, "Directed Diffusion for Wireless Sensor Networking", IEEE/ACM *Transaction on Networking*, 11: 529–551, 2003.

[10] C. Schurgers and M.B. Srivastava, "Energy efficient routing in wireless sensor networks", in the *MILCOM Proceedings on Communications for Network-Centric Operations*: Creating the Information Force, McLean, VA, 2001.

[11] Miao, Lusheng, et al. "Network coding and competitive approach for gradient based routing in wireless sensor networks." *Ad Hoc Networks* 10.6 (2012): 990-1008.

[12] Patra, Mr Chiranjib, Parama Bhaumik, and Debina Chakroborty. "Modified rumor routing for wireless sensor networks." *IJCSI* (2010).

[13] Kouassi, N. W., K. Djouani, and A. Kurien. "Performance Study of an Improved Routing Algorithm in Wireless Sensor Networks." *Procedia Computer Science* 19 (2013): 1094-1100.

[14] N.W. Kouassi, K. Djouani, A. Kurien, " An energy efficient rumor based routing protocol for Wireless Sensor Networks", proceeding of the *16th Southern Africa Telecommunication Network and Application conferences*, 1-4 Sept. 2013, Stellenbosch, ISBN 978-0-620-57882-0.

[15] C. Chou, Cheng-Fu, Jia-Jang Su, and Chao-Yu Chen. "Straight line routing for wireless networks." *Computers and Communications, 2005. ISCC 2005. Proceedings. 10th IEEE Symposium on*. IEEE, 2005.

[16] Chen, Min, et al. "Virtual coordinates based routing in wireless sensor networks." *Sensor Letters* 4.3 (2006): 325.

[17] R. Draves, J. Padhye, and B. Zill, "Routing in multi-radio, multihop wireless mesh network", Mobicom, pp.114-128, 2004.

[18] Yang, Yaling, Jun Wang, and Robin Kravets. "Designing routing metrics for mesh networks." *IEEE Workshop on Wireless Mesh Networks (WiMesh)*. 2005.

[19] Chen, Min, et al. "Virtual coordinates based routing in wireless sensor networks." *Sensor Letters* 4.3 (2006): 325.

[20] N. Javaid, A. Bibi, K. Djouani " Interference and bandwidth adjusted ETX in Multi-hop networks", USA, GLOBECOM Workshops, 2010 IEEE, pp1638-1643



**N'guettia William KOUASSI** received his undergraduate degree in 2009 in Computer System Engineering in Ivory Coast. He obtained his MTech degree in Electrical Engineering at TUT and jointly with an MSc degree through F'SATI at TUT. His research interests include Wireless Sensor Networks, routing protocols and energy efficiency.

# Collaborative Incentive Schemes and Virtual Coordinate Routing in Sensor Networks

D.J. Brand and A.E. Krzesinski
Department of Mathematical Sciences
Stellenbosch University, 7600 Stellenbosch, South Africa
Email: aek1@cs.sun.ac.za Tel.: +27 21 808 4232 Fax: +27 21 882 9865

***Abstract*—This paper describes two incentive schemes to promote collaboration in wireless sensor networks (WSNs). We investigate the effect of the incentive schemes in a WSN where the nodes have limited bandwidth and energy resources and do not have IP addresses. We assume that the sensor nodes do not have an on-board capability of establishing their geographic location, nor do they have the computational capacity to compute their location by means of triangulation.**

**We present a virtual coordinate routing (VCR) algorithm which is used to compute near optimal routes. The VCR algorithm occasionally computes a route prefix which cannot reach the intended destination. We present a simple method to recover from most of such situations. We show that routes computed by the VCR algorithm yield performance results approximately equal to that provided by least hop count routing and geographic based routing. We show that the incentive scheme reduces the network delay and affords consistent performance over a wide range of node speeds when the nodes are mobile.**

***Keywords*—*Virtual coordinate routing, geographic based routing, least hop count routing, sensor networks, incentive schemes to improve wireless network performance.***

## I. Introduction

Consider a wireless sensor network (WSN) where the mobile nodes consist of relatively simple low cost sensor nodes. Each sensor is equipped with a wireless transceiver with limited range, a processor of limited capacity and a modest amount of storage. We assume that the sensor nodes are mobile, have limited bandwidth and energy resources and do not have IP addresses. The sensor nodes do not have an on-board global positioning system (GPS) capability to establishing their geographic location, nor do they have the computational capacity to compute their location by means of triangulation.

Standard methods for computing routes in wireless networks cannot be used in a sensor network since the sensor nodes do not have IP addresses nor can they afford the bandwidth and energy overhead incurred by the frequent transmission/reception of network state update messages. One method of calculating routes would be to use geographic based routing (GBR) where a node forwards a packet to a neighbour that is closer to the intended destination of the packet. However, we assume that the sensors do not know nor can they estimate their geographic coordinates so GBR cannot be used.

We present a virtual coordinate routing (VCR) algorithm which is used to assign a unique identifier to each sensor node and compute near optimal routes. The VCR algorithm relies on the designation of certain nodes as *landmark* nodes. Our VCR algorithm chooses landmark nodes at random. The VCR algorithm occasionally computes a route prefix which cannot reach the intended destination. We present a simple method to recover from most of such situations and to find a route to the intended destination.

An ad hoc network requires that the nodes act as relays to form multi-hop routes connecting the origin-destination node pairs that are out of radio transmission range with respect to each other. It may be necessary to give the nodes incentives to spend their resources in forwarding packets that originate at other nodes. This can be done by introducing a credit balance for each node, where the nodes use credits to pay for the costs of sending their own traffic, and earn credits by forwarding traffic from other nodes. We describe two variants of a credit-based incentive scheme to promote collaboration in wireless sensor networks.

We present a simulation study of a 50-node network model and we show that the routes computed by our VCR algorithm yield a performance that is approximately as good as that provided by geographic based routing (GBR) and by least hop count (LHC) routing when the nodes are fixed and when the nodes are mobile.

The remainder of the paper is organised as follows. Section III presents the credit-based incentive schemes. Section II summarises the properties of GBR and VCR. Section IV presents our VCR algorithm. Section V describes a discrete event simulator that was used to investigate the properties of our VCR algorithm. Experimental results are presented in Section VI. Conclusions are presented in Section VII.

## II. Routing in wireless sensor networks

Consider a WSN where data packets are sent from the sensor nodes to a designated node that collects the data. This can in principle be done by flood routing where the data packets are forwarded to every neighbour. Under this scheme a copy of the data packet will arrive at its intended destination along the shortest possible path, along with many other copies of the packet which will be discarded. However, the bandwidth and energy cost of transmitting multiple copies of each packet precludes the use of flood routing in a WSN.

This work is based on research supported in part by the National Research Foundation of South Africa (Grant specific unique reference number (UID) 83965), Nokia-Siemens Networks and Telkom SA Limited.

Standard wireless network routing protocols such as Dynamic Destination-Sequenced Distance-Vector routing [4], Temporally-Ordered Routing Algorithm [9], Ad hoc On-Demand Distance Vector routing [10] and Destination-Sequenced Distance-Vector routing [11] cannot be used to find routes in a WSN since sensor nodes do not have IP addresses.

Geographic based routing [5, 6, 13] requires that the nodes are aware of their geographic location. This can be done by means of on-board GPS receivers. Not all of the nodes need be GPS enabled. Some GPS enabled nodes can be designated as *anchor* nodes and the non GPS enabled nodes can estimate their location by means of triangulation among themselves and the anchor nodes. However, packet forwarding by means of geographic based routing (GBR) is not possible if the sensor nodes have no means of knowing or estimating their geographic location.

Unlike GBR where the nodes are located in a three-dimensional coordinate system, virtual coordinate routing [2, 7, 12, 14] designates $m$ *landmark* nodes and establishes an $m$-dimensional *hopId* coordinate system. Nodes are located within this coordinate system and VCR is used to find routes connecting nodes in this coordinate system.

VCR routes are not optimal in terms of minimising the route hop count. In addition, VCR has some unique problems. Some VCR algorithms require careful selection of the landmark nodes, and the landmark nodes may have to be reselected if the nodes are mobile. VCR sometimes does not find a route even though a route exists – this is referred to as the *dead end* problem. Some VCR algorithms contain elaborate methods to avoid the dead end problem. Other VCR algorithms have inefficient methods of dealing with dead ends such as rerouting packets via the landmark nodes.

## III. The credit-based incentive schemes

Two incentive schemes are investigated. The *free-for-all* protocol provides a baseline: this scheme does not promote cooperation among the nodes. The *origin pays* protocol when combined with credit redistribution (see below) is designed promote cooperation among the nodes.

### A. Free-for-all

The *free-for-all* protocol takes no measures to regulate the willingness of the nodes to forward packets on behalf of other nodes. This protocol is not fair since nodes that act as relays use more of their resources on behalf of the community than the other nodes, yet they receive insufficient compensation in return.

### B. Origin pays constant pricing

Consider a packet that is offered to the originating node $i$ of route $r$. The originating node $i$ is required to pay $2|r|$ credits. If the originating node has insufficient credit, the packet is dropped, else the credit counter $B_i$ at the originating node is debited $2|r|$ credits and the packet purse $b_p$ is initialised to $2|r|$.

When a packet completes transmission from node $i$ to node $j$, the credit counters $B_i$ and $B_j$ at nodes $i$ and $j$ are each incremented by 1 and the packet purse $b_p$ is decremented by 2. If the queue at node $j$ is full then the packet is dropped, else the packet is queued.

### C. Origin pays congestion pricing

Consider a packet that is offered to the originating node $i$ of route $r$. The originating node $i$ is required to pay $2|r|$ credits. If the originating node has insufficient credit, the packet is dropped, else the credit counter $B_i$ at the originating node is debited $2|r|$ credits and the packet purse $b_p$ is initialised to $2|r|$.

When at time $t$ packet $p$ is successfully transmitted from node $i$ to node $j$ of route $r$, the credit counter $B_i$ at node $i$ is incremented by an amount $c_i(t)$ for transmission, and the credit counter $B_j$ at node $j$ is incremented by an amount $c_j(t)$ for reception and an amount $c_i(t)+c_j(t)$ is deducted from the packet purse $b_p$ where a congestion charge $c_i(t) = 0.5 + n_i(t)/n_i$ is applied at node $i$ at time $t$ to receive or to transmit a packet where $n_i(t)$ is the number of packets queued at node $i$ at time $t$ and $n_i$ is the maximum size of the packet queue at node $i$.

### D. Credit redistribution

Credits are periodically destroyed at nodes that have a surplus of credits and credits are periodically created at nodes that have a deficit of credits. The redistribution process works as follows. Consider a credit redistribution event at node $i$ at time $t$. Let $\delta$ denote the discount rate per unit time. Let $\Delta$ denote the time interval between successive discount events at node $i$. Let $B$ denote the target credit balance and let $B_i(t)$ denote the credit balance at node $i$ at time $t$. The credit balance at node $i$ is adjusted

$$B_i(t + \Delta) = B_i(t) + \delta\Delta(B - B_i(t))$$

so that nodes that possess an amount of credit that exceeds the target credit balance will destroy a fraction $\delta$ of the surplus per unit time, while nodes whose credit balance is less than the target credit balance will create a fraction $\delta$ of the deficit per unit time.

Credit redistribution ensures that under-provisioned nodes are able to send some traffic, while at the same time providing over-provisioned nodes with a mechanism for disbursing a fraction of their credits for the common good rather than accumulating them. No mechanism is needed to transfer the redistributed credit. A node that has a surplus of credits will over a period of time destroy a fraction of that surplus. Likewise a node that has an under-supply of credits will over a period of time create a fraction of that deficit.

## IV. Virtual coordinate routing

In this section we present an implementation of VCR which differs from that of Zhao et al. [14] and Cao et al. [2] in several ways. First, the landmark nodes are chosen at random. Second, landmark management is not needed since the landmarks do not change as the sensors move. Third, a simple method is presented for dealing with (most) dead end problems. Finally, our method trivially avoids circular routes.

### A. Virtual coordinate routing

Consider a network consisting of a set $\mathcal{N}$ mobile nodes. A subset $\mathcal{M} \subset \mathcal{N}$ of the nodes are designated as landmark nodes. Each node $i \in \mathcal{N}$ is assigned a coordinate $H^i = (H^i_m)_{m\in\mathcal{M}}$ referred to as the *hopId* of node $i$, where $H^i_m$ is the least hop

count distance from node $i$ to the landmark node $m$. For example, if there are three landmark nodes $\mathcal{M} = \{m_1, m_2, m_3\}$ then node $i$ is assigned a hopId $H^i = (H^i_{m_1}, H^i_{m_2}, H^i_{m_3})$ which denotes its hop distance from the landmark nodes $m_1$, $m_2$, and $m_3$ respectively.

The distance between two nodes is determined as follows. Let $H$ denote the hop count distance between nodes $i$ and $j$. The following triangulation inequality holds [14]

$$\max_m(|H^i_m - H^j_m|) \leq H \leq \min_m(H^i_m + H^j_m).$$

This defines an upper bound and lower bound for the distance between two nodes. The upper bound is is not suitable as a distance metric in greedy routing, since it can assign a large distance between two nodes that are directly connected. The lower bound is a better metric to use, although it does not correspond to the true hop count distance. It was found [14] that the $p$-norm

$$D = \left( \sum_{m=1}^{M} |H^i_m - H^j_m|^p \right)^{1/p} \quad (1)$$

where $M = |\mathcal{M}|$ affords a good metric for the distance between nodes $i$ and $j$, where typically $p = 10$.

### B. The VCR algorithm

Fig. 1 presents the greedy routing algorithm.

Consider a network with $N$ nodes. Let COUNT[$i$] record the number of nodes that are one hop away from node $i$. Let NEIGHBOURS[$i$] record the set of nodes that are one hop away from node $i$. Let HOPIDDISTANCE[$i$][$j$] record the the hopId distance between nodes $i$ and $j$ computed using Eqn. (1).

Let LABEL[$i$] record whether the route from $S$ to $i$ has been found or not. A route is encoded as an $N \times N$ matrix where PRED[$S$][$j$] = $i$ denotes that node $i$ is the penultimate node on the route from $S$ to $j$. Thus PRED[$S$][$j$] = $i$ denotes that the route from $S$ to $j$ is $S \to \cdots \to i \to j$.

The LABEL and PRED data structures are initialised in lines 1 to 8 of the VCR algorithm to indicate that all routes outbound from $S$ are unknown apart from the route from $S$ to $S$. The state variable $C$ is initialised to indicate that a route has been found from $S$ to $S$.

The main loop of the VCR algorithm between lines 9 and 30 is repeatedly executed until $C$=$D$ which indicates that a route from $S$ to $D$ has been found and a value for PRED[$S$][$D$] has been computed. Thus at line 9 the VCR algorithm has computed the route from $S$ to $C$. Node $C$ has more than one neighbour (see line 15) so node $C$ is not a dead end (see line 11).

Lines 13 to 23 search the neighbourhood of node $C$ for a node $N$ that has more than one neighbour (node $N$ is not a dead end) and is closest in terms of hopId distance to the destination node $D$. If node $N$ is already in the route from $S$ to $D$ then it is not a candidate. In this way circular routes are avoided.

Lines 24 to 30 check that the newly-found node $N$ is closer to the destination node $D$ than node $C$ is. If so, node $N$ is the successor of node $C$ on the route to node $D$. If not, a dead end is declared at line 28 and PRED[$S$][$D$] = 0.

```
// Find a route from node S to node D
1: function GETROUTE(Node S, Node D)
// initialisation
2:     for all i in N do
3:         LABEL[i] := false
4:         for all j in N do
5:             PRED[i][j] := 0
6:         end for
7:     end for
8:     LABEL[S] := true; C := S
// find a route from C to D
9:     while C ≠ D do
10:        if C ≠ S then
11:            assert COUNT[C] > 1
12:        end if
// find a neighbour N of C closer to D
13:        MINDIST := ∞
14:        for all N in NEIGHBOURS[C] do
//  avoid circular routes and dead ends
15:            if LABEL[N] = true or COUNT[N] = 1 then
16:                continue
17:            else
18:                NTOD:= HOPIDDISTANCE[N][D]
19:            end if
20:            if MINDIST > NTOD then
21:                MINDIST := NTOD; B := N
22:            end if
23:        end for
// greedy routing: N is closer to D than C is
24:        if MINDIST < HOPIDDISTANCE[C][D] then
25:            PRED[S][B] := C
26:            LABEL[B] := true; C := B
27:        else
28:            PRED[S][D] := 0; break        // dead end
29:        end if
30:    end while
31: end function
```

Fig. 1. Greedy hop ID routing.

### C. The dead end problem

Zhao et al. [14] observe that VCR routing using Eqn. (1) as a distance metric "effectively avoid(s) the dead ends, even for a sparse network. The intrinsic reason is that the Hop ID coordinates are constructed based on the topology, and the virtual coordinate distance between any two nodes is very close to the shortest path length between them."

Nonetheless dead ends, although infrequent, can occur. The dead end solution method proposed by Zhao et al. [14] involves rerouting via the closest landmark node to the destination. We implemented a simple and effective way of resolving most dead end problems. The test at line 15 for a successor to node $C$ excludes successors that lead immediately to dead ends. This one-step look-ahead procedure does not guarantee that a dead end will never occur.

### D. Computational complexity of the VCR algorithm

The computational complexity of the VCR algorithm presented in Fig. 1 is $O(N^2)$ where $N = |\mathcal{N}|$ is the number of nodes. This implies that VCR, like other interior gateway protocols, does not scale. We will therefore investigate the performance of VCR in relatively small network models.

## V. THE SIMULATION MODEL

Previous investigations into VCR [12, 14] reported that VCR performed well in large networks containing thousands of nodes. However, these simulation studies of large network models either ignored the effects of radio interference or used simple probabilistic packet loss models.

We developed a discrete event simulator that models contention-based access according to the IEEE 802.11 DCF. The simulator uses the ns2 [8] default radio attenuation model with transmission range parameters adapted for a WSN.

Given the complexity of simulating the DCF function, and the fact that VCR does not scale, our experiments address a relatively small network model with $N = 50$ nodes. The model assumes a flat $105\,m \times 105\,m$ terrain that is partitioned into a grid of $7 \times 7$ cells. The radio transmission range is $30\,m$ and the cell size is $15\,m$. One node is initially placed at random in each cell. The remaining node is placed at random. The nodes are placed so that the network is connected. The average number $30^2\pi/15^2$ of neighbours for a node is approximately 12. Our initial simulation experiments did not model the effect of obstacles as was done in [14]. Voids were not expressly modelled although voids can form as the nodes move.

Each node has a bandwidth of $2\,Mbps$. Each node has a packet queue which can store 50 packets. Each node attempts $x$ packet transmissions per second. The transmissions take place at the instants of a Poisson process. Each transmission selects a random originating node and a random destination node. Each packet is 1,554 bytes long. RTS/CTS is disabled.

The nodes are either all immobile (static) or all mobile in which case they move according to the random waypoint mobility model [1]. The destination $(x, y)$ is chosen at random, the node speed is uniformly distributed in the range [2.5, 7.5) *cm/sec* and the pause time is 10 *secs*.

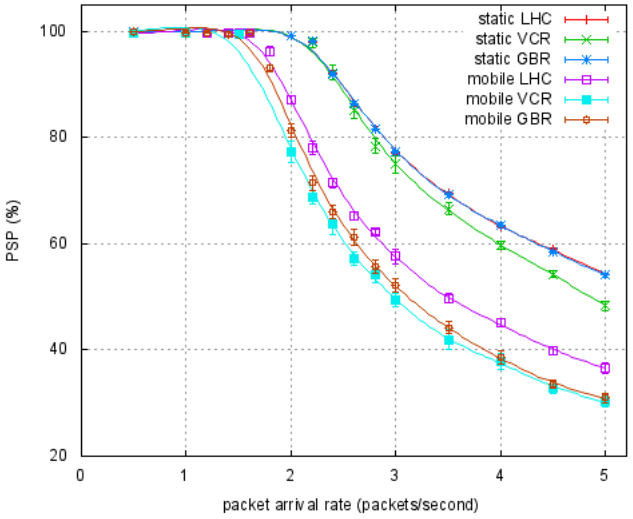
Three routing algorithms were simulated namely least hop count (LHC) routing, virtual coordinate routing (VCR) and geographic based routing (GBR). In the case of VCR some 10% of the nodes are selected at random as landmark nodes.

The parameter values of the incentive schemes are as follows. The discount rate $\delta$ is 5% *per second*, the discount interval $\Delta$ is 10 *msec* and the target credit balance $B$ at each node is 100.

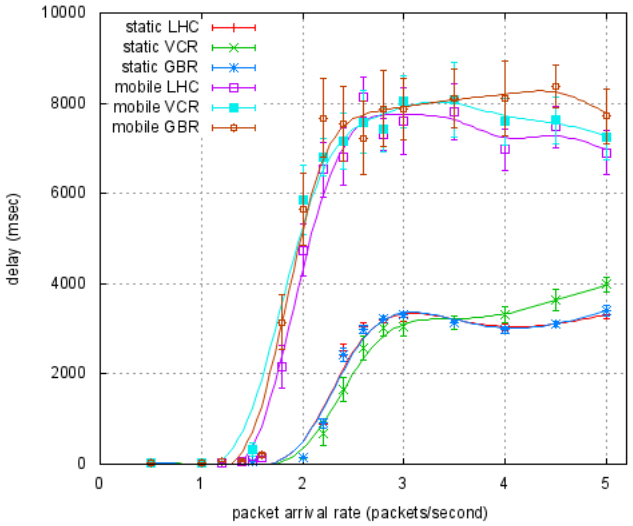
Each simulation experiment models 2,500,000 packet transmissions. The experiments were replicated 12 times to compute 95% confidence intervals.

## VI. WSN PERFORMANCE

Figs. 2, 3 and 4 show the performance of the 50-node network model in terms of (a) the packet success probability (the probability that a packet is successfully delivered to its destination) and (b) the network delay as a function of the packet sending rate when the *free-for-all*, the *origin pays* constant pricing and the *origin pays* congestion pricing incentive schemes are used respectively. Performance data are presented for the case when all the nodes are static (immobile), and for the case when all nodes are mobile. The figures show the following.



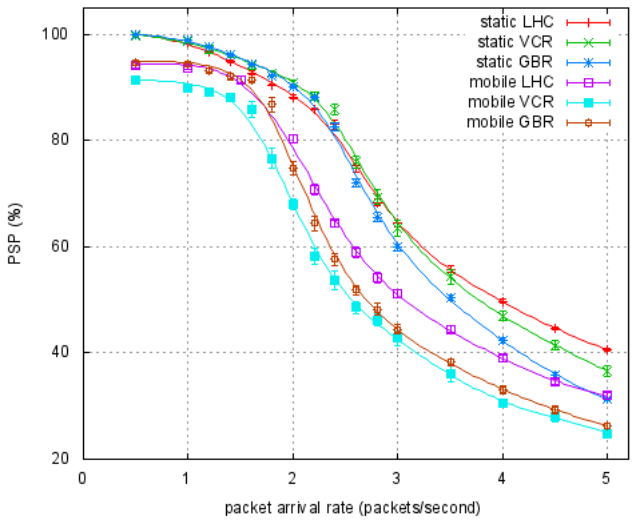

(a) Packet success probability



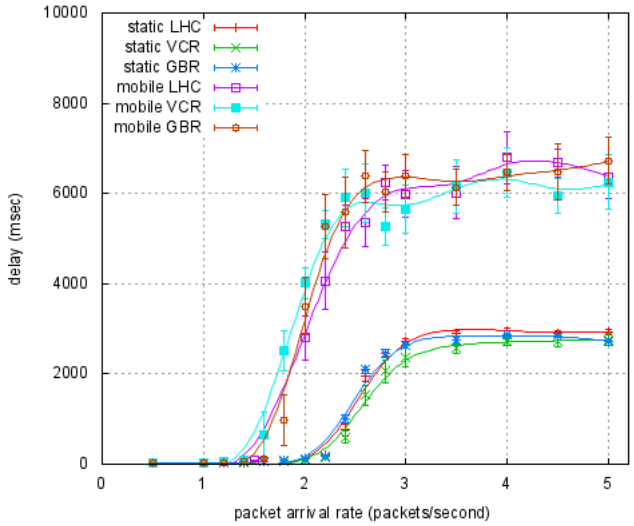

(b) End-to-end delay

Fig. 2. *Free-for-all*: 50-node network model.

- The LHC, VCR and GBR routing algorithms yield approximately the same packet success probability (PSP) and approximately the same end-to-end delay.
- The network performance improves when the nodes are immobile: the PSP is larger and the network delays are lower. This could be due to the location of the immobile nodes in a grid of cells where one node is placed at random in each cell. When the nodes are mobile they may temporarily form clusters with strong radio interference causing frequent packet drops, repeated packet retransmissions, congested packet queues and long delays.
- The PSP decreases rapidly when the average packet sending rate at each node exceeds 2 packets per second. In the case of *free-for-all*, the deterioration in the PSP is due to congested packet queues and the resulting tail drops. In the case of the *origin pays* incentive scheme, the deterioration in the PSP is due to a lack of credit at the originating nodes, in particular at the nodes at the edge

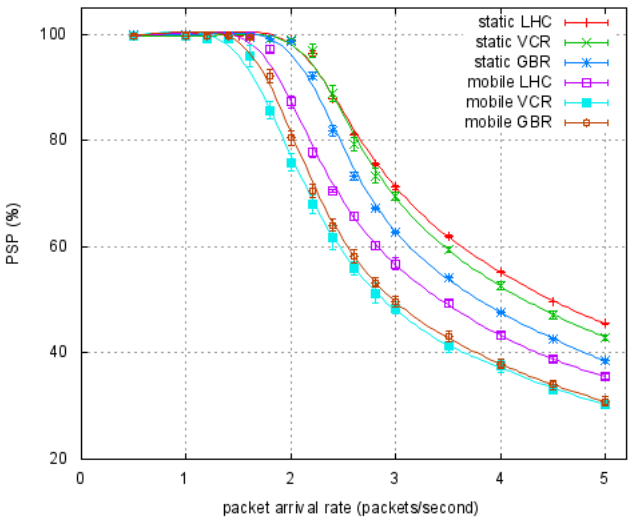



(a) Packet success probability

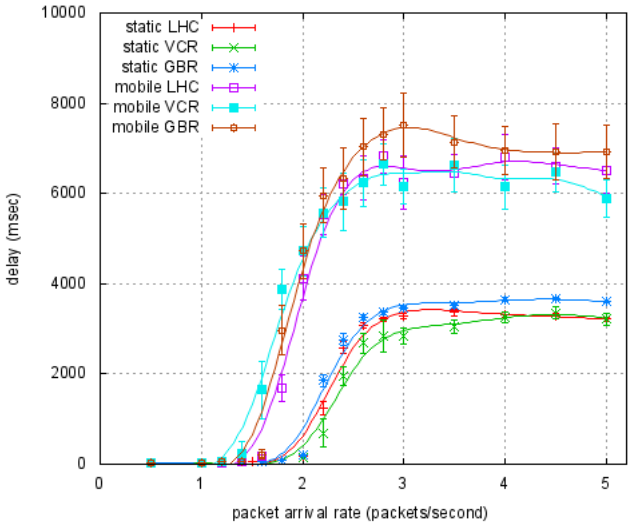



(b) End-to-end delay

Fig. 3. *Origin pays* constant pricing: 50-node network model.




(a) Packet success probability




(b) End-to-end delay

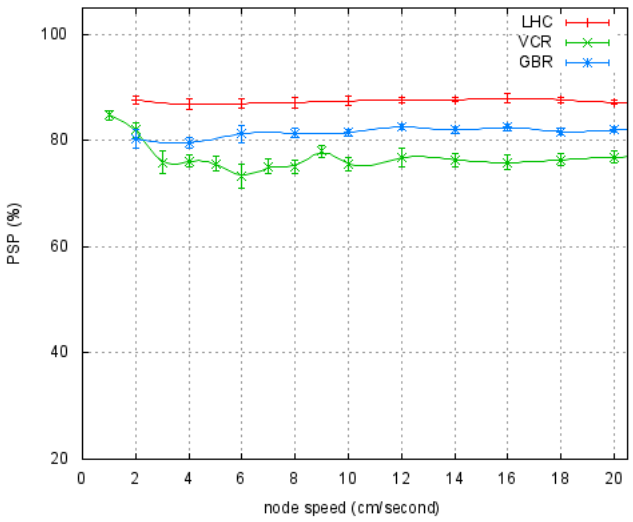
Fig. 4. *Origin pays* congestion pricing: 50-node network model.

of the network, which prevents the packets from being accepted for transmission.

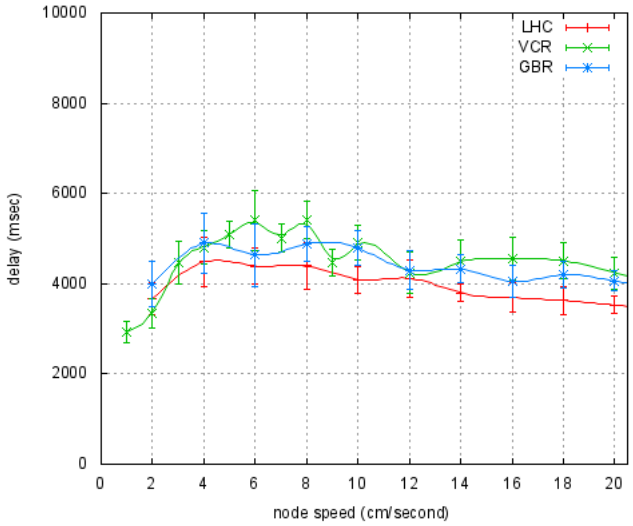
- The network delay increases rapidly when the packet sending rate exceeds 2 packets per second, due to congested packet queues particularly at nodes in the centre of the network which host many transit routes.
- The *origin pays* with constant prices incentive scheme yields (especially at low packet sending rates) a slightly worse PSP and a slightly lower delay that *free-for-all*.
- The *origin pays* with congestion prices incentive scheme provides a PSP equivalent to *free-for-all* and, due to the less congested packet queues, a modest reduction in the average network delay. See Section VIII.

For the network models and mobility models investigated, the credit-based incentive scheme achieves a modest reduction in the network delay. However, what is significant is that LHC routing, VCR and GBR yield approximately the same performance. It is somewhat surprising that GBR performs so well. GBR like VCR uses greedy routing which does not find optimal routes. In the case of VCR the landmark nodes are chosen at random and are not re-selected as the mobile nodes move. In addition the dead end recovery procedure, although used infrequently, can find inefficient (large hop count) routes. Nonetheless, the performance obtained from the VCR compares well against LHC routing, which supports the claim [14] that the virtual coordinate distance between any two nodes as computed from Eqn. (1) is very close to the shortest path length between them.

Finally, Fig. 5 shows the PSP and the delay for the 50-node network when all nodes are mobile. The node speed is varied from 1 to 20 *cm/sec*. The packet sending rate is 2 packets per second. The figure shows that the PSP and the delay are relatively insensitive to the node speed.




(a) Packet success probability




(b) End-to-end delay

Fig. 5. *Origin pays* congestion pricing: 50-node network model, all nodes mobile.

## VII. Conclusion

This paper applies several variants of a credit-based incentive scheme to promote collaboration in WSNs. We present a virtual coordinate routing (VCR) algorithm which is used to compute near optimal routes. The VCR algorithm occasionally computes a route prefix which cannot reach the intended destination. We present a simple method to recover from most of such situations and to find a route to the intended destination.

For the 50-node network model under consideration, we show that the credit-based incentive scheme achieves a modest reduction in the network delay, and that the routes computed by the VCR algorithm yield performance results approximately equal to that provided by least hop count routing and geographic based routing.

## VIII. Future work

It may be appropriate to compare the average performance of the nodes at the periphery of the network versus the average performance of the nodes at the centre of the network as an indicator of the effectiveness of the incentive scheme, rather than the average performance of all the nodes. This remains to be investigated.

## References

[1] T Camp, J Boleng and V Davies. A survey of mobility models for ad hoc network research. *Wireless Communications and Mobile Computing*, 2 pp. 483–502, 2002.

[2] Q Cao and T Abdelzaher. Scalable logical coordinates framework for routing in wireless sensor networks. *ACM Transactions on Sensor Networks (TOSN)*, Volume 2 Issue 4, November 2006 pp. 557–593 ACM New York, NY, USA

[3] Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. IEEE Std 802.11-2012 (Revision of IEEE Std 802.11-2007 ) IEEE, 3 Park Avenue, New York, NY 10016-5997, USA 29 March 2012.

[4] D Johnson, Y Hu and D Maltz. RFC 4728: The Dynamic Source Routing Protocol (DSR) for Mobile Ad Hoc Networks for IPv4 (Feb 2007).

[5] B Karp and HT Kung. Gpsr: Greedy perimeter stateless routing for wireless networks. In *Proceedings 6th Annual International Conference on Mobile Computing and Networking (MobiCom 2000)*, pp. 243–254 (2000).

[6] F Kuhn, R Wattenhofer, Y Zhang and A Zollinger. Geometric ad-hoc routing: Of theory and practice. In *Proceedings 22nd Annual Symposium on Principles of Distributed Computing*, pp. 63–72 (Apr 2003).

[7] T Moscibroda, R O'Dell, M Wattenhofer and R Wattenhofer. Virtual coordinates for ad hoc and sensor networks. In *Proceedings 2004 Joint Workshop on Foundations of Mobile Computing (DIALM-POMC'04)*, pp. 8–16 (Oct 2004).

[8] http://www.isi.edu/nsnam/ns/

[9] V Park and S Corson. INTERNET-DRAFT draft-ietf-manet-tora-spec-04.txt (2001), Temporally-Ordered Routing Algorithm (TORA) Version 1 Functional Specification.

[10] C Perkins, E Belding-Royer and S Das. RFC 3561, Ad hoc On-Demand Distance Vector (AODV) Routing (July 2003).

[11] CE Perkins and P Bhagwat. Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for Mobile Computers, In Proceedings Conference on Communications Architectures, Protocols and Applications (SIGCOMM 94), London England UK (Oct 1994) pp. 234–244.

[12] A Rao, S Ratnasamy, C Papadimitriou, S Shenker and I Stoica. Geographical routing without location information. In *Proceedings 9th Annual International Conference on Mobile Computing and Networking (MobiCom '03)*, San Diego, California, USA, pp. 98–99 (Sept 2003).

[13] I Stojmenovic, P Bose, P Morin and J Urrutia. Routing with guaranteed delivery in ad hoc wireless networks. In *Proceedings 3rd International Workshop on Discrete Algorithms and Methods for Mobile Computing and Communications (DIALM '99)*, pp. 48–55 (1999).

[14] Y Zhao, Y Chen,B Li and Q Zhang. Hop id: A virtual coordinate-based routing for sparse mobile ad hoc networks. *IEEE Transactions on Mobile Computing*, Volume 6, Issue 9 pp. 1075–1089 (Sept 2007).

Dirk Brand is an MSc student in the Department of Mathematical Sciences (Computer Science Division) at Stellenbosch University. Anthony Krzesinski is a Professor of Computer Science at Stellenbosch University, South Africa.

# Rural Wireless Mesh Network Analysis On Mobile Devices

**Ghislaine L. Ngangom Tiemeni**, Isabella M. Venter and William D. Tucker
Department of Computer Science
University of the Western Cape, Private Bag X17, Bellville 7535 South Africa
Tel: +27 21 9593010, Fax: +27 21 9591274
Email: {**3261404**, iventer, btucker}@uwc.ac.za

**Abstract-Traffic-generator software is a valuable tool for generating synthetic yet realistic workloads that can be analyzed to test communication networks' quality of service. This paper describes the research and design of a packet-level traffic generator (known as MTGawn) on a mobile platform. The main objective is to describe a way to adapt a well-known packet generator—designed for a personal computer (PC)—for use with mobile devices. This will simplify the feasibility testing and monitoring of wireless mesh networks deployed in remote areas where mobile devices are more practical and affordable than PCs or laptops, i.e. ease of battery charging and of usability of a touchscreen given the resource constraints of a mobile device. In order to achieve this objective, a suitable model for emulating realistic workloads randomized in terms of packet size and time between packets was designed using various statistical distributions such as Constant, Uniform, Pareto and Normal. The most common transport protocols, transmission control protocol (TCP) and user datagram protocol (UDP), were used to enable the generation of accurate and representative traffic patterns that were characteristic of user behaviour. The paper covers work done in the laboratory with a mesh network testbed. We employ design science research in a cyclical fashion to move toward demonstrating that a mobile generator can provide acceptable packet generation and analysis functionality on a mobile platform in order to move from the laboratory to rural in-field use.**

**Index Terms—network performance, end user mobile application, quality of service, traffic emulation, wireless mesh network**

## I. INTRODUCTION

A wireless mesh network (WMN) is a communication network system in which all nodes communicate together without any centralized infrastructure to control the network. This lack of dependency on any pre-installed infrastructure makes WMNs suitable for addressing connectivity issues in disaster and field scenarios, allowing, for instance, the network to handle geographic challenges in dispersed communities that typically experience large amounts of attenuation or noise due to the proximity of mountains (Gunashekar, Das, Erlebach, & Warrington, 2014; Carrano, Bletsas, & Magalhães, 2007). In addition, WMNs are known to be self-healing and extremely reliable because of their ability to maintain connectivity even if a node fails and the capacity of all nodes to obtain access to a wireless point if one node has access to it (Gunashekar et al., 2014; Hamidian, Palazzi, Chong, Navarro, Korner, & Gerla, 2009).

In comparison to traditional, cable and digital subscriber line broadband networks, which require a substantial investment in order to begin offering services, WMNs are very flexible, hence they can be built progressively and can grow as the number of users grows. Further benefits of WMNs include ease of installation, the absence of cable cost, automatic connection among nodes, the discovery of newly added nodes and redundancy (Carrano et al., 2007; Hamidian et al., 2009; Sichitiu, 2005). However, like many other evolving technologies, WMNs come with advantages and drawbacks. As user data travels through multiple hops, the complexity of the routing protocol impacts on performance and makes it challenging to provide high-level security to end user. Factors that directly affect the performance of WMNs include load balancing, avoiding congested routes and dealing with interference patterns (Carrano et al., 2007).

Nevertheless, the decentralized topology of WMNs, added to their flexibility, low cost and ease of deployment, has made them a useful of providing broadband connectivity to people living in rural areas (Gunashekar et al., 2014; Carrano et al., 2007).

In recent years the deployment of multimedia applications such as video conferencing and voice-over Internet protocol (VoIP)—in addition to traditional data services—has significantly increased in WMNs. As demand for multimedia services evolves, it becomes fairly difficult to maintain quality of service (QoS) in multihop WMNs where dynamic environments cause fragile links and high packet-loss ratios. These factors degrade the QoS of multimedia services and significantly affect user satisfaction (Cheng, Mohapatra, Lee, & Banerjee, 2008). For this reason network managers have to keep track of network evolution in order to identify and resolve possible problems that may occur so as to efficiently deliver network services to the end user. To deal with this issue, many networking experts rely on performance evaluation tools such as traffic generators and packet sniffers.

Literature on some of the most interesting tools used to estimate the performance of networks was reviewed in the initial stages of this study. Despite the powerful features of these tools, the literature survey indicated that they are

unable to simplify the feasibility testing and monitoring of WMNs deployed in rural areas. For this reason we propose a new tool that we call MTGawn, i.e. 'mobile traffic generator for analysis of wireless networks'. A prototype of this tool has been designed and deployed on a mobile device, and is being tested using an experimental mesh network as testbed.

The paper is organized as follows. Section II covers a relevant sample of work related to network performance analysis. Section III describes the methods used to follow a design science research (DSR) approach to move toward producing a traffic emulator with analysis capabilities on a mobile form factor. Section IV describes the architecture of a prototype deployed in a laboratory wireless mesh testbed. Section V presents the results of preliminary tests, and finally, Section VI draws conclusions and identifies future work.

## II. Related Work

The monitoring and evaluation of WMNs is essential to ensure that the QoS—required for widely used protocols, such as VoIP—is satisfied. To do so, networking experts rely on traffic-generator software to generate synthetic but realistic traffic patterns that can assist in predicting and estimating the performance of networks. These are discussed briefly below.

According to Nicola, Giordano, Procissi, and Secchi (2005), it is essential to evaluate the performance of high-speed networks either because of the lack of reliable tools to generate traffic workloads at high rates or because of the inaccessibility of network equipment. For these reasons they implemented a tool called Brute (browny and robust traffic engine), a Linux application allowing high-speed packet generation on personal computers (PC).

Similar to Brute, Harpoon (Sommers & Barford, 2004; Sommers, Kim, & Barford, 2004) is a flow-level traffic generator for router and network tests that focuses on the generation of transmission control protocol (TCP) and user datagram protocol (UDP) packet flows. These packet flows have the same characteristics as routers for the purpose of showing the empirical behaviour of routers under actual conditions. Another significant feature of Harpoon is its ability to self-configure by automatically extracting parameters from standard net flow logs or packet traces.

Avallone, Pescape, and Ventre (2003) developed D-ITG, a tool for the generation of transport-layer traffic (TCP and UDP) and other types of traffic, including VoIP, Telnet, and domain name service. D-ITG has numerous features such as allowing the measurement of round-trip time and one-way delay, while it has the capacity to keep information about received and transmitted packets. This feature allows the evaluation of important network QoS metrics such as throughput, jitter, packet loss and average bit rate. D-ITG has additional functionality such as using different network loads or different network configurations to study scalability problems. It allows the generation of complex and varied traffic sources, and offers the option to repeat exactly the same traffic pattern (Avallone et al., 2003).

## III. Design and Methodology

DSR methodology was used to build a mobile traffic generator. Each cycle of the DSR iterative process consists of six phases—identify, build, document, select, evaluate and communicate (Brocke & Buddendick, 2006) (see Figure 1).

### A. *Identify*

During this phase methods such as a literature survey and document analysis were used to identify and clarify our main concern, which is the lack in the literature of a tool capable of running efficiently on devices with limited screen interface and computing power such as mobile devices. Our aim is to make available a cost-effective tool that generates typical traffic and that will run on affordable and easily accessible devices.

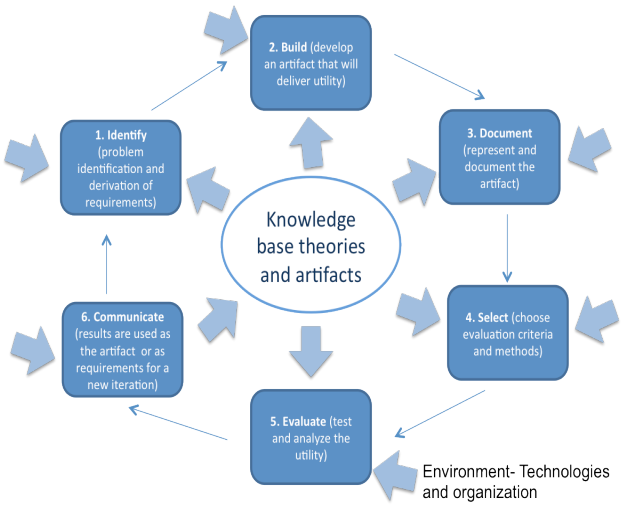



**Figure 1: The basic form of DSR cycles**
Source: adapted from Brocke and Buddendick (2006)

### B. *Build and document*

These phases guided us through the development and representation of a model capable of representing the relevant features of real-life traffic flows. To create a successful model it is essential to classify a particular network's activities, because different user activities produce different traffic patterns and each traffic pattern can be characterized by various parameters. Besides, different traffic can flow in the network at the same time, because users often browse the web, read emails, send text messages, play games, make voice calls and stream videos simultaneously. In this case data traffic is a result of parallel user activities (Varga & Olaszi, 2013).

The conceptual model for this research categorizes traffic in five different types of flows—custom (UDP and TCP flows using various statistical distributions), voice, data, game, and text, and incorporates a mixture of these flows in order to represent real-life traffic. In a communication network traffic flows circulating between a sender and a receiver are characterized by two significant parameters: the size of each transmitted packet and the elapsed time between packet transmissions (as shown in Figure 2). In this research a stochastic model was developed to represent both packet size and time between two consecutive packets in the network. Other significant aspects to consider while defining a model are the facts that packets circulating in the network have different sizes, while the time between two consecutive packets is not always constant. Our model was

implemented in order to emulate these two parameters using various statistical distributions (Constant, Uniform, Pareto, Normal, etc.) to randomize them. The distributions chosen to emulate each type of traffic should be able to capture the relevant and representative characteristic of the emulated traffic.

For the sender, each time period represents the elapsed time between the transmission of the current packet and the transmission of the next packet, while for the receiver each time period represents the elapsed time between the reception of the previous packet and the reception of the current packet.

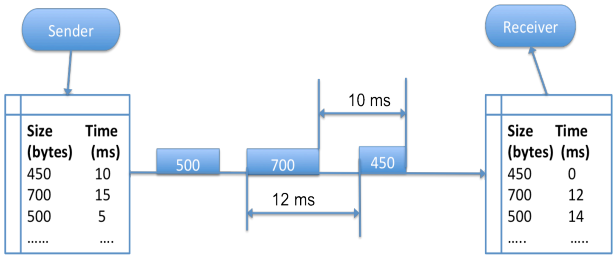



**Figure 2: Example of traffic flow circulating in the network between a sender and receiver**

### C. *Select and evaluate*

After designing and implementing the traffic-generator tool based on the conceptual model defined earlier, the next step is to select the appropriate performance metrics (packet loss, delay and jitter) to represent the service under evaluation. Then the system is tested. The sender and receiver generate a trace file containing detailed information about each packet sent and received during the generation process. Both trace files are then analyzed to determine the functioning of the network or protocol under evaluation.

### D. *Communicate*

This step includes a qualitative analysis in order to verify the efficacy of the tool. In the present research several experiments will be completed both in a laboratory testbed environment and in an actual WMN deployed in a rural area. The results obtained using the MTGawn tool will be compared to results obtained using well-recognized computer-based traffic generators such as D-ITG in the same experiments. In this step, if the results are not satisfactory the cycle is repeated and more experiments are done until the usefulness of the tool has been proved.

## IV. ARCHITECTURE OF THE SYSTEM

The proposed MTGawn system was divided into four main elements: modelling, sender, receiver and analyzer. Each component plays a specific role, but all the components communicate simultaneously to provide the underlying features. The system structure follows a linear form, i.e. the flow is first created, then sent over the network and received at the other endpoint of the network, and then analyzed (see Figure 3).

### A. *Modelling*

The 'modelling' module is in charge of emulating real-life traffic, i.e. it is responsible for defining appropriate stochastic models to generate realistic traffic in terms of packet size and the time between packets.

### B. *Sender*

The 'sender' module is the core function responsible for sending the flows over the network. To do so, it uses the type of transport protocol appropriate for the specific type of flow (UDP or TCP). For example, in case of voice traffic, either RTP (real-time protocol) or cRTP (compressed real-time protocol) packets are created and encapsulated in the UDP packet to carry the voice packet. Both the sender and the receiver module integrate another component ('request manager') whose role is to manage parallel incoming and outgoing traffic involving single or multiple senders or receivers. The request manager's purpose is to allow multiple flows to be sent and received simultaneously.

### C. *Receiver*

The 'receiver' module's role is to receive the flow, manage it and control the parallel incoming flows. The relevant information about each packet received is saved in a file. Each packet includes information such as the flow identifier, the packet identifier, the time sent, the time received and the payload size of the packet.

### D. *Analyzer*

The 'analyzer' module is responsible for analyzing the sent and received flows in order to compute the QoS of the network under consideration.

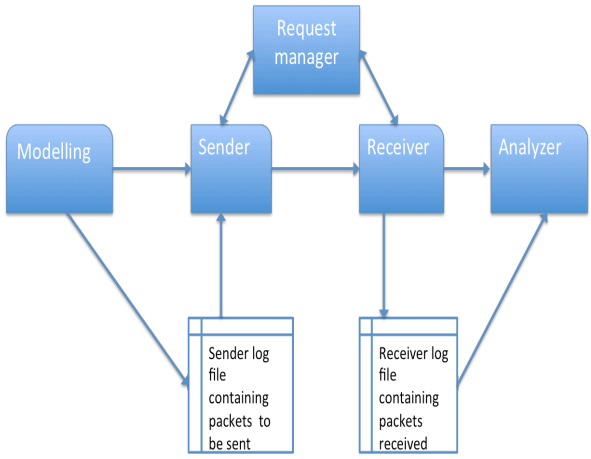



**Figure 3: Basic architecture of the MTGawn system**

## V. RESULTS

The experiments conducted so far in this research focused on estimating the QoS of TCP and UDP in a testbed consisting of a WMN deployed in a laboratory. The testbed to carry out the measurement consisted of a WMN with three mesh potato nodes and two android phones connected to the network wirelessly. Figure **4** depicts the testbed environment.

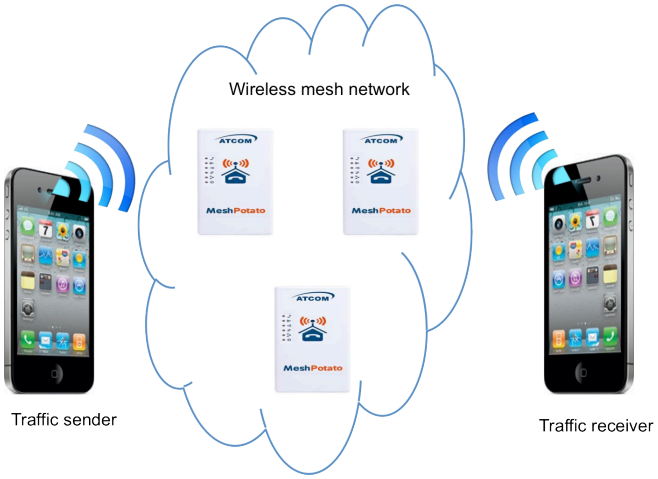



**Figure 4: Testbed environment**

The purpose of these experiments was to generate packets in the network without impacting on the services offered to end users. In addition, this testbed will serve as a platform for assessing the performance of UDP and TCP traffic flows in terms of various pre-defined scenarios without affecting the QoS of other applications. In order to do this the application was deployed on two android devices connected to the testbed WMN. One of the devices was configured as a traffic sender and the other as a traffic receiver.

Different applications (data/voice/multimedia) have different requirements for network service in order to be useful to users—some applications are impaired by message loss and others are impaired by delay or jitter (Marsic, 2013). For this reason, we focused on three performance metrics: delay, jitter and packet loss.

These performance metrics were computed as follow:

i. **Delay:** If $S_i$ is time transmitted for packet *i* and $R_i$ is the time of arrival for packet *i*, then the delay $D_i$ is determined as the difference between the received and transmitted time.

$$D_i = R_i - S_i$$

ii. **Jitter:** The inter-arrival jitter is defined as the difference (D) in packet spacing at the receiver (R) compared to the sender (S) for a pair of packets. For two packets (*i* and *j*), D is expressed as:

$$D(i,j) = (R_j - R_i) - (S_j - S_i) = (R_j - S_j) - (R_i - S_i)$$

$$\textbf{Average Jitter} = \frac{\sum_i^n |D(i-1,i)|}{n}$$

where *n* is the total number of packet sent.

iii. **Packet loss:** The amount of packet loss is the difference between the numbers of transmitted and received packets. If *n* represent the number of packets transmitted and *m* represent the number of packets received, then the percentage of packet loss is defined as:

$$\textbf{Packet loss (\%)} = \left(\frac{n-m}{n}\right) \times 100 = 100 - \frac{m}{n} \times 100$$

For simplicity, time synchronization between the mobile sender and the mobile receiver was not considered during these experiments.

The first DSR cycle focused on the design of an MTGawn prototype for a single TCP and UDP flow using two different distributions—Constant and Uniform—for the representation of both time between packets and packets size. During this phase, a series of experiments was completed:

1. The first experiment was carried out to test the performance of a single UDP flow over the WMN. A total number of 1000 UDP packets were sent at a constant rate of 100 packets per second (a constant time of 10 millisecond (ms) between two consecutive packets) and the size of each packet was equal to 500 bytes.
2. The same experiment was repeated for a single TCP flow. A total number of 1000 TCP packets were sent at a constant rate of 100 packets per second (a constant time of 10 ms between two consecutive packets) and the size of each packet was equal to 500 bytes. Table 1 shows the results obtained in the testbed environment.

TABLE 1: SINGLE TCP AND UDP FLOW WITH CONSTANT DISTRIBUTION FOR BOTH PACKET SIZE AND TIME BETWEEN PACKETS

| | TCP | UDP |
|---|---|---|
| **Minimum delay (ms)** | 1045 | 956 |
| **Maximum delay (ms)** | 1310 | 1002 |
| **Average jitter (ms)** | 1.028 | 0.928 |
| **Packet loss** | 0% | 0% |

3. Another experiment was carried out to evaluate the performance of a single UDP traffic flow using the Uniform distribution. A total number of 1000 UDP packets were sent with uniformly distributed time between packets (between 10 and 20 ms) and uniformly distributed packet sizes (between 500 and 1000 bytes).

4. The previous experiment was repeated for a single TCP flow. A total number of 1000 TCP packets were sent with uniformly distributed time between packets (between 10 and 20 ms) and uniformly distributed packet sizes (between 500 and 1000 bytes). Table 2 shows how the testbed WMN performs with the Uniform distribution.

TABLE 2: SINGLE TCP AND UDP FLOW WITH UNIFORM DISTRIBUTION FOR BOTH PACKET SIZE AND TIME BETWEEN PACKETS

| | TCP | UDP |
|---|---|---|
| **Minimum delay (ms)** | 1151 | 1008 |
| **Maximum delay (ms)** | 1276 | 1115 |
| **Average jitter (ms)** | 1.235 | 1.017 |
| **Packet loss** | 0% | 0% |

In order to emulate multiple user activities simultaneously, a new DSR cycle was iterated in order to estimate the performance of testbed WMN networks under more realistic conditions. The purpose of this phase was to generate complex and varied TCP and UDP traffic sources

representative of a wide range of traffic conditions. Different traffic patterns were produced to simulate different user activities. In this case, four distinctive traffic patterns were generated and sent simultaneously between the mobile sender and mobile receiver; the results are given in Table 3.

TABLE 3: MULTIPLE UDP AND TCP FLOWS SENT SIMULTANEOUSLY

| | UDP (a) | UDP (b) | TCP (c) | TCP (d) |
|---|---|---|---|---|
| **Minimum delay (ms)** | 3170 | 3050 | 4178 | 3620 |
| **Maximum delay (ms)** | 3450 | 3152 | 4276 | 4515 |
| **Average jitter (ms)** | 2.581 | 2.397 | 3.032 | 3.593 |
| **Packet loss** | 39% | 26% | 0% | 0% |

- One flow for UDP traffic pattern (a): using a Constant distribution for packet size (500 bytes) at a uniform rate of between 1000 and 2000 packets per second.
- One flow for UDP traffic pattern (b): using the Uniform distribution for packet size (between 500 and 1000 bytes) at a constant rate of 1000 packets per second.
- One flow for TCP traffic pattern (c): using the Constant distribution for packet size (500 bytes) at a uniform rate of between 1000 and 2000 packets per second.
- One flow for TCP traffic pattern (d): using the Uniform distribution for packet size (between 500 and 1000 bytes) at a constant rate of 1000 packets per second.

From the results of these experiments general conclusions can be drawn about how UDP and TCP traffic flows perform under diverse scenarios. According to the above results it can be concluded that packet loss was observed only in cases of UDP traffic, while larger delay and jitter were observed in TCP traffic.

Another observation is the increase in delay and packet loss when many flows are sent simultaneously. This observation is reasonable even in a real-life scenario, where the growth of users using the network leads to the growth of traffic flowing through the network. This growth in traffic load generally results in non-negligible service delay and packet loss.

## VI. CONCLUSION AND FUTURE WORK

This paper presents preliminary results obtained during the development of a prototype for a mobile traffic generator called MTGawn. The tool is intended to ease feasibility testing and performance evaluation of a rural WMN by implementing a 'stripped down' version of a packet generation and monitoring system to a mobile platform with functionality found in common open source tools. The primary phase of the design process of this prototype involved modelling traffic patterns illustrative of realistic TCP and UDP traffic flows. Each traffic pattern was characterized by packet size and inter-arrival distributions. For this purpose, it is important to appropriately depict both parameters. These parameters were chosen such that they helped to build an effective traffic model for a given traffic pattern (Varga & Olaszi, 2013). From this perspective, a model was defined to represent traffic patterns in different scenarios by using various distributions. Consequently, this mobile system is able to generate diverse TCP and UDP traffic patterns over any wireless network interface on a mobile device and calculate standard QoS metrics such as packet loss, delay and jitter. All experimentation was conducted in a laboratory testbed consisting of mesh network nodes.

To move toward *in situ* network activities, the next steps involve the following: (a) the modelling and generation of other type of flows such as VoIP and games; and (b) comparing the packet generation and analysis capabilities on mobile devices against a standard PC-based performance tool such as D-ITG. The functionality of (a) can be added to the testbed, while (b) can also be accomplished using the testbed. Finally, in-the-field testing will be pursued on an actual rural WMN to assess how the mobile prototype behaves under real physical conditions.

## VII. REFERENCES


Avallone, S., Pescape, A., & Ventre, G. (2003). Distributed Internet Traffic Generator (D-ITG): analysis and experimentation over heterogeneous networks. *International Conference on Network Protocols (ICNP 2003 Poster Proceedings), 7 November 2003*. Atlanta, GA. IEEE Computer Society.

Brocke, J. V., & Buddendick, C. (2006). Reusable conceptual models–requirements based on the design science research paradigm. In S. Chatterjee and A. Hevner (Eds.), *First International Conference on Design Science Research in Information Systems and Technology (DESRIST 2006), February 24-25 2006* (pp. 576-604). Claremont, CA.

Carrano, R. C., Bletsas, M., & Magalhães, L. C. (2007). Mesh networks for digital inclusion-testing OLPC's XO mesh implementation. *Internacional do Workshop do 8 Fórum Internacional do Software Livre (FI3.SL)*. Porto Alegre.

Cheng, X., Mohapatra, P., Lee, S.-J., & Banerjee, S. (2008). Performance evaluation of video streaming in multihop wireless mesh networks. *Proceedings of the 18th International Workshop on Network and Operating Systems Support for Digital Audio and Video* (pp. 57-62). New York, NY: ACM.

Gunashekar, S. D., Das, A., Erlebach, T., & Warrington, E. M. (2014). An experimental study of small multi-hop wireless networks using chirp spread spectrum. *Wireless Networks*, *20*(1), 89-103.

Hamidian, A., Palazzi, C. E., Chong, T., Navarro, J., Korner, U., & Gerla, M. (2009). Deployment and evaluation of a wireless mesh network. *Second International Conference on Advances in Mesh*

*Networks, MESH 2009* (pp. 66-72). Athens, Greece: IEEE.

Marsic, I. (2013). *Computer networks: Performance and quality of service.* New Brunswick, NJ: Department of Electrical and Computer Engineering, Rutgers University.

Nicola, B., Giordano, S., Procissi, G., & Secchi, R. (2005). BRUTE: A high performance and extensible traffic generator. *Proceedings of SPECTS* (pp. 839-845). Philadelphia, PA.

Sichitiu, M. L. (2005). Wireless mesh networks: Opportunities and challenges. *Proceedings of World Wireless Congress*, *2*. San Francisco, CA.

Sommers, J., & Barford, P. (2004). Self-configuring network traffic generation. *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement* (pp. 68-81). New York, NY: ACM.

Sommers, J., Kim, H., & Barford, P. (2004). Harpoon: A flow-level traffic generator for router and network tests. *ACM SIGMETRICS Performance Evaluation Review, 32*(1), 392.

Varga, P., & Olaszi, P. (2013). LTE core network testing using generated traffic based on models from real-life data. *IEEE International Conference on Advanced Networks and Telecommuncations Systems (ANTS), 2013* (pp. 1-6). Chennai, India: IEEE.

**ACKNOWLEDGEMENTS**

Telkom, Cisco, Aria Technologies and THRIP provide financial support. This work is based on research supported in part by the National Research Foundation (NRF) of South Africa (grant-specific unique reference number UID) 75191. The grantholder acknowledges that opinions, findings and conclusions or recommendations expressed in any publication generated by NRF-supported research are those of the author(s) and that the NRF accepts no liability whatsoever in this regard.

**Ghislaine L. Ngangom** received a BSc in 2009 from the University of Yaoundé I (Cameroon), completed a BSc Honours at the University of the Western Cape (UWC) in 2012 and is presently an MSc student in Computer Science at UWC. Her research interests include mobile computing, network performance optimization, Internet programming and mobile security.

# Dispersive Characteristics for Broadband Indoor Power-Line Communication Channels

Modisa Mosalaosi and Thomas J. O. Afullo
Department of Electrical, Electronic & Computer Engineering
University of KwaZulu-Natal, Private Bag X54001, Durban 4001
Tel: +27 31 2602730, Fax: +27 31 2602740
email: {201508647@stu.ukzn.ac.za, Afullot@ukzn.ac.za}

**Abstract-A study of time dispersion of multitudes of indoor powerline communications (PLC) channels in the 1-30 MHz band is presented in this paper. The dispersive characteristics of the PLC channel are derived from the measured complex channel transfer function (CTF) by evaluating its impulse response (IR). The impulse response provides a wideband characterization of the propagation channel and provides the basis for which the channel power-delay profile (PDP) is determined. The expected degree of dispersion is derived from the PDP of the channel, hence time-delay spread parameters such as the root mean square (rms) delay spread, mean excess delay, maximum excess delay and the first arrival delay are obtained. The paper thus presents the statistics of time delay spread parameters for all measured PLC channels. Finally, the paper presents the coherence bandwidth and evaluates its dependence on the rms delay spread. Results show that the two parameters have an inverse relationship.**

**Index Terms—rms delay spread, coherence bandwidth, power delay profile**

## I. INTRODUCTION

The dramatic increase in the number of digital radio communication systems within indoor environments has raised the demand for broadband characterization of indoor PLC channels. Signal time dispersion remains the core study issue owing to its limitation of the achievable maximum symbol rate without inter-symbol interference (ISI) [1]. The impulse response (IR) of the PLC channel along with other parameters such as the mean delay $(\tau_e)$ and the root mean square (rms) delay spread $(\tau_{rms})$ [2], [3], are usually used to characterize the channel's time dispersion. The frequency selectivity nature of the PLC channel can also be described by the coherence bandwidth $(B_c)$, a parameter with a close relation with $\tau_{rms}$ [4].

Previous works have investigated the time delay spread parameters for indoor PLC channels in different parts of the world and for different frequency bands [5], [6], [7], [8]. Nonetheless, few experimental results exist that describe these parameters. In this paper, an experimental study on the dispersive characteristics of indoor PLC channels is presented. The paper also presents the relationship between $B_c$ and $\tau_{rms}$. The results are based on broadband measurements taken at various environments at our University. The channel transfer characteristics depends on the loading of the channel itself. Typically, the type of loads includes fluorescent bulbs, small-to-medium size air conditioners, personal computers, laboratory equipment, and small-size motors used to perform curricular tests. The measurement system based on a vector network analyzer (VNA) was used to measure the complex frequency response of the PLC channel in the 1-30 MHz frequency band. Both $B_c$ and IR are obtained from the channel frequency response (CFR). The statistics of the delay parameters are given as well as their 90$^{th}$ percentiles. The coherence bandwidth is given at 0.9 correlation level and it is found to vary inversely with $\tau_{rms}$.

## II. MEASUREMENT SETUP & FUNDAMENTALS

A Rhode & Schwartz ZVL13 VNA was used to measure the channel complex frequency response in the 1-30 MHz frequency band. The measurement setup is as shown in fig. 1. The transmitting and receiving ends of the PLC channel are connected to the transmitter and receiver through couplers respectively. It is necessary to use coupling circuits to protect the measurement equipment from damage when connecting it to the electrical network to launch or receive the information signal. The measured results represents the frequency response of all the devices connected between the two ports of the VNA, comprising the channel, measurement cables, coupling circuitry, and the frequency response of the VNA itself. The effect of the measuring system on measurements needs to be eliminated; thus a calibration is required to be carried out prior to channel measurements. With this purpose, the couplers were connected directly between the two ports of the VNA, with the frequency response $H(f,t)_{system}$ being measured. This result was automatically subtracted from all subsequent channel measurements, minimizing the effects of the measuring system on channel transfer characteristics. The true channel frequency response takes the following form [4]:

$$H(f,t)_{channel} = \frac{H(f,t)_{measured}}{H(f)_{system}} \quad (1)$$

## III. CHANNEL IMPULSE RESPONSE (CIR)

The random and complex nature of the PLC channel can be characterized using the impulse response approach. The channel impulse response *h(t)* can be determined by means of inverse Fourier transform (IFT) derived from the absolute value and phase of a measured transfer function [9]. It provides a broadband characterization of the propagating

channel, and provides necessary information for the analysis of radio transmission through the channel. Thus the time domain channel impulse response including cable loss is [10]:

$$h(t) = \sum_{i=1}^{N} I_i \,.\, [A_t(t, vT_i) \otimes \delta(t - T_i)] \qquad (2)$$

where $I_i$, and $T_i$ are magnitude and delay of the $i$th path respectively. $v$ is the TEM wave propagation speed in the cable which can be calculated according to the permittivity of the insulating material of the cable $(v = c/\sqrt{\varepsilon_r})$. $A_t(t, vT_i)$ is the cable loss effect in the time domain evaluated as the inverse Fourier transform of the cable attenuation [10]. A sample channel obtained through the IFFT of the channel transfer function (CTF) is shown in figure 2. Due to multiple reflections experienced by the signal as it propagates through the network, echoes appear at the receiver. Thus, multiple delayed versions of the transmitted signal appear at the receiver with reduced amplitudes as shown in figure 2.

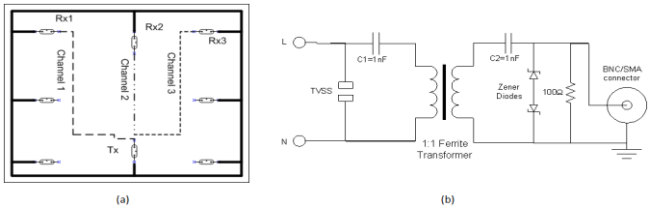



Figure 1: (a) Channel definition & (b) Coupling Circuitry

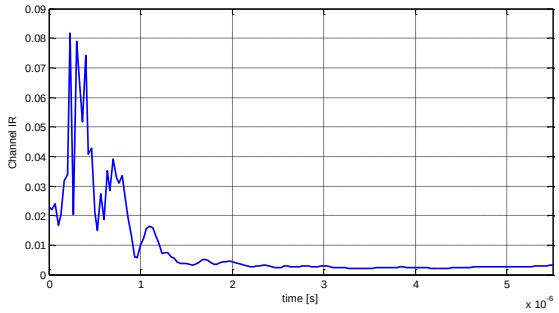



Figure 2: Sample Channel IR (absolute)

## IV. POWER DELAY PROFILE (PDP)

Multipath propagation causes severe dispersion of the transmitted signal [11]. The severity of the dispersion is determined through the channel power delay profile (PDP). The PDP provides an indication of the distribution of the transmitted power over various paths in a multipath environment. The channel PDP is considered as the spatial average of $|h(t)|^2$, a density function of the form [9]:

$$p(\tau) = \frac{|h(\tau)|^2}{\int_{-\infty}^{\infty} |h(\tau)|^2 \, d\tau} \qquad (3)$$

An ensemble of PDPs is built each representing a multipath propagation PLC channel. A typical plot of the power delay profile is shown in figure 3, corresponding to the IR of figure 2. Time delay multi-path channel parameters are derived from the PDP. Time dispersion varies widely in a PLC channel due to the multiple reflections in the power network and random loading profiles; resulting in random channel response.

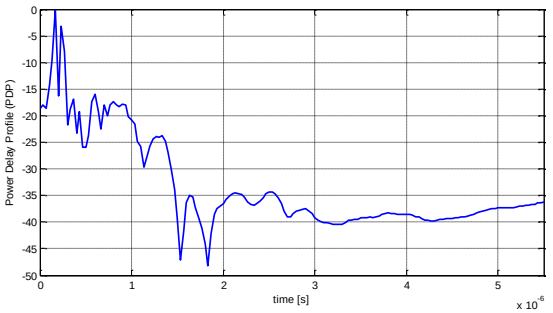



Figure 3: A typical PDP plot

## V. TIME-DELAY SPREAD PARAMETERS

Since time dispersion is dependent on factors such as the network topology, loading characteristics, transmitter-receiver distance, just to name a few; some parameters that can be used to grossly quantify the PLC multipath channel are described [11].

### A. *First-Arrival Delay* $(\tau_A)$

This is the first arrival delay. This delay corresponds to the arrival of the first transmitted signal at the receiver. It translates to the minimum possible propagation path delay from transmitter to receiver. This parameter serves as a reference i.e. all other delay parameters are measured relative to it. Any delay measured beyond this reference is considered excess delay.

### B. *Mean Excess Delay* $(\tau_e)$

The mean excess delay represents the first moment of the power-delay profile with respect to the first arrival delay. It can be computed as follows:

$$\tau_e = \int (\tau - \tau_A) P(\tau) d(\tau) \qquad (4)$$

### C. *RMS Delay Spread* $(\tau_{rms})$

This is the square root of the second central moment of the power-delay profile. It is the standard deviation about the mean excess delay and is expressed as follows:

$$\tau_{rms} = \left[ \int (\tau - \tau_e - \tau_A)^2 p(\tau) d\tau \right]^{1/2} \qquad (5)$$

The RMS delay spread is a good measure of the multipath spread. The possibility of inter-symbol interference (ISI) is determined from this parameter. Signal delays of high magnitude (relative to the shortest path) with long delay time contribute immensely to the $\tau_{rms}$ [11]. Regardless of the shape of the PDP, the dispersion effects on digital receiver performance are related only to the RMS delay

spread. ISI will be avoided as long as the RMS delay spread is smaller compared to the symbol period (T) of the digital modulation. Data rates for transmission can also be estimated using this parameter.

### *D. Maximum Excess Delay ($\tau_m$)*

The maximum excess delay is specified as the excess delay for which $p(\tau)$falls below a specified threshold of the signal. Researchers have considered different threshold levels, but mostly consider -30dB [12] with respect to the peak value and the same is adopted in this work. The lower signal levels are then processed as noise. Consequently, the mean excess delay ($\tau_e$) and the RMS delay spread ($\tau_{rms}$) are calculated based on channel time coefficients lower than $\tau_m$.

## VI. RESULTS ANALYSIS

The time dispersion characteristics of PLC channels are presented statistically in table 1. The analysis evaluates the minimum, maximum, mean, standard deviation and the 90$^{th}$ percentiles of each time-delay parameter. Alongside our results is that of an extensive measurement campaign carried out in France for comparison (here shown with a subscript (1)). It should be noted that their measurements extends to frequencies up to 100 MHz, though for the same threshold signal level, similar delay characteristics are expected and it is macroscopically the case here.

Table 1: Statistics of the time delay parameters

| $\tau$ (us) | Min (us) | Max (us) | Mean (us) | STDev (us) | 90% above | 90% below |
|---|---|---|---|---|---|---|
| $\tau_A$ | 0.033 | 0.534 | 0.179 | 0.143 | 0.067 | 0.37 |
| $\tau_{A1}$ | 0.010 | 0.410 | 0.152 | 0.097 | 0.050 | 0.30 |
| $\tau_e$ | 0.00082 | 0.97 | 0.163 | 0.216 | 0.020 | 0.40 |
| $\tau_{e1}$ | 0.00030 | 0.88 | 0.182 | 0.157 | 0.025 | 0.36 |
| $\tau_{rms}$ | 0.0215 | 1.16 | 0.287 | 0.222 | 0.04 | 0.55 |
| $\tau_{rms1}$ | 0.026 | 1.039 | 0.309 | 0.212 | 0.06 | 0.60 |
| $\tau_m$ | 0.534 | 4.535 | 1.749 | 0.867 | 0.83 | 3.27 |
| $\tau_{m1}$ | 0.18 | 6.26 | 2.228 | 1.327 | 0.55 | 3.81 |

In our study, the first arrival delay ($\tau_A$) was observed to have a minimum value of 0.033µs, standard deviation of 0.143µs, and a mean value of 0.179µs. For 90% of the time, the value of $\tau_A$ was found to be less than 0.37µs and above 0.06µs. On the other hand, the mean excess delay ($\tau_e$) has a mean value of 0.163µs, and a standard deviation of 0.216µs. Though $\tau_e$ is confined between 0.00082µs and 0.97µs, it is found that for 90% of the time, its values are between 0.02µs and 0.4µs.

Concerning the maximum excess delay ($\tau_m$) 90% of the measured channels were observed to lie between 0.83µs and 3.27µs. Its mean and standard deviation were found to be 1.749µs and 0.867µs respectively. Of the measured channels, 90% of them exhibit an RMS delay spread between 0.04µs and 0.55µs. The percentiles for the time delay parameters are displayed in figure 4 in the form of cumulative distributions (CDFs). Other percentiles not discussed in this paper can easily be estimated from the CDFs.

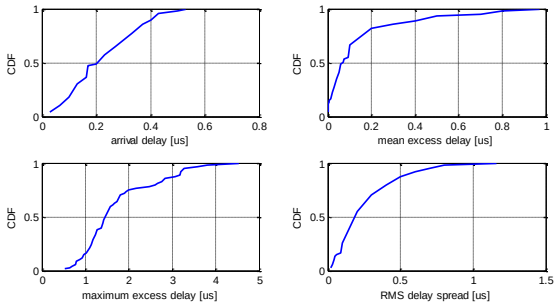



Figure 4: Cumulative distributions (CDFs) of the time delay parametrs

## VII. COHERENCE BANDWIDTH

An important parameter in characterizing radio communication channels is the coherence bandwidth. In this paper, we present an analysis of the coherence bandwidth in PLC networks. When designing robust and reliable communication systems, a significant amount of effort is placed on choosing modulation, coding and receiver architecture schemes that mitigates the deleterious effects of the propagating channel [13]. The PLC propagation environment is characterized by multipath effects which results in significant nulls in the amplitude frequency response. One measure of the varying frequency response is the coherence bandwidth ($B_c$).

In the frequency domain, the magnitude spectrum has spectral peaks that are quasi-constant over a minimum band that is the inverse of the maximum delay spread, the same for the phase spectrum where it is linear only in such a band. The coherence bandwidth statistically quantifies this band and is therefore a function of the RMS delay spread. It is the measure of the magnitude correlation between the channel responses at two spaced frequencies, thus, it statistically quantifies the range of frequencies over which the frequency correlation function (FCF) can be considered flat [9]. The frequency selective nature of the PLC channel can be described in terms of the auto-correlation function for a wide sense stationary uncorrelated scattering (WSSUS) channel. The frequency correlation function is given by [9]:

$$R(\Delta f) = \int_{-\infty}^{\infty} H(f)H^*(f+\Delta f)df \qquad (6)$$

Where *H(f)* is the complex transfer function of the channel, $\Delta f$ is the frequency shift and * denotes the complex conjugate. Frequency correlation functions obtained for five sample transmitter-receiver scenarios are shown in figure 5. We observe the rapid degradation of the FCF with respect to frequency separation. There is no definitive value of correlation that has been put forward for specification, but generally accepted coefficients are 0.5, 0.7, and 0.9 [5], [6], [7], [8], [9]. In this work we have considered the latter, which will further be referred to as $B_{0.9}$. Due to the presence of multipath replicas at the receiver in PLC channels; the decrease of FCF with increasing frequency is non-monotonic. In figure 5, the upper graph (blue) represents a good channel due to its high coherence bandwidth for a given RMS delay spread.

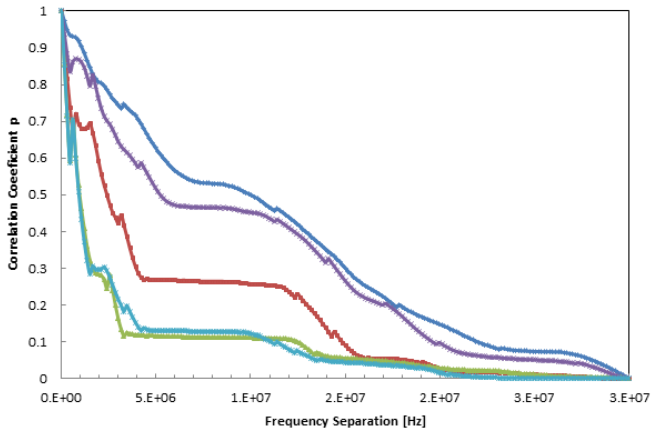



Figure 5: Frequency correlation functions of sample measured channels

The worst channel would be the lower-most graph (green) as it exhibits the lowest coherence bandwidth for a given RMS delay spread. A good channel can be assumed to have the least multipath contributions. The correlation coefficient was computed using the smallest frequency separation, 150 kHz in this work. The estimate of the coherence bandwidth is derived from the FCF graph.

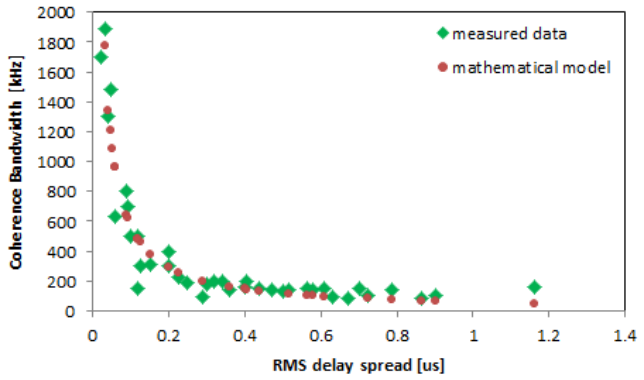



Figure 6: Scatter plot of coherence bandwidth against RMS delay spread

Figure 6 shows the scatter plot of the RMS delay spread against the coherence bandwidth $B_{0.9}$ for the measured PLC channels. An approximation of their relationship (shown in red) was determined to be:

$$B_{0.9}(kHz) = \frac{60}{\tau_{rms}(\mu s)} \qquad (7)$$

The results show that RMS delay spread values less than 0.09µs achieve a coherence bandwidth of at least 600 kHz. On the other hand, in the range 0.09µs – 1.16µs, the coherence bandwidth is between 80 kHz and 600 kHz. In terms of system design, desired values of $\tau_{rms}$ are those that result in high coherence bandwidth as this translates to faster symbol transmission rates [14]. The coherence bandwidth of the channel is particularly relevant to frequency-hopping spread spectrum (FHSS) systems [15] and other multi-carrier systems such as OFDM. Coherence bandwidth is one of the key parameters which provides a good indication towards the possibility for achieving the PLC wideband performance as envisaged for PLC networks. In this frequency range, the transmission rates of 200Mbs are expected at the physical layer with about 80Mbs at the MAC layer [17].

## VIII. Conclusions

An indoor PLC channel measurement campaign has been established to determine the time dispersive characteristics of these channels. This paper includes the statistics of time delay parameters as well as coherence bandwidth and its relation to RMS delay spread in the frequency range up to 30 MHz. We have observed the inverse relationship between coherence bandwidth and RMS delay spread as given in equation (7). The frequency correlation function is also found to decrease rapidly with frequency separation. The 90$^{th}$ percentile of the RMS delay spread was found to be between 0.04µs and 0.55µs with a mean of 0.287µs and standard deviation of 0.222µs. 80% of the measured channels exhibit estimated values of coherence bandwidth between 600 kHz and 90 kHz. We thus, however, take note that for improved accuracy and consistency a lot more data still needs to be captured and incorporated into our analysis. Also, for improved coherence bandwidth estimation, smaller frequency steps are required for increased data points within the band of interest.

Comparatively, the relationship between coherence bandwidth and RMS delay spread as shown in equation (7) is very close to that obtained by [9]. The constant in the numerator (60 in this case and 55 in [9]) is expected to differ from site to site but with more measurement campaigns, this number could be defined as bound within a reasonable range just like it is the case with wireless networks. The comparisons in table 1 also show closeness in terms of bounds for other time delay parameters for indoor PLC channels. We also take note of the difference in bandwidth consideration in our case and that of [9], which considers frequencies up to 100 MHz.

**Modisa Mosalaosi** received his undergraduate degree in 2010 from the University of KwaZulu-Natal and is presently studying towards his Master of Science degree at the same institution. His research interests include electromagnetic propagation, satellite communication, radar, and power line communications.

## References

[1] J. G. Proakis, "Digital Communications", 2nd ed, New York, McGraw Hill, 1989.

[2] W. C. Jakes, "Microwave Mobile Communications", New York, Willey, 1974.

[3] W. C. Y. Lee, "Mobile Communications Design Fundamentals", 2nd ed, New York, Willey, 1993.

[4] M. S. Varela and M. G. Sanchez, "RMS Delay and Coherence Bandwidth Measurements in Indoor radio Channels in the UHF Band", *IEEE Transactions on veh., Tech.*, vol. 50, No. 2, March 2001.

[5] H. Philipps, "Development of a Statistical Model for Power Line Communications Channels", *Proceedings of ISPLC 2000*, Limerick, Ireland, April 2000.

[6] V. degardin, M. Lienard, A. Zeddam, F. Gauthier, and P. Degauque, "Classification and Characterization of

Impulse noise on Indoor Power Lines and for Data Communications', *IEEE Transactions on Consumer Electronics*, vol. 48, November 2002.

[7] T. Esmalian, F. R. Kschischang, and P. Glenn Gulak, "In-building Power Lines as High Speed Communication Channels: Channel Characterization and a test Channel Ensemble", *Int. J. Comm. Sys. 2003*.

[8] T. V. Pasad, S. Srikanth, C. N. Krishnan, and P. V. Ramakrishna, "Wideband characterization of Low Voltage outdoor Powerline Communication Channels in India", *International Symposium on Power-Line Communications (ISPLC'2001)*, Sweden, April 2001.

[9] M. Tlich, G. Avril, A. Zeddam, "Coherence Bandwidth and its Relationship with the RMS delay spread for PLC Channels using Measurements up to 100 MHz", *1st International Home Networking Conference (IHN 2007)*, Paris-France, 10-12 December 2007.

[10] Bo Tan, John S. Thompson, "Power line Communications Channel Modeling Methodology Based on Statistical Features", *Proceedings of the IEEE*, March 2012.

[11] T. K. Sarkav, Zhong Ji, Kyungjung Kim, A. Medouri, and M. Salazar-Palma, "A Survey of Various Propagation Models for Mobile Communication", *IEEE Antenna and Propagation Magazine*, vol. 45, No. 3, June 2003.

[12] Mohamed Tlich, Ahmad Zeddam, Fabienne Moulin, and Frederic Gauthier, "Indoor Power-Line Communications Channel Characterization up to 100 MHz-Part II: Time-Frequency Analysis", *IEEE Transactions on Power delivery*, vol. 23, No. 3, July 2008

[13] H. R. Anderson and J. P. McGeehan, "Direct Calculation of Coherence Bandwidth in Urban Microcells Using a Ray-Tracing Propagation Model", *IEEE Int. Symp. On Wireless Networks*, vol. 1, pp. 20-24, September 1994.

[14] H. Lutz, J. Lampe and Johannes B. Huber, "Bandwidth Efficient Power Line Communications Based on OFDM", *AEU Int. J. Electr. Commun.*, 1999.

[15] D. J. Purle, A. R. Nix, M. A. Beach, and J. P. McGeenhan, "A Preliminary Performance Evaluation of a linear frequency Hopped Modem", *proceedings of the 1992*, veh. Tech. Society conf., Denver, pp. 120-124, May 1992.

[16] "Seventh Framework Programme: Theme 3. ICT-213311 OMEGA, PLC Channel Characterization and Modelling." [Online]. Available: http://www.ict omega.eu/fileadmin/documents/deliverables/Omega D3.2 v1.1.pdf.

[17] HomePlud Powerline Allience,"HomePlug AV Specification, Version 1.0.05", October 2006.

# Performance Analysis of Dynamic Switching between Spatial Multiplexing and Diversity over Rayleigh Fading Channels in MIMO-OFDM Systems using QPSK Modulation Scheme

Jamal R. Elbergali[1], Neco Ventura[2]
Electronic Engineering Department
College of Industrial Technology[1], Misurata-Libya
Tel: +218 91 3220149
and Department of Electrical Engineering
University of Cape Town[2], Private Bag X3, Rondebosch, South Africa
Email: jelbergali@yahoo.com[1]; neco@crg.ee.uct.ac.za[2]

**Abstract-Multiple Input Multiple Output (MIMO) systems are wireless systems with multiple antenna elements at both ends of the link. Wireless communication systems demand high data rate and spectral efficiency with increased reliability. MIMO systems have been popular techniques to achieve these goals because increased data rate is possible through spatial multiplexing scheme and diversity. In this paper, we propose a dynamic MIMO mode switching scheme between spatial multiplexing (SM) and diversity. The proposed dynamic switching based on computing the instantaneous channels conditions and compare it with a chosen threshold level. A tradeoff between spectral efficiency and Bit Error rate (BER) performance should be considered. We analyze the spectral efficiency and the (BER) performance of a switching algorithm between SM and diversity for an OFDM-MIMO system over Rayleigh Fading Channel, using QPSK Modulation scheme.**

**Index Terms— SNR, BER, QPSK, MIMO, OFDM, Modulation, spectral efficiency, and spatial multiplexing (SM).**

## I. INTRODUCTION

The choice of an appropriate modulation and multiple-access technique for mobile wireless data communications is critical to achieving good system performance. In particular, typical mobile radio channels tend to be dispersive and time-variant, and this has generated interest in multicarrier modulation. In general, multicarrier schemes subdivide the used channel bandwidth into a number of parallel subchannels. Ideally the bandwidth of each subchannel is such that they are, ideally, each non-frequency-selective (i.e. having a spectrally flat gain); this has the advantage that the receiver can easily compensate for the subchannel gains individually in the frequency domain.

Orthogonal Frequency Division Multiplexing (OFDM) is a special case of multicarrier transmission where the non-frequency-selective narrowband subchannels, into which the frequency-selective wideband channel is divided, are overlapping but orthogonal. This avoids the need to separate the carriers by means of guard-bands, and therefore makes OFDM highly spectrally efficient. The spacing between the subchannels in OFDM is such that they can be perfectly separated at the receiver. This allows for a low complexity receiver implementation, which makes OFDM attractive for high-rate mobile data transmission such as the Long Term Evolution (LTE) downlink.

Multiple input multiple output (MIMO) is a wireless technology which enables the use of multiple transmitting and receiving antennas to transfer more data in less time. MIMO system takes advantage of the spatial diversity that is obtained by spatially separated antennas in a dense multipath scattering environment [1].

OFDM transforms frequency selective MIMO channels into set of parallel frequency flat MIMO channels and increases frequency efficiency. MIMO-OFDM technology has been researched as the infrastructure for next generation wireless networks. A major drawback of wireless communication system is the effect of fading. Fading occurs due to multipath propagation and shadowing from obstacles affecting wave propagation. To overcome the detrimental effects of fading, multiple copies of data are transmitted from transmitter to receiver [2, 3].

Today's wireless communication systems demand high data rate and spectral efficiency with increased reliability. MIMO systems have been popular techniques to achieve these goals through both spatial multiplexing and diversity schemes [1].

The rest of the paper is structured as follows. In section II, we discuss the background followed by the system model description in section III. In section IV, we explain our simulation results and observations. Finally section V concludes the paper.

## II. BACKGROUND

MIMO systems are wireless systems with multiple antenna elements are used at both the transmitter and receiver, and can be used to:

1- Increase the system reliability (decrease the bit or packet error rate).

2- Increase the achievable data rate and hence system capacity.
3- Increase the coverage area.
4- Decrease the required transmit power.

However, these four desirable attributes usually compete with one another; for example an increase in data rate often will require an increase in either the error rate or transmit power [4].

MIMO systems can be used for beamforming, diversity combining, or spatial multiplexing. The first two applications are the same as for the smart antennas, while spatial multiplexing is the transmission of multiple data streams on multiple antennas in parallel, leading to a substantial increase in capacity. MIMO technology and turbo coding are the two most prominent recent breakthroughs in wireless communication. MIMO technology promises a significant increase in capacity.

Depending on the availability of multiple antennas at the transmitter and/or the receiver, such techniques are classified as Single-Input Multiple-Output (SIMO), Multiple-Input Single-Output (MISO) or MIMO. When a multi-antenna terminal is involved, a full MIMO link may be obtained, although the term MIMO is sometimes also used in its widest sense, thus including SIMO and MISO as special cases [5].

The use of multiple antennas allows independent channels to be created in space and is one of the most interesting and promising areas of recent innovation in wireless communications. The spatial diversity, which can be created without using the additional bandwidth that time and frequency diversity both required. In addition to providing spatial diversity, antenna arrays can be used to focus energy (beamforming) or create multiple parallel channels for carrying unique data streams (spatial multiplexing).

Let us consider a MIMO channel (Fig. 1) with $N$ transmit and $M$ receive antennas (Note if $N = 1$, it is Single Input Multiple Output SIMO, if $M = 1$, it is Multiple Input Single output MISO, and if $N = M = 1$, it is a Single input Single Output SISO system). There are $N \times M$ paths and each path has a channel response denoted by $h_{ji}$, which is between $j^{th}$ receiver and $i^{th}$ transmitter.

The MIMO channel ($H$) is shown below,

$$H = \begin{bmatrix} h_{11} & \cdots & h_{1N} \\ \vdots & \ddots & \vdots \\ h_{M1} & \cdots & h_{MN} \end{bmatrix} \quad (1)$$

where the values $h_{ji}$ are the channel response of the paths consisting of complex-Rayleigh elements [11].

Now if the transmitted signal is,

$$X = [x_1, x_2, \ldots\ldots, x_N\ ]^T \quad (2)$$

where $x_i$ is the transmitted symbol using Quadrature Phase Shift Keying (QPSK) modulation scheme.

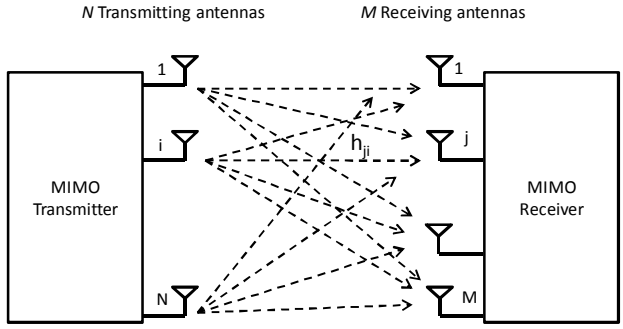



**Fig. 1: Simplified transmitted model for MIMO system with *N* transmit antennas and *M* receive antennas, giving rise to an *M* X *N* channel matrix, *MN* links.**

The signal received at the receive antenna is as follows:

$$Y_{MX1} = H_{MXN} X_{NX1} + n_{MX1} \quad (3)$$

where $(n)$ is the noise vector consisting complex-Gaussian elements with zero mean and variance $\sigma_n^2$. Sufficient antenna separation (typically half the carrier wavelength) makes elements of $h_{ji}$ independent, zero-mean, complex Gaussian random variables (Rayleigh fading). However, at a given time, $h_{ji}$ varies over frequency and time depending on multipath and Doppler spread respectively [6, 11].

### *A. Spatial Multiplexing*

From data rate standpoint, the most exciting type of MIMO communication is spatial multiplexing (SM), which refers to breaking the incoming high rate-data streams into $N$ independent data streams. Assuming that the streams can be successfully decoded, the nominal spectral efficiency is thus increased by a factor of $N$.
The standard mathematical model for SM is similar to what was used for space/time coding.

$$y = Hx + n \quad (4)$$

Using basic linear algebra argument, it is straightforward to confirm that decoding $N$ streams is theoretically possible when there exist at least $N$ nonzero eigenvalues in the channel matrix, that is rank$(H) \geq N$ [7, 8].

SM can be performed with or without channel knowledge at the transmitter. We consider the principle open-loop techniques; we always assume that the channel is known at the receiver. The open-loop technique for spatial multiplexing attempts to suppress the interference that result from all $N$ streams being received by each of the $M$ antennas [9, 10].

### *B. Linear Detectors*

In MIMO communication, the linear detectors can capable of recovering the transmitted vector $x$. The most obvious such detector is the zero-forcing (ZF) detector, which sets the receiver equal to the inverse of the channel.
Now, let the transfer function of the ZF detector as following,

$$G_{zf} = (H^*H)^{-1}H^* \quad (5)$$

Or,

$$G_{zf} = H^{-1}, \text{ for } N = M \quad (6)$$

The ZF detector completely removes the spatial interference from the transmitted signal, giving an estimated receiver vector,

$$\hat{x} = G_{zf}y = x + (H^*H)^{-1}H^*n \quad (7)$$

Or,

$$\hat{x} = x + H^{-1}n, \text{ for } N = M \quad (8)$$

Inverse of $H$ might boost up the noise as bad subchannels that have lower eigenvalues are inverted. This can easily amplify the noise [11].

### C. Switching Between Diversity and Multiplexing

In order to achieve the reliability of diversity and the high raw data rate of spatial multiplexing, these two MIMO techniques can be used simultaneously or alternately, based on the channels conditions. There is a fundamental trade-off between diversity and multiplexing: One cannot have full diversity gain and also attempt spatial multiplexing. Essentially, the choice comes down to the following question: Would we rather have a lower spectral efficiency but good BER performance or high spectral efficiency with low BER performance ? [12, 13, 14].

## III. THE SYSTEM MODEL

We consider a MIMO system with $N$ transmit antennas and $M$ receive antennas. The block diagram of the proposed system is shown in Fig. 2. The system consists of a transmitter with a switch between a SM and a diversity mode, a receiver unit with the corresponding pair of receivers, and MIMO mode selector, and a feedback path. At the receiver side, the MIMO mode are selected according to the current channels conditions. The information about the MIMO mode is sent to the transmitter through the feedback path. The transmitter then switches the MIMO mode switching based on the feedback information.

Using equation (7), the estimated received signal based on SM, using ZF detector can be calculated as following,

$$\begin{bmatrix} \hat{x}_1 \\ \vdots \\ \hat{x}_M \end{bmatrix} = \begin{bmatrix} x_1 \\ \vdots \\ x_M \end{bmatrix} + (H^*H)^{-1}H^* \begin{bmatrix} n_1 \\ \vdots \\ n_M \end{bmatrix} \quad (9)$$

Where $\hat{x}_1$, $\hat{x}_1$, ...,and $\hat{x}_M$ represents the estimated received data in the $i^{th}$ spatial subchannel), and let us assume that,

$$(H^*H)^{-1}H^* = \begin{bmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{M1} & \cdots & a_{MN} \end{bmatrix} \quad (10)$$

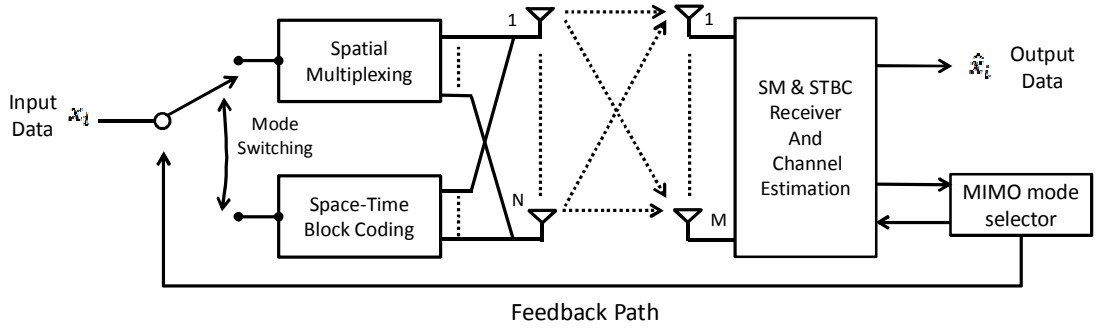



**Fig. 2: Block diagram of MIMO mode dynamic switching scheme.**

Then,

$$\hat{x}_i = x_i + a_{i1}n_1 + a_{i2}n_2 + \cdots + a_{iN}n_i, \; i = 1,2, \ldots M \quad (11)$$

Now let us define a new variable,

$$\gamma_i, i = 1,2, \ldots M$$

Where $\gamma_i$, represents the instantaneous received signal-to-noise ratio per bit per spatial subchannel, $i$.

For QPSK signal, we have

$$\gamma_i = 1/(\rho_i \sigma_n^2), i = 1,2, \ldots M \quad (12)$$

Where,

$$\rho_i = f(h_{i1}, h_{i2}, \ldots h_{iN}) = |a_{i1}|^2 + |a_{i2}|^2 + \cdots + |a_{iN}|^2 \quad (13)$$

And , $i = 1,2, \ldots M$.

The values of $\rho_i$, represents the instantaneous channels conditions and play the main role in defining the dynamic mode selector.

The MIMO mode dynamic switching algorithm working as following

1- using the channel estimation values, we calculate the instantaneous channels conditions, $\rho_i, i = 1,2, \ldots M$.

2- Define a threshold value $\rho$, where $\rho$ controls both the spectral efficiency and BER performance. If $\rho = 0$, then only diversity mode is considered; if $\rho = \infty$, then only SM is considered. Usually $0 < \rho < \infty$.

3- Now for any value of $\rho_i > \rho$, we consider that spatial path, $i$ is bad. Otherwise the spatial path considered to be good.

4- The receiver sends an $M$ bits to the transmitter via a feedback path. The bit "0" means bad spatial path $i$, and the bit "1" means good spatial path $i$, where $i = 1,2, \ldots M$.

5- If the number of 1's is greater than $M/2$, then the SM algorithm will be implemented through the mode switching, and the data streams corresponding to the bad spatial paths should not be sent.

6- If the number of 1's is less than or equal to $M/2$, then the orthogonal space-time block codes (OSTBC) diversity algorithm will be implemented.

## IV. Simulation Results And Observations

The modulation scheme used is Quadrature Phase Shift Keying (QPSK). We have performed these simulations using MATLAB. In this model we assume $N = 2$, and $M = 2$ (see Figs. 1, and 2). where the complex values $h_{11}, h_{12}, h_{21}, h_{22}$ are modeled as Rayleigh random variables. Assuming that the channels are constant over at least two adjacent OFDM symbols [16 ].

Both the spectral efficiency and the BER were used to analyze the system performance. The BER is an important parameter which is used to analyze the transmission impairments like noise, jitter and interference in wireless communication systems. The bit error rate is the probability that any given bit of the received data will be in error. A bit error rate of $10^{-6}$ means that one bit in $10^6$ will be in error. The simulation results obtained by plotting the BER against the Signal to Noise Ratio (SNR) [15].

### *A. Spectral Efficiency*

Spectral efficiency measures how much the available bandwidth is optimized for the maximum transmission of data. For spectral efficiency, we have analyzed the relationship between the threshold value $\rho$, versus spectral efficiency. This can be shown in Fig. 3.

Theoretically, the minimum spectral efficiency, 100% occurs when $\rho = 0$. The maximum spectral efficiency, $(100 \times M)\%$ occurs when $\rho = \infty$.

We can see from Fig. 3, that spectral efficiency increases by increasing the threshold value, $\rho$. The spectral efficiency increased rapidly for values of $0 < \rho \leq 15$, and the spectral efficiency increased very slowly when $\rho > 5$. We can get 190% spectral efficiency for the value of $\rho = 15$. It means that by increasing the threshold value, of $\rho > 15$, we can only benefit 10% of spectral efficiency.

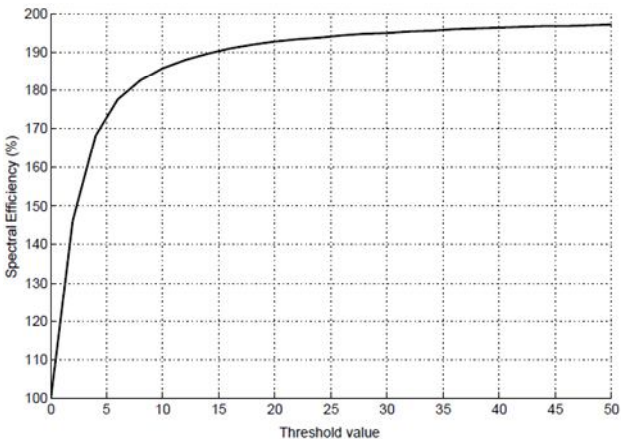



**Fig. 3: The spectral efficiency versus threshold value $\rho$.**

### *B. BER Performance Evaluation*

For the BER performance comparison purposes, we have subdivided our simulation to three scenarios (SC-1 to SC-3).

1- SISO through Gaussian Noise channel Model (SC-1)

In this scenario model, we assume $N = 1$, and $M = 1$ (see Fig. 4). Only Gaussian noise is considered in this scenario, *i.e.*, $h_{11} = 1$.

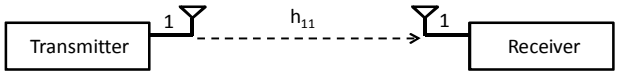



**Fig. 4: The SISO Gaussian channel model.**

The estimated received signal can be determined as following:

$$\hat{x}_1 = x_1 + n_1 \quad (14)$$

Where $n_1$ is a Gaussian noise.

2- SISO Flat fading through Gaussian noise channel Model (SC-2)

We use the same Fig. 4, where the complex value $h_{11}$ is modeled as Rayleigh random variable.
The estimated received signal can be calculated as following [11]:

$$\hat{x}_1 = x_1 . |h_{11}|^2 + (h_{11})^* n_1 \quad (15)$$

3- Switching Between Diversity and Multiplexing through Flat Fading Channel Model (SC-3)

In diversity mode, we have used Space Time Block Code (STBC) ( Alamouti's Code) ) [17], the estimated received signals can be determined as following:

$$\hat{x}_1 = (1/\sqrt{2}).x_1.(|h_{11}|^2+|h_{12}|^2 + |h_{21}|^2+|h_{22}|^2) + (h_{11})^*n_1 + (h_{12})(n_3)^* + (h_{21})^*n_2 + (h_{22})(n_4)^* \quad (16)$$

$$\hat{x}_2 = (1/\sqrt{2}).x_2.(|h_{11}|^2+|h_{12}|^2+|h_{21}|^2+|h_{22}|^2) + (h_{12})^*n_1 - (h_{11})(n_3)^* + (h_{22})^*n_2 - (h_{21})(n_4)^* \quad (17)$$

Where $n_1$ and $n_2$ are Gaussian noise.

For the comparison purposes, we run the simulation program for three threshold values; the first value ($\rho = 1$ ) reflects a little spectral efficiency effect, the second value ($\rho = 5$) reflects moderate spectral efficiency effect, and finally the third value ($\rho = 20$) reflects a good spectral efficiency effect. Figures 5,6, and 7 show the simulation results for the different threshold values, for different spectral efficiency effects.

It can be shown from the three figures that the BER performance degrades when threshold value ($\rho$) increases, but by increasing the threshold value, the spectral efficiency is enhanced. This is a tradeoff between having a good spectral efficiency or good performance of BER.

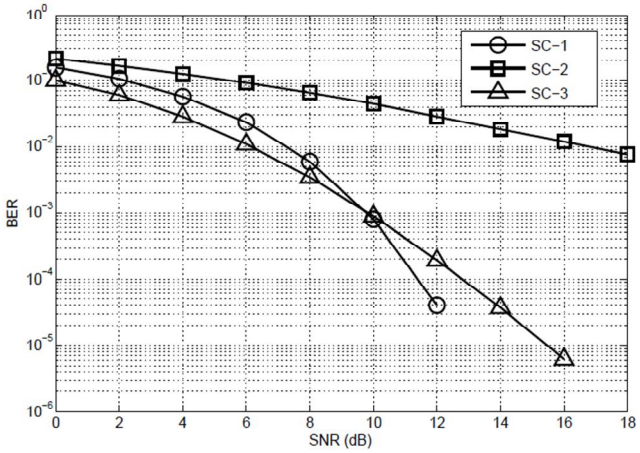



**Fig. 5: The simulation results for the threshold value $\rho = 1$.**

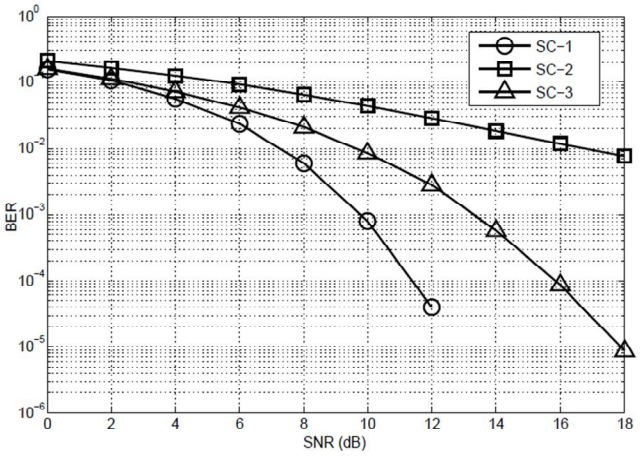



**Fig. 6: The simulation results for the threshold value $\rho = 5$.**

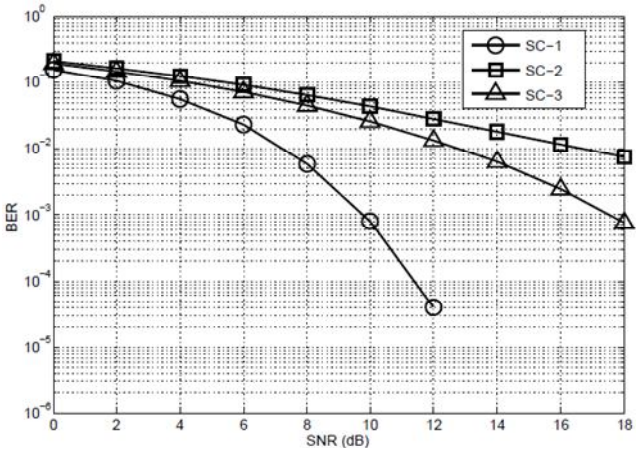



**Fig. 7: The simulation results for the threshold value $\rho = 20$.**

It can also be noticed that the performance of the proposed system is still better than the performance of a single path flat fading channel (SISO system), even for a good spectral efficiency ($\rho = 20$).

## V. Conclusion

In this paper, we propose a dynamic MIMO mode switching scheme between spatial multiplexing and diversity. The proposed scheme switches dynamically based on the instantaneous channels conditions. The spectral efficiency and Signal to Noise Ratio (SNR) versus Bit Error Rate (BER) were evaluated under QPSK modulation scheme for OFDM-MIMO system over Rayleigh fading channel. There is a tradeoff between having a good spectral efficiency and high BER performance. Increasing the spectral efficiency will reduce the BER performance, but the results show that our scheme outperforms a single path flat fading channel.

## VI. References


[1] V. Nayak, G. Bhangle , and M Raje, "Diversity Performance of MIMO-OFDM systems", *International Journal of Engineering Research & Technology* (IJERT), Vol. 1 Issue 8, October – 2012.

[2] Changick Song, Kyoung-Jae Lee, Inkyu Lee, "Performance Analysis of MMSE-Based Amplify and Forward Spatial Multiplexing MIMO Relaying Systems", *IEEE Transaction on Communications*, vol. 10, pp. 2445 - 2450, July 2011.

[3] Prof. Lajos Hanzo, Dr. Yosef (Jos) Akhtman, Dr. Li Wang and Dr. Ming Jiang, "*MIMO-OFDM for LTE, Wi-Fi and WiMAX*", John Wiley & Sons,UK,2011.

[4] J. Andrews, "Fundamentals of WiMAX Understanding Broadband Wireless Networking", PRENTICE HALL, 2007.

[5] S. Sesia, I. Toufik, and M. Baker, "LTE – The UMTS Long Term Evolution: From Theory to Practice", John Wiley & Sons Ltd, 2011.

[6] L. Korowajczuk, "LTE, WiMAX and WLAN Network Design, Optimization and Performance Analysis", John Wiley & Sons, Ltd, 2011.

[7] G. Foschini and M. Gans, "On limits of wireless communications in a fading environment when using multiple antennas", *Wireless Personal Communications*, 6:311–335, March 1998.

[8] E. Teletar, "Capacity of multi-antenna gaussian channels", *European Transactions Telecommunications*, 6:585–595, November–December 1999.

[9] J. G. Proakis, "*Digital Communications*", 3rd ed., McGraw-Hill, 1995.

[10] S. Verdu, "*Multiuser Detection*. Cambridge University Press", 1998.

[11] M. Ergen, "Mobile Broadband Including WiMAX and LTE", Springer Science+Business Media, LLC 2009.

[12] R. W. Heath and A. J. Paulraj, "Switching between multiplexing and diversity based on constellation distance", In *Proceedings, Allerton Conference on Communications, Control, and Computing*, September 2000.

[13] L. Zheng and D. Tse, "Diversity and multiplexing: A fundamental trade-off in multiple antenna channels", *IEEE Transactions on Information Theory*, 49(5), May 2003.

[14] C. Kim, and J. Lee, "Dynamic rate-adaptive MIMO mode switching between spatial multiplexing and diversity", EURASIP Journal on Wireless Communication and Networking, July 2012, 2012:238.

[15] L. J. Cimini, "Analysis and simulation of a digital mobile channel using orthogonal frequency division multiplexing", *IEEE Transaction on Communications*, Vol. 33, Issue 7, pp. 665–675, July 1985.

[16] F. Khan, "LTE for 4G Mobile Broadband Air interface Technologies and Performance", Cambridge University Press 2009.

[17] S. M. Alamouti, "A simple transmit diversity technique for wireless communications", *IEEE Journal on Selected Areas in Communications*, vol.16, pp.1451-1458, Oct.

**Jamal Elbergali** received his B.Sc. degree in 1990 from Tripoli University, Tripoli, Libya and M.Sc in 1993 from Aligarh Muslim University, India. His Ph.D was received in 2003 from the University of Cape Town (UCT). Currently he is a staff member at the College of Industrial Technology. His research interests on the Evolved Packet Systems (EPS), and Long Term Evolution (LTE advanced).

# Performance Evaluation of RSSI based CCA-Map Localisation Algorithm in Wireless Sensor Networks

Omotayo G. Adewumi*, Karim Djouani and Anish M. Kurien
F'SATI / Department of Electrical Engineering
Tshwane University of Technology, Private Bag X680, Pretoria 0001, South Africa
Tel: +27 12 3825911, Fax: +27 12 3825114
email: {adewumiog, djouanik, kurienam}@tut.ac.za

**Abstract-Wireless Sensor Networks are currently a very active area of research and are broadly being used in diverse smart environments to accomplish numerous monitoring tasks such as search, rescue, disaster relief and target tracking. In order to accomplish these monitoring tasks, wireless sensor nodes are required to know the information about their geographical locations. So therefore in this paper, the performance of RSSI based localisation algorithm to accurately estimate the positions of wireless sensor nodes is evaluated. This approach is based on the Curvilinear Component Analysis Mapping (CCA-MAP) algorithm which applies an efficient non-linear data mapping techniques for position estimation. The results obtained from the RSSI based localisation algorithm shows that it is able to provide improved position accuracy and computational efficiency.**

**Index Terms—WSN, RSSI, CCA-MAP, Localisation.**

## I. INTRODUCTION

Recent advances in wireless communications and electronics has enabled the development of low cost sensor networks and it has attracted a lot of interest in recent times. A wireless sensor network is a network that consists of a large number of wireless radio nodes that are equipped with sensing devices and are densely distributed for specific applications. Each sensor node is equipped with advanced sensing functionalities (thermal, pressure, acoustic, and so on), a small processor, battery supply and a short-range wireless transceiver. An example of a typical wireless sensor node is shown in figure 1 below. Wireless sensor nodes exchange information in the environment in order to build a global view of the monitored region in which is made accessible to the external user through a gateway node [1].

Figure 1: Figure taken from [2], *XM2110 Crossbow Iris* Mote with Standard Antenna.

A number of issues arise when WSN is designed. All the nodes in the network must have the ability to communicate with one another and send data to a central station. Due to the significance of some routing protocols and data sampling in WSN, all the sensor nodes must have the ability to know what the current time is [3], [4].

Wireless sensor nodes are usually deployed in different topologies. Each of these deployed sensor nodes has the ability to collect data and route data back to the sink or base station and to end users. Data is routed back to the end user by a multi-hop infrastructure less architecture through the sink. The sink may communicate with the task manager node through the Internet or through Satellite communications. This is illustrated in Figure 2.

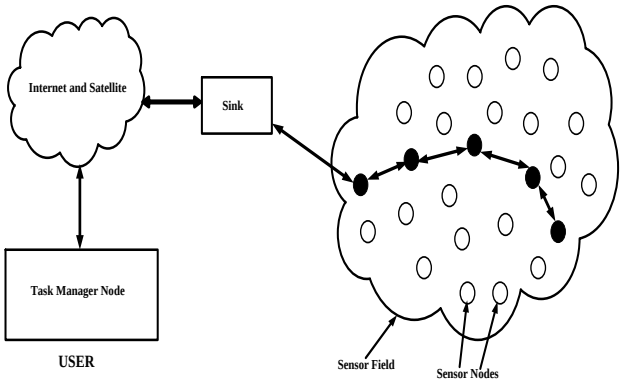



Figure 2: Figure taken from [1], Wireless Sensor Nodes Scattered in a Sensor Field.

Some of the applications of WSN include disaster and relief operations, biodiversity mapping for wildlife observation, intelligent building and bridges, military operations, health applications where motes may be deployed to collect vital information such as pulse and heart rate. Some WSN applications have been applied in underground environments such as earthquake monitoring and mining applications, soccer fields, locating people in collapsed buildings, e.t.c. In addition to the applications already listed, WSN have also been used for under-water applications and have been implemented for ocean sampling networks, disaster prevention, assisted navigation, pollution monitoring specifically for chemical and biological spillage [1].

In many WSN applications, sensed information only becomes useful when it is accompanied by the location of the area where such information is sensed, and hence, sensor nodes need to know their location. This need drives the assumption that the location information of sensor nodes is available by some means such as Global Positioning System (GPS) or by manual entry. The challenges faced in entering the position information of nodes manually is that it limits the size and scalability of a sensor network; as a result, it reduces the strengths of WSNs.

The problem with assuming GPS enabled sensor nodes is that the cost of the sensor nodes increases considerably, and the appropriate environment in which the sensor nodes can be deployed is limited. As a result of the constraints of power consumption as well as the limitations that could be caused due to the lack of visibility of satellites require the need to propose an alternative solution. In this case, methods for self-determination of position information are required for achievable placement of wireless sensor nodes [5]. In this paper, we evaluated the performance of CCA-Map localization algorithm using the Received Signal Strength Indicator (RSSI) technique which is suitable for indoor and outdoor propagation modeling to determine the location of individual sensor nodes in the network.

The remainder of this paper is organized as follows. In section 2, we present the related work on localisation algorithm and section 3 emphasizes on RSSI based ranging model which is the model used as an input to our proposed localisation algorithm. Section 4 focuses on the proposed approach which is the localisation algorithm to find the positions of the nodes in WSN. In section 5, the simulation results of the proposed approach are presented. Finally, the conclusion and future work of this research study are summarized in section 6.

## II. RELATED WORK ON LOCALISATION ALGORITHMS IN WSN

Many localisation algorithms have been proposed in the literature to calculate relative positions among the nodes of a network. Triangulation and Optimisation based methods proposed by [6], [7] are examples of work done on localisation and utilises a neighbouring distance measurement method for local estimation of a node's position and makes use of iterative steps to localise all nodes. The Distance Vector Hopping (DV-Hop) algorithm proposed by [8] is a hop count based localisation technique where a limited fraction of nodes has a self-positioning capability; but however, the localisation accuracy of this algorithm is poor.

The Monte Carlo Localisation (MCL) proposed by [9] is designed to obtain the probabilistic distribution of the node's possible location by using Sequential Monte Carlo (SMC) and uses mobility of nodes to attain better position accuracy; this algorithm suffers from large latency because of the computation intensive sampling methods used and the high seed or anchors density employed. Approximate Point in Triangulation (APIT) proposed in [10] uses the point in triangulation technique; however, a high anchor density and a long range of anchors are required. The Multidimensional Scaling (MDS) approach proposed by [11] relies on central processing node that collates and performs computation of location assignments; but it has a problem of scalability. Cooperative localisation techniques based on MDS was also proposed by [12] and concerns distributed approach for node location estimation either in range based or range free conditions with minimal number of anchors required; this algorithm however has the problem of time complexity.

The major challenges in all these methods lie in designing an effective and robust localisation scheme that takes cognisance of the form factor and computing power of the node, power usage at a minimum level, scalability of the network, and of course, the accuracy of the localisation technique. Based on the limitations of the localisation algorithm that was discussed in section 2, the need to find an alternative approach to eradicate some of these limitations is identified. As a result, in section three, an RSSI based ranging model is described and in section 4 an improved position accuracy and distance based cooperative localisation algorithm is presented.

## III. RSSI BASED RANGING MODEL

Received Signal Strength Indicator (RSSI) is one of the simplest methods of distance measurement and can be easily implemented in real systems. This technique is based on a standard feature found in most wireless devices, a received signal strength indicator (RSSI). RSSI is defined as the amount of power present in a received radio signal. Due to radio-propagation path-loss, RSSI decreases as the distance of the radio propagation increases. Therefore, the distance between two sensor nodes can be compared using the receive signal strength value at the receiver assuming that the transmission power in the sender is either fixed or known [13].

In distance prediction based methods, raw RSSI data are directly mapped to distances through a signal propagation model. In some cases, distances are not explicitly measured, but are taken into consideration through assumptions on the underlying distributions. In the case of outdoor deployments, where there are no obstacles, a single signal propagation model is all what the profiling data is required [14], [15]. In the case of indoor deployments the same applies. However, the signal propagation model might be different across different buildings with different furniture and wall arrangements and therefore, when an electromagnetic signal propagates, it may be diffracted, reflected and scattered. These effects have two important consequences on the signal strength. First, the signal strength decays exponentially with respect to distance. Second, for a given distance *d*, the signal strength is random and log-normally distributed about the mean distance-dependent value [15].

Given an underlying signal propagation model, raw RSSI data can be mapped to specific distances. These distances can be either used directly in order to perform localisation or assigned a probability and used in a learning based/probabilistic localisation algorithm. Obviously, the accuracy of the distance prediction based methods depends heavily on the accuracy of the signal propagation model used for translating RSSI values to distances. One of the most common radio propagation models is the log-normal shadowing model proposed by [16] and is given in Equation 1.

The Log-Normal Shadowing model is a radio propagation model that predicts received signal strength inside a building or densely populated areas over distance. The model is applicable to indoor and outdoor propagation modelling [17].

$$PL(d) = PL(d_0) + 10n\log_{10}(\frac{d}{d_0}) + X_\sigma \quad (1)$$

where $d$ is the transmitter-receiver distance, $n$ is the attenuation constant (rate at which signal decays), $X_\sigma$ is a zero-mean Gaussian (in dB) with standard deviation $\sigma$ (multi-path effects), $d_0$ is a reference distance and $PL(d_0)$ is the power decay for this distance. Usually, $n$ and $\sigma$ are obtained through curve fitting of empirical data. The received signal strength ($P_r$) at a distance $d$ is the output power of the transmitter ($P_t$) minus $PL(d)$, that is $P_r = P_t - PL(d)$, (all powers in dB) [17], [18].

Equation 2 is derived from Equation 1 which is the simplified and commonly used model for calculating a distance $d$ from a given RSSI value using a measured RSSI value at distance of 1 meter ($A$) and the attenuation constant $n$.

$$d = 10^{\frac{(A-RSSI)}{10n}} \quad (2)$$

## IV. RSSI BASED CCA-MAP ALGORITHM

The approach in this paper is an RSSI - distance based cooperative localization scheme called Curvilinear Component Analysis Mapping (CCA-MAP). CCA-MAP algorithm was chosen among other localisation algorithms because it has high accuracy with low number of anchor nodes and without any additional refinement process. CCA-MAP algorithm is also computationally efficient resulting in faster mapping process [19].

CCA-MAP technique is a cooperative node localization scheme that applies an efficient non-linear projection method [20], to deliver accurate data dimension reduction and to localise nodes in a WSN using distance measurement. CCA delivers accurate data dimension reduction while preserving distances between the data points during the reduction process at a computational cost that is the least among the various reduction methods. CCA looks for configuration of points in the output space that preserves the original distances as much as possible while focussing on small distances in the output space. This cooperative localisation scheme formulates the localisation problem as a joint estimation problem and applies optimisation techniques to derive location coordinates considering all constraints on inter-node distances, rather than considering only constraints between the sensor nodes and anchor nodes. It uses a variant of the stochastic gradient descent method to create a mapping of data.

The goal of the CCA is to minimise a cost function shown in equation 3 based on inter-point distances between the original input space and projection output space [19].

$$E = \frac{1}{2}\sum_i \sum_{j\neq i}(A_{ij}-B_{ij})^2 F(B_{ij},\lambda_b) \quad (3)$$

$A$ represents the distance matrix in the input space and $B$ represents the distance matrix in the output space. $A_{ij}$ stands for inter-point distance that forms a $N\times N$ distance matrix of $A_{ij} = d(a_i a_j) = \sqrt{\sum_{k=1.....n}(a_{ik}-a_{jk})^2}$ in the input space and $B_{ij}$ also stands for inter-point distance that forms a $N\times N$ distance matrix of $B_{ij} = d(b_i b_j) = \sqrt{\sum_{k=1.....n}(b_{ik}-b_{jk})^2}$ in the output space. The weighting function is often bounded and monotonically decreasing in time with each computing cycle in order to favour the local topology conservation. CCA uses a modified stochastic gradient descent method to improve computation efficiency in its update cycle such that it pins one $b_i$ and moves all other $b_j$ around. This means that only the $b_i$ distance from node $i$ to other $N-1$ nodes are computed instead of all $\frac{N(N-1)}{2}$ distances in the input and output spaces [20].

The updated cycle is;

$$\Delta b_j = \alpha(t)F(B_{ij},\lambda_y)(A_{ij}-B_{ij}).\frac{b_j-b_i}{B_{ij}} \quad \forall j \neq I \quad (4)$$

The decrease exponential function, $F(B_{ij},\lambda) = e^{-\frac{B_{ij}}{\lambda(t)}}$ is selected with the fact that both $\alpha(t)$ and $\lambda(t)$ decreases with time along each computing cycle in order to conserve the local topology. The complexity in terms of adaptation cycle of all nodes is $O(N)$ instead of $O(N^2)$ as in most Non-Linear Mapping (NLM) algorithms. Contrast to most NLM technique, CCA minimisation allows the cost $E$ to increase temporarily, but bind it to decrease on average. The computation applying CCA minimisation not only converges much faster, but it also escapes from local minima to reach a much deeper minimum.

The projection of node coordinates formally starts by considering the following problem. If we have a distance matrix $D_{(N\times N)}$ of $N$ nodes, the coordinates of all nodes are sought to obtain;

$$\min \sum (d_{ij}-p_{ij})^2 \text{ for } i,j=1, 2, \ldots, N \quad (5)$$

In equation 5 above, $d_{ij}$ is the measured or known distance between node $i$ and $j$, while $p_{ij}$ represent the distance between nodes $i$ and $j$ computed using the calculated position coordinates of $i$ and $j$. Taking $d_{ij}$ to be the distance matrix of the input data set and $p_{ij}$ to be the distance matrix of the output set, CCA will force equation 5 to a minimum in order to maximize the cost function in equation 3.

### A. Steps To Project Node's Position using CCA

There are two simple steps involved in projecting the node's coordinates provided that the node distance matrix is given. The distance matrix is calculated using RSSI which is one of the efficient range based method to estimate distance in WSN.

- The initial output estimation of $b_{(N\times 2)}$ is set. This is done by using the mean values of the first two columns of the input data set $a_{(N\times N)}$. It is then adjusted by a Gaussian distribution with a standard deviation of the same column.
- In each cycle, select node $i$ and compute for each node $j(j\neq i)$ the new $b_j(t+1)$ from the current value of $b_j(t)$ using equation 6 below.

$$b_j(t+1) = b_j(t) + \alpha(t)\, e^{-\frac{B_{ij}}{\lambda(t)}} \left(\frac{A_{ij}}{B_{ij}} - 1\right)(b_j - b_i) \qquad (6)$$

The decrease exponential function, $F(B_{ij},\lambda) = e^{-\frac{B_{ij}}{\lambda(t)}}$ is selected based on the fact that both $\alpha(t)$ and $\lambda(t)$ decreases with time along each computing cycle in order to conserve the local topology.

### B. Implementing the MAP Algorithm for CCA

CCA-MAP builds a local map at each node within the sensor field and puts them together to form a global map. Additionally, in computing the node coordinates in the local map, CCA is also implemented to mean that every single node computes its own local map using only the information obtained locally. In a situation where accurate ranging capability (RSSI model) is implemented in the network, the distance between pairwise neighbour's nodes will be measured and known. Then, each node applies the CCA algorithm generating the relative coordinates of each node in its local map. The local maps are then merged. The merged map transforms into an absolute map based on the positions of the anchor nodes. While a minimum of three anchor nodes is required for a 2-dimensional space. In the merging step, for the starting map, the local map of a randomly selected node is used. After this, the neighbour node whose local map shares the most nodes in the current map are selected to merge its local map into the current map. Using the coordinates of their common nodes, two maps are merged. A linear transformation is used for merging a new local map into the current map. This scheme allows local maps to be merged in parallel in different parts of the network since CCA-MAP is implemented in a distributed fashion

The CCA-MAP requires no further optimisation because the results obtained are largely optimized over the given distance information which is satisfactory in this regard.

## V. SIMULATION RESULT

In this section, the performance of the proposed RSSI based localisation algorithm is evaluated in order to measure the effectiveness of the algorithm on wireless sensor networks.

The performance of the RSSI based CCA-MAP was evaluated using different randomly deployment scenarios implemented in MATLAB version 7.10 on a Pentium M processor of 2GHz with 2GB RAM. Sensor nodes were randomly deployed according to a specified connectivity level. The node that performs the map merging is randomly selected and the distance measurement between the sensor nodes were computed using a range based distance estimation technique called Received Signal Strength Indicator (RSSI) model. RSSI model used the Lognormal Shadowing model (LNSM) to compute the range [21]. This model is applicable for both indoor and outdoor propagation modelling.

### A. Simulation Experiment

In this research work, we considered a network topology in which the coordinates of each node in the topology are known and the distance matrix between each node is also known from the RSSI measurement which was taken from the research work in [21]. Only part of the distance matrix is used for the simulation input with three nodes selected as anchors. The localisation goal is to determine the coordinates of all the nodes in the network.

### B. Localisation Accuracy

To evaluate the performance of the proposed algorithm, the mean error was used between the estimated and the true position of the non-anchor nodes in the network.

$$\text{Mean Error} = \frac{1}{N-M} \frac{\sum_{i=M+1}^{N} (\|\hat{x}_i - x_i\|)}{r}, \qquad (7)$$

Where $N$ and $M$ are the total number of sensor nodes and the number of anchor nodes respectively. $x_i$ denotes the true position of sensor $i$ in the network, $\hat{x}_i$ is the estimated location of sensor $i$ and $r$ is the radio transmission range.

Obtaining low error would be indication of good performance.

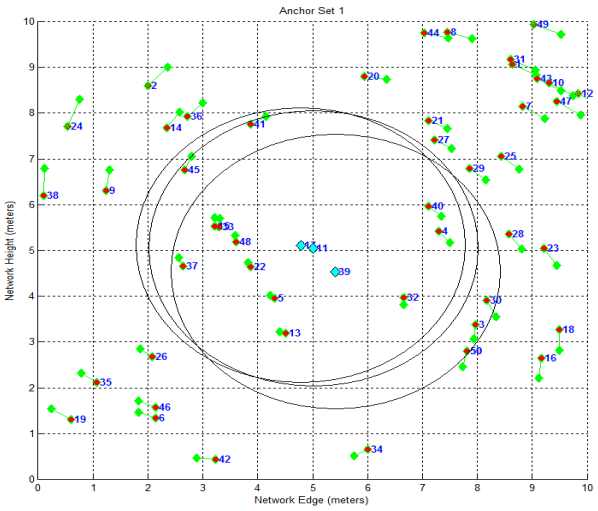



Figure 3: Plot of the network difference between the real position and the estimated positions of 50 nodes.

The simulation results shown in figure 3 is the plot of the network difference between the real positions of the 50 sensor nodes randomly deployed in the network and the estimated positions which are also the mapped points generated after applying the RSSI based CCA-MAP localisation algorithm.

The green shaded diamonds are the real positions of the sensor nodes, the red shaded circles are the estimated positions of the sensor nodes while the cyan shaded diamonds are the anchor node positions. The black big circles represent the network coverage of each anchor node in the network. The mean error in the network is 0.153.

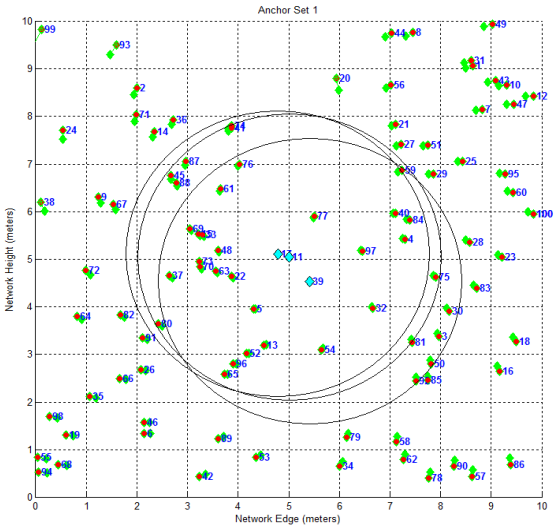



Figure 4: Plot of the network difference between the real position and the estimated positions of 100 nodes.

We further evaluated the performance of the proposed algorithm by increasing the density of the sensor node. The total number of sensor nodes was increased to 100 and the simulation result is shown in figure 4 above which shows the network difference between the real points and the mapped points of 100 randomly deployed sensor nodes. The mean error in the network is 0.042 and it is observed that the localisation error of the network containing 100 sensor nodes is lower than the network of 50 sensor nodes.

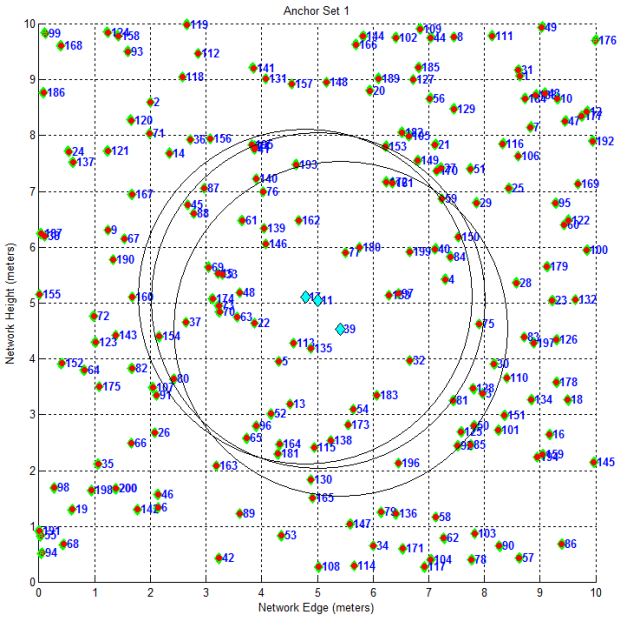



Figure 5: Plot of the network difference between the real position and the estimated positions of 200 nodes.

Figure 5 above shows the network difference between the real points and the mapped points of 200 randomly deployed sensor nodes. The mean error of the network is 0.004. It can also be observed that the localisation error of network containing 200 sensor nodes is lower than the network of 100 sensor nodes and 50 sensor nodes. It was observed that as the density of the node is increased (scalability), the localisation error decreases. This shows that the proposed distance based algorithm accurately localises the positions of the sensor nodes.

## VI. CONCLUSION AND FUTURE WORK

In this paper, a comprehensive performance of the CCA-Map localisation algorithm which is based on RSSI measurements was evaluated and from the simulation results obtained, it was observed it achieved a desirable result and improved position accuracy compared to other localisation algorithms that was discussed in the section 2.

Currently, we are working on implementing the CCA-Map localisation algorithm based on RSSI measurements on a real test-bed to localise static and mobile sensor nodes in WSN.

## ACKNOWLEDGEMENTS

"This work is based on the research supported in part by the National Research Foundation of South Africa (Grant reference number (UID) 80050)".

## VII. REFERENCES


[1] I. F. Akyildiz, W. Su, Y. Sankarasubramamiam, and E. Cayirci, "Wireless Sensor Networks: A Survey". Proceedings of IEEE Communications Magazine, August 2002, Vol. 40(8):104-112.

[2] CROSSBOW. 2007, "MPR-MIB User's Manual: Revision A". 2007.

[3] Kiran, M, Kamal, K. and Nitin, G, "Application Based Study on Wireless Sensor Networks". International Journal of Computer Application, May, 2011. Vol. 21(8):0975-8887.

[4] Danielle, P., Martin, H, "Wireless Sensor Networks: Applications and Challenges of Ubiquitous Sensing". In: Journal of IEEE Circuits and Systems Magazine, 2005. Vol. 5(3):19-31.

[5] A. S. Roger, "Clustering Based Localization For Wireless Sensor Networks," A Thesis Submitted in Partial Fulfilment of Master of Science, School of Electrical Engineering and Computer Science, USA. May 2006.

[6] Michael, C., Holly, J., Michael, S. and Yinyu, Y, "Spaceloc: An Adaptive Sub-Problem Algorithm for Scalable Wireless Sensor Network Localization". In: Society for Industrial and Applied Mathematics, 2006. Vol. 17(4):1102-1128.

[7] Doherty, L., Ghaoui, L. and Pister, K, "Convex Position Estimation in Wireless Sensor Networks". In: Proceedings of 20th Annual Joint Conference of the IEEE Computer and Communications Societies, Alaska, USA, April 2001, 1655–1663.

[8] Niculescu, D., Nath, B, "Dv Based Positioning In Ad Hoc Networks". In: The Kluwer Journal Of Telecommunication Systems, 2003. Vol. 22(1-4):267-280

[9] Hu, L., Evans, D. 2004, "Localization For Mobile Sensor Networks". In: Proceedings Of The Mobicom'04 Of The 10th Annual International Conference On Mobile Computing And Networking, New York, Usa. 2004, 45-57.

[10] Tian, H., Chengdu, H., Brian, M. B., John, A. S. and Tarek, A, "Range-Free Localization Schemes For Large Scale Sensor Networks". In: Proceedings Of The Mobicom'03, International Conference On Mobile Computing And Networking, San Diego, Ca, Usa, September, 2003, 81-95.

[11] Ji, X., Zha, H, "Sensor Positioning In Wireless Ad Hoc Sensor Networks Using Multidimensional Scaling". In: Proceedings Of Ieee International Conference On Computer Communications (Infocom'04), Hong Kong, China, March, 2004, 2652-2661.

[12] Shang, Y., Rumi, W, "Improved Mds-Based Localization". In: Proceedings Of IEEE International Conference On Computer Communications (Infocom), March 2004, 2640-2651.

[13] Stojmenovic, I, "Handbook Of Sensor Networks: Algorithms And Architecture". Isbn: 978-0-471-68472-5, Ed.: Wiley Inter-Science, 2005.

[14] Sharly, J. H., Joon-Goo, P. and Wooju, K, "Adaptive Filtering For Indoor Localization Using Zigbee RSSI And LQI Measurement, Adaptive Filtering Applications". Dr Lino Garcia (Ed.), Isbn: 978-953-307-306-4, In-Tech.

[15] Shashrank, T, "Indoor Local Positioning System For Zigbee, Based On RSSI". Master Of Science Thesis Report, Department Of Information Technology And Media (Itm), Mid Sweden University, October, 2006.

[16] Rappaport, T. S, "Wireless Communications Principles And Practice". Prentice Hall, Upper Saddle River, Nj, 1996.

[17] Jiuqiang, X., Wei, L., Fenggao, L., Yuanyuan, Z. and Chenglong, W, "Distance Measurement Model Based On RSSI in WSN". In: Wireless Sensor Network, June, 2010. Vol. 2(8):606-611.

[18] Zuniga, M., Krishnamachari, B, "Link Layer Models For Wireless Sensor Networks". Usc, 1st Edition, December, 2005.

[19] L. Li, And T. Kunz, "Cooperative Node Localization Using Non-Linear Data Projection," In Acm Transactions On Sensor Networks, February 2009. Vol. 5(1): 1-26.

[20] P. Demartines, And J. Herault, "Curvilinear Component Analysis: A Self-Organizing Neural Network For Nonlinear Mapping Of Data Sets". In IEEE Transactions On Neural Networks, January, 1997. Vol. 8(1):148-154.

[21] Omotayo G. Adewumi, Karim Djouani and Anish M. Kurien, "RSSI based Indoor and Outdoor Distance Estimation for Localization in WSN". In proceedings of the IEEE International Conference on Industrial Technology {ICIT), Capetown, South Africa. February 2013, 1534 -1539.



**Adewumi Omotayo** received his undergraduate degree in 2010 from the Tshwane University of Technology and has already completed his Master of Technology degree at the same institution. He is currently enrolling for his Doctor of Technology in the same institution. His research interests include Localisation in WSNs.

# An investigation into the accuracy of the kriging method for multiple Wi-Fi access point RSSI estimation

PJ Joubert
School of Electrical, Electronic and Computer Engineering
North-West University
Email: 21570434@nwu.ac.za

ASJ Helberg
School of Electrical, Electronic and Computer Engineering
North-West University
Email: albert.helberg@nwu.ac.za

***Abstract*—Kriging was originally developed for geographical interpolation purposes, but has proven to be a powerful tool in many other applications as well. In this study we evaluate the accuracy of kriging for Wi-Fi signal strength estimation in a complex indoor environment using multiple Wi-Fi Access Points. The empirical investigation of several scenarios is described and the results analysed. Kriging is shown to be a valid method for multiple access point Wi-Fi signal strength estimation and provides a 84% to 88% level of accuracy. Some further comments are made on the practical use of kriging for Wi-Fi signal strength estimation.**

***Index Terms*—Interpolation, Kriging, Signal Strength Estimation.**

## I. Introduction

The high demand for mobile networking and the many applications such as coverage analysis, localization, fast hand-off, and security auditability has led to increased attention to signal strength estimation for indoor wireless communications. Signal Strength estimation is needed for generating radio signal maps for planning wireless networks and in addition it can reduce the cost of time consuming site surveys and can be used to determine the network coverage where measurements can not be taken [5].

A variety of approaches can be followed to estimate signal strength in an indoor environment and can be classified as either empirical model based or deterministic model based signal strength estimations. Empirical models are known to have low accuracy compared to deterministic models and are usually replaced with deterministic models. Deterministic models however can become very complex especially in an indoor environment where input parameters are difficult to obtain and not necessarily reliable [6].

## II. Problem statement

A typical modern office environment can be described as a complex indoor environment containing many different wireless devices causing interference. Reflections caused by different kinds of walls with different attenuation factors are unavoidable. A number of different wireless Access Points may be used to connect to an extensive network that can be overlapped by other networks sharing the same frequency bands.

In an ideal environment a floor plan indicating the position of each access point and characteristics of the walls can provide the information needed to set up a simulation model to predict signal strength coverage for a building. However, in a real life environment it will be impractical to take every single influence into account to set up a deterministic model for estimating signal strength throughout a building. If such a model could be constructed, it will be very processing intensive as well. In addition, a floor plan usually consists of a 2D representation of a floor in a building, so adding a third dimension for different stories of a building only adds to the complexity.

A simple way to take every factor that has an influence on the signal strength into account is to physically measure the signal strength in the area of concern. It is also impractical to measure the signal strength at every point in a building.

In this study[1] we evaluate the use of kriging for Wi-Fi received signal strength indicator (RSSI) estimation in a complex indoor environment. This idea originated from the fact that Wi-Fi has a geographical property that obeys the first law of geography: All places are related, but nearby places are more related than distant places [1]. The random spatial distribution of signal strength in an environment as explained above further suggests the use of a geostatistical model.

The aim of this study is to determine the accuracy of the kriging interpolation method for RSSI estimation in a complex indoor environment taking an empirical approach.

Although the term signal strength is generally used, the 802.11 standard does not define signal strength as measuring RF energy in mW or dBm. The reported RF energy (RSSI) is an integer value between 0 and RSSI_Max intended for use internally by the physical and data link layers. This value can be converted to represent the user of a utility tool with an indication of the signal strength measurement presented in

[1]This work was completed at the Telkom Centre of Excellence at the North-West University.

mW or dBm.

Measuring RF energy requires expensive equipment and are subject to variations. In this study we use the reported RSSI value provided by a Wi-Fi device as an indicator of the perceived signal strength.

## III. Kriging

In 1951 a South African Mining Engineer, D.G. Krige, published a seminal paper in the journal of the Chemical, Metallurgical and Mining Society of South Africa, where he pursued a statistical exploration of the conditional biases in ore block valuations. This formed the basis of the interpolation method known today as kriging [2]. The French mathematician G. Matheron, who is known for laying the foundation of geostatistics, adapted and formalized the work of Krige in 1963. [3]. Kriging is defined as an optimal interpolation method based on regression against observed values of surrounding data points, weighted according to spatial covariance values [4].

## IV. Method

Kriging can be divided into three main types: simple kriging, ordinary kriging, and universal kriging. Cross validation of initial empirical investigations showed that universal kriging gave the best and most consistent results for the problem scenario presented here. This section will briefly explain how universal kriging is applied.

The first step in universal kriging is to record a scatter point set to be interpolated and to construct an experimental variogram from the data [7]. The experimental variogram is used to construct a model variogram which will be used to determine the weights used in kriging. A variogram is a visual representation of the variance in data as a function of the distance between samples.

The general formula from [7] used in universal kriging to interpolate a value $F$ in point $(x, y)$ is shown in (1).

$$F(x,y) = \sum_{i=1}^{n} w_i f_i \tag{1}$$

In (1), $n$ is the number of scatter points used, $w_i$ is the weight of each scatter point with each $w_i < 1$ and $\sum_{i=1}^{n} w_i = 1$ and $f_i$ is the corresponding value of the scatter point.

The weights are determined using the model variogram and solving the matrix in (2):

$$\begin{bmatrix} w_1 S(d_{11}) & w_2 S(d_{12}) & \dots & 1 & x_1 & y_1 \\ w_1 S(d_{21}) & w_2 S(d_{22}) & \dots & 1 & x_2 & y_2 \\ w_1 S(d_{31}) & w_2 S(d_{32}) & \dots & 1 & x_3 & y_3 \\ \vdots & \vdots & & \vdots & & \\ 1 & 1 & \dots & 0 & 0 & 0 \\ x_1 & x_2 & \dots & 0 & 0 & 0 \\ y_1 & y_2 & \dots & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ \lambda \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} S(d_{1p}) \\ S(d_{2p}) \\ S(d_{3p}) \\ \vdots \\ 1.0 \\ x_p \\ y_p \end{bmatrix} \tag{2}$$

In the matrix above, $S(d_{nm})$ represents the model variogram value at the distance that samples $n$ and $m$ are from each other and $S(d_{np})$ represents the model variogram value at the distance between samples and the new interpolated point. Variables $w_1$ .. $w_n$ represent the weighting of each sample when calculating the value of the new interpolated point. The constraint of the weights summing to 1.0 introduces one more equation than there are variables. In order to solve this matrix, the variable $\lambda$ is added to obtain a unique solution and is called a Lagrange multiplier, used to minimize possible estimation error. The variables $\alpha_1$ and $\alpha_2$ are the local trend coefficients of the first order trend. Including the $x$ and $y$ coordinates in the matrix is unique to universal kriging and is used when a trend is present in the data.

Solving (2) for $w_1$ .. $w_n$ enables one to calculate from (1) the specific interpolation point $f_p$ as in (3).

$$f_p = w_1 f_1 + w_2 f_2 + w_3 f_3 + \dots + w_n f_n \tag{3}$$

## V. Experimental Setup

Wi-Fi signal strength was measured at 110 random points throughout the engineering campus's main building. The building is two storeys high, approximately 30m in width, 100m in length and has open volumes shared by three floors in two areas. Each floor has two Wi-Fi Access Points and the top floor has a third access point. The RSSI at each sample point was determined by taking four measurements, about 20cm from each other and calculating the average value of the four measurements to represent the value of that sample point [8]. This was done in an attempt to minimise the effect of interference patterns that can influence the samples.

A single round of interpolations consisted of eliminating a number of points from the data set and to estimate their values by applying the kriging interpolation method to the remaining points. The interpolated points were then compared to the original measured points to determine the accuracy of kriging for this application.

In order to determine the consistency of the method for this application it first had to be tested in a number of different scenarios. The most practical approach for determining the consistency was through cross validation [9]. This was achieved by repeating the rounds a number of times for a given number of random samples taken from the set of 110 measurements. These samples were used to interpolate the rest of the points in the set of measurements. The average of the errors obtained from cross validation was then used as a reference error for that number of samples.

The process of using a random set, with a certain number of samples, to interpolate remaining points was repeated while increasing the number of random samples with each repetition. The average error when using a certain number of samples could then be compared to the average error when using fewer samples in order to find a trend. The trend could then be used to find the optimal number of samples needed to represent the given environment, after which more samples do not have a significant influence on the accuracy of the model.

A number of models are available when constructing a model variogram and it is important to choose the model

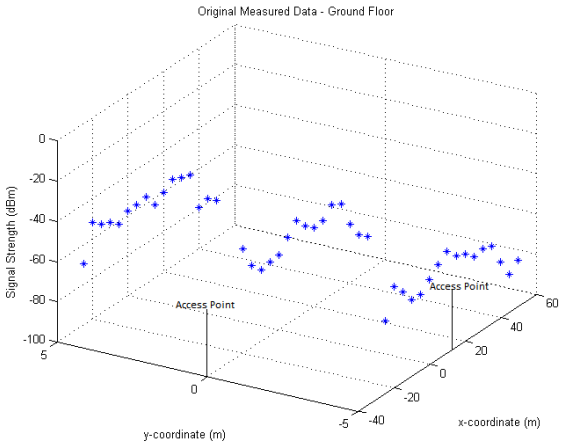



Fig. 1. Measured Wi-Fi signal strength at the ground floor

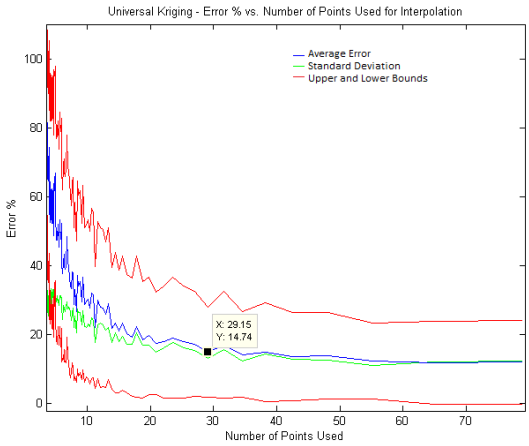



Fig. 2. Average error as sampling points increase

that is most representative of the given data [10]. The above experiment was done using the model variograms suggested by [10]. The exponential model variogram gave the best and most consistent results and was used to obtain the results presented here.

To address a single round, one must consider that in a real life scenario, a human will determine samples that will be representative of the full data set. The final step of the experiment was to determine the accuracy of the model when choosing random samples that are well distributed throughout the building instead of using an algorithm to choose random samples, that may result in a localized sample set.

## VI. Results

This section provides the results of the experiment described in the previous section. First the results of the random samples will be shown, demonstrating the consistency of the method and how accuracy improves as more samples are used, followed by the results obtained from well distributed points.

To illustrate the measurements that were taken throughout the building, only the measurements of the ground floor will be shown here. Since the Access Points on the other floors are placed in the same (x,y) positions, the distribution looks very similar to that of the ground floor. Figure 1 shows a 3D view of the distribution of the Wi-Fi signal strength on the ground floor of the building where the data was sampled.

Figure 2 shows the average error of each instance where a random set of samples is used to interpolate the remaining points in the full set of 110 points. Since the points used as samples are randomly chosen, the average error is influenced since random samples can be locally grouped and not necessarily representative of the full data set. If we assume that in practice there is some intelligence behind the sample selection process, cases where the average error is very high as a result of locally grouped samples can be considered as outliers.

In Figure 2 the blue line indicates the average error. The green line represents the standard deviation and the red lines show the upper and lower bounds of the error. The average

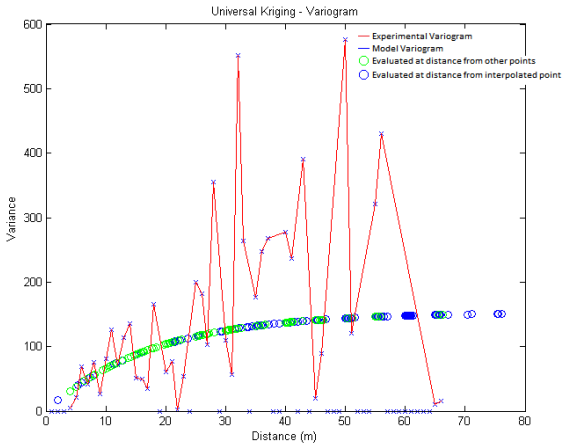



Fig. 3. Experimental variogram with exponential model variogram

accuracy of this experiment, with outliers included, increases as the number of samples increases, but only up to a certain point at around 29 samples. Using more than 29 samples in this specific environment will not have a significant improvement on the accuracy of the model.

An example of the variogram that was used in this test is shown in Figure 3. An exponential model variogram was used. The red line connects the values of the experimental variogram obtained from the data. The green circles represent the model variogram evaluated at the distance each point was from the other samples. The blue circles represent the model variogram evaluated at the distance the samples were from the new interpolated point. Since each interpolated point is in a different position, this variogram changes a little for each new point to be interpolated.

In Figure 4 the values of the originally measured points are compared to the values obtained from kriging by only using the 29 samples to estimate the remaining 81 points. The points are sorted according to their measured RSSI in ascending order. Figure 5 shows a histogram of the error that was made

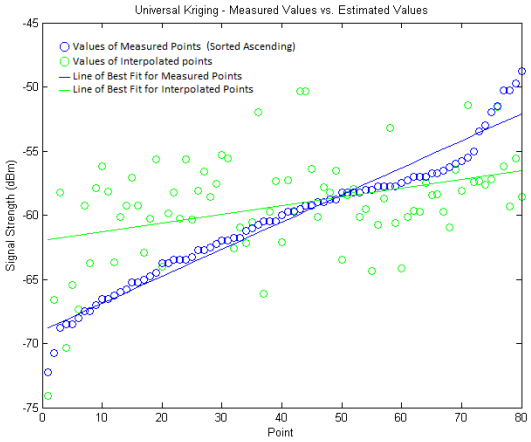



Fig. 4. Comparison between interpolated points and measured points

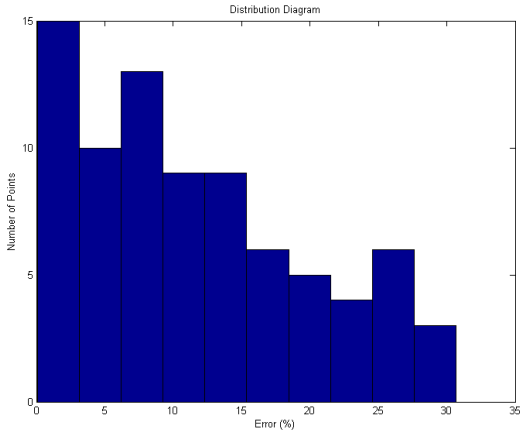



Fig. 5. Distribution of errors when using prudently selected samples

in each point. The error percentage ranges from 0.1% to 31% with an average error of 11.7% and a standard deviation of 8.4%. Therefore, as few as 29 strategically measured points are sufficient to describe the total interpolated area with an accuracy of 88.3%.

One would like to generalize the fact that 29 samples of the 110 total measurements were needed to estimate the signal strength with accuracy higher than 80%. However, if an infinite amount of samples were measured, 29 samples would still be enough to estimate the remaining points with an accuracy of more than 80%. Through cross validation it was found that 29 samples are sufficient to represent the distribution of Wi-Fi RSSI for this building. In practice one has to be able to determine the optimal number of samples, without measuring the rest of the building first.

## VII. Conclusion

In our investigation it was found that choosing a model variogram is an important aspect to consider as it has a significant influence on the accuracy of the model. The developed model will have similar results for instances where multiple Access Points are used in an indoor and partially outdoor environment, if samples are carefully considered and spatially diverse measurements are employed.

For this particular study, a simple sampling method was followed by choosing samples closest to the Access Points and some far away from them, with a few scattered points in between. The results of the cross validation using random points of the measured data confirmed that the method is consistent. To increase the accuracy and the confidence level of the model in general for different scenarios, it is important to set up a specific sampling plan. The sampling plan should include methods to determine the minimum number of samples necessary and methods to find the most effective positions for the samples.

Randomly selecting points results in an average accuracy of 84%, but with the possibility of significantly lower accuracies in worst case scenarios. An accuracy of 88% was achieved with as few as 29 prudently selected sampling points in a building with multiple Access Points using kriging for Wi-Fi signal strength estimation.

The results confirm that kriging is a remarkably accurate method for multiple access point Wi-Fi RSSI estimation. Kriging can be recommended for practical applications for its simplicity of taking measurements and applying the model, compared to the complexity of deterministic models.

## References


[1] H. J. Miller, "Toblers first law and spatial analysis," *Annals of the Association of American Geographers*, vol. 94, no. number, pp. 284–289, June 2004.

[2] P. R. Minnitt and D. W. Assibey-Bonsu. (2010) Professor d.g. krige - his contributions to research and engineering. [Online]. Available: http://www.gasa.org.za/wp-content/uploads/Tribute-to-Prof-Danie-Krige-2010-Edited.pdf

[3] J. Kruisselbrink. (2010) Kriging. [Online]. Available: http://www.liacs.nl/~jkruisse/material/Kriging_December_6_2010.pdf

[4] G. Bohling. (2005) Kriging. [Online]. Available: http://people.ku.edu/~gbohling/cpe940/Kriging.pdf

[5] A. Konak, "Estimating path loss in wireless local area networks using ordinary kriging," in *Simulation Conference (WSC), Proceedings of the 2010 Winter*, 2010, pp. 2888–2896.

[6] S. Latif, M. Ghazanfar, A. Memon, B. Chowdhry, and J. Ahmed, "Comparison of d-model and wall-attenuation model for signal strength estimations in indoor environment," in *Computational Intelligence, Modelling and Simulation (CIMSiM), 2012 Fourth International Conference on*, 2012, pp. 336–340.

[7] Ordinary kriging. [Online]. Available: http://www.ems-i.com/gmshelp/Interpolation/Interpolation_Schemes/Kriging/Ordinary_Kriging.htm

[8] D. Tang, G. Zhang, and J. Qin, "On the combination of spatial diversity and multiuser diversity under spatial correlation," in *Mobile Technology, Applications and Systems, 2005 2nd International Conference on*, 2005, pp. 4 pp.–4.

[9] K. Yang, H. Wang, G. Dai, S. Hu, Y. Zhang, and J. Xu, "Determining the repeat number of cross-validation," in *Biomedical Engineering and Informatics (BMEI), 2011 4th International Conference on*, vol. 3, 2011, pp. 1706–1710.

[10] V. Gandhi. Semivariogram modeling. [Online]. Available: http://www-users.cs.umn.edu/~gandhi/courses/CS8701/g4_semivariogram_final_draft.pdf

[11] F. P. Agterberg. Georges matheron - founder of spatial statistics. [Online]. Available: http://www.geostatscam.com/Adobe/G_Matheron.pdf

[12] A. Cantoni, "Optimal curve fitting with piecewise linear functions," *Computers, IEEE Transactions on*, vol. C-20, no. 1, pp. 59–67, 1971.

[13] W. Huizan, Z. Ren, L. Kefeng, L. Wei, W. Guihua, and L. Ning, "Improved kriging interpolation based on support vector machine and its application in oceanic missing data recovery," in *Computer Science and Software Engineering, 2008 International Conference on*, vol. 4, 2008, pp. 726–729.

P.J. Joubert is currently pursuing an M.Eng degree at the North-West University. He received his B.Eng in Computer Engineering at the North-West University in 2013.

# Effect of Node Pause Time and Speed on Routing Protocols in Mobile Ad-Hoc Networks

Botshelo T. Nokane[1], Michel Mbougni[2], Obeten Ekabua[1], Nosipho Dladlu[1], William Montshosi[1]
[1]Department of Computer Science
North West University, Mafikeng Campus,
Private Bag X2046, Mmabatho, 2735
[2]Department of Computer Science and Software Engineering
Faculty of Engineering and Computer Science
Concordia University, Montreal

Emails: nosipho.dladlu@nwu.ac.za; m_mbougn@encs.concordia.ca

**Abstract- Mobile Ad-hoc networks (MANETs) are self-configuring infrastructure of less networks of mobile devices connected by wireless medium forming rapidly changing topologies. In MANETs, every node contributing in the network acts both as host and a router; that is, every node is willing to forward and receive packets to and from other nodes. This paper presents a comprehensive performance evaluation of four different routing protocols namely; Ad-hoc On Demand Distance Vector (AODV), Dynamic Source Routing (DSR), Temporally-Ordered Routing Algorithm (TORA) and Optimized Link State Routing (OLRS), under different speed and pause time. The analysis was done using performance metrics throughput and delay under OPNET modeler 14.0. The overall results showed that the proactive routing protocol OLSR outperformed the reactive routing protocols AODV, DSR, and TORA under different speed and pause time.**

**Index Terms— MANETs, Routing protocols, Mobility Model, Random Waypoint,**

## I. INTRODUCTION

Mobile Ad-hoc Networks (MANETs) in present years, has found application particularly to defeat the limitation of bandwidth in wireless communication [1, 2]. A MANET is a collection of independent mobile nodes dynamically forming a temporary network without the fixed infrastructure and can survive rapid changes in the network topology [1, 3, 4]. The connection point in ad-hoc networks can be represented as both routers and hosts, so that nodes may develop packets between the other nodes as well as run user application [5]. The examples of nodes in such networks are personal digital assistants (PDAs), laptop [4, 5] and cellular phone [5]. There are various applications of MANETs such as military deployment [4, 5], mobile sensor networks [4], vehicle-to–vehicle communication and commercial application.

In MANETs, the movement of the nodes are unpredictable; so, steady routing protocols should be able to adapt to the unpredictable and dynamic topology of the [4, 6] caused by the random displacement of mobile nodes within a particular area. In MANETs, each mobile node moves randomly and at any time [4, 5, 6], it makes the network topology highly dynamic, thus making packets routing challenging. Accordingly, it is required for MANET to have routing protocols which can become accustomed to mobility and dynamically-changing topology. Hence, the mobility of nodes becomes very important when evaluating routing protocols. In this paper, an extensive performance analysis of four different routing protocols namely Ad-hoc On Demand Distance Vector (AODV), Dynamic Source Routing (DSR), Temporally-Ordered Routing Algorithm (TORA) and Optimized Link State Routing (OLRS) is done under different mobility scenarios; that is under different speeds and pause time.

## II. ROUTING PROTOCOLS

The network research community has been intensively working on modelling, designing and implementing new routing protocols for MANETs. Widely routing protocols are characterized into three categories proactive routing protocol, reactive routing protocol and hybrid routing protocol [4]. The three well-known reactive routing protocols, AODV, DSR, TORA and the well-known proactive routing protocol OLSR will be discussed in the next subsections.

### *A. Reactive Routing Protocols*

Reactive routing protocols find the new route only when needed and when there is no need for a node to keep route to destinations that are not in active communication. If a source node wants to know the route to a destination node, it actually sends the route request message in different path network and chooses the best route afterwards.

#### Ad hoc On Demand Distance Vector (AODV)

AODV [7] is one of the popular routing protocols in wireless networks. One of the most important characteristics of AODV is that its uses a destination sequence number for each route entry and for a guaranteed loop freedom. The source node will choose one of the greatest sequence numbers if the destination node is given the variety of multiple routes. The mainly current destination sequence number joining together with the destination and lifetime of the route is accumulated in the table. Throughout lifetime, if the route is not being used, the routing table entry is rejected. Once the error arises or the present selected routes have a change, the AODV have capacity to generate a new route for the rest of transmission for organization and maintenance.

**Dynamic Source Routing (DSR)**

DSR allows the network to be completely self-configuring and self-organizing [4, 5]. The source node identifies the entire path in the header of the data packet which includes the classified list of the nodes from the beginning to the end where the packet should pass through. Nodes work together by forwarding packets for every other to permit communication over multiple hops among nodes that are outside the wireless transmission range of every node. When the network alters the nodes approaching or leaving the network, or continues living nodes moving, all routing is automatically established by the DSR protocol. Nodes are needed to keep collection that can include multiple source routes to any destination. When the new routes are discovered, entries in the route store are continually informed.

**Temporally-Ordered Routing Algorithm (TORA)**

TORA is a routing protocol that is based on link reversal. It has the following characteristics [8]:

- It is an adaptive protocol, that is, it finds out the routes when needed
- It reacts minimally to topological changes and thus minimizes the communication overhead
- Multiple path routing distribution
- Routes are not necessarily optimal
- It uses a loop-free algorithm for routing
- It is a fast route finder algorithm.

*B. Proactive Routing Protocols*

Proactive routing protocols keep the routing information for each node within the network and inform their routing information at all-time irrespective of the routing request.

**Optimized Link State Routing (OLSR)**

OLSR is proactive routing protocol that uses the concept of Multipoint Relays (MPRs). MPR is an optimized flooding control protocol used by OLSR to construct and maintain routing tables by diffusing partial link state information to all nodes in the network [4].

The functioning of OLSR can be divided into the following three mechanisms:

• Neighbor/Link sensing.
• Efficient control flooding using MPR.
•Optimal route calculation using the shortest route algorithm.

## III. MOBILITY MODELS

Mobility models characterize the movement of node, behavior and their position; speed and acceleration change over time. Mobility models have a very essential role in determining the protocol performance in MANETs from other simulation parameters. This section presents an overview of two popular mobility model namely the random walk and random waypoint mobility models.

**Random Walk Model**

Random walk model is the simple mobility model from the random model that is based on random direction and speed. In particular, the Random Walk Model's node moves from current position to new position by randomly choosing the speed and direction for its path. Figure 1 shows an example of a node movement in Random Walk Model. The speed and direction of the node is selected according to a predefined range among [0, 2π] and [Speedmin, Speedmax] respectively [9]. Every movement in Random Walk Mobility Model happens either at a uniformly interval time t or uniformly distanced travelled d, at the end of which new direction and speed are calculated.

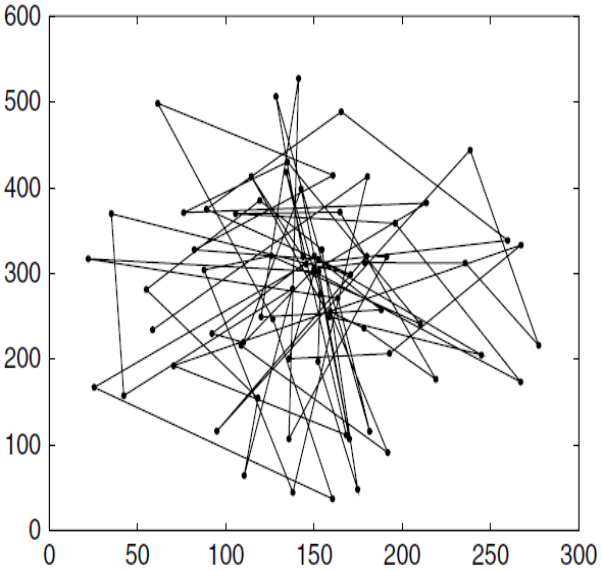



Figure 1: example of node movement Random Walk Model [10]

**Random Way Point Mobility Model**

The most commonly used synthetic mobility model used for wireless communication is the Random Waypoint model [11]. Every node selects a random destination contained by the simulated field and speed among the several minimum and maximum limits. Figure 2 shows an example of node movement in Random Way Point Mobility Model.

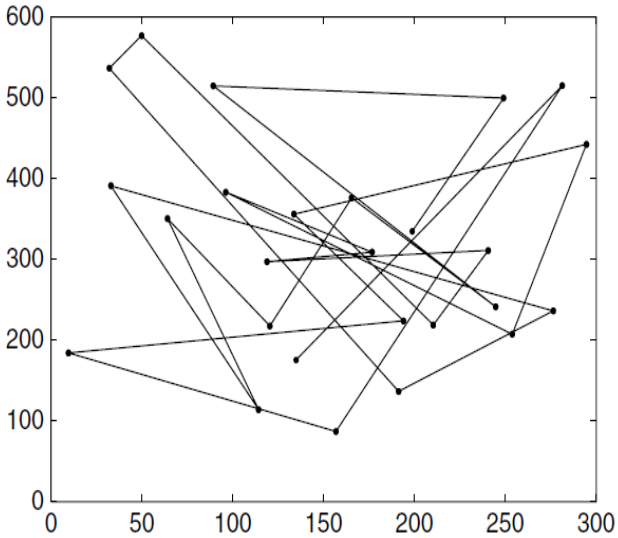



Figure 2: example of node movement in Random Waypoint [12]

In this paper, the random waypoint is used as the mobility model due to its popularity and simplicity in implementation.

## IV. SIMULATION SETUP

The simulation in this paper was run using OPNET Modeler 14.0 [13]; with 50 nodes randomly distributed in an area of 2500 m × 2500 m. The varying speed used was 2 metres per second and 10 metres per second with the varying pause time of 100 seconds and 500 seconds. The nodes moved following the random waypoint mobility model. The simulation parameters are summarized below:

**Table 1: Simulation parameters**

| Parameters | Values |
|---|---|
| Protocols | AODV, DSR, TORA, OLSR |
| Simulation Area | 2500 x 2500 meter |
| Number of Nodes | 50 |
| Transmission Power | 0.005 watts |
| Transmission Rate | 1 Mbs |
| Mobility Model | Random Waypoint |
| Type of Traffic | http (light browsing condition) |
| Transmission Range | 100m |
| Packet Size | 512 bytes |
| Performance Metrics | Throughput, Delay |

The profiles and scenarios modelled in this paper are represented in Table 2

**Table 2: Summary of profiles and scenario**

| Profile | Scenario |
|---|---|
| Profile1:Speed = 2 m/s | Scenario1: pause time = 100sec<br>Scenario2: pause time = 500sec |
| Profile2:Speed = 10 m/s | Scenario3: pause time = 100sec<br>Scenario4: pause time = 500sec |

## V. RESULTS AND DISCUSSION

This section presents the simulation results and analysis of routing protocols mentioned earlier.

### A. *Profile1: Speed of 2m/s.*

**Delay comparison of Scenario 1 and Scenario 2**

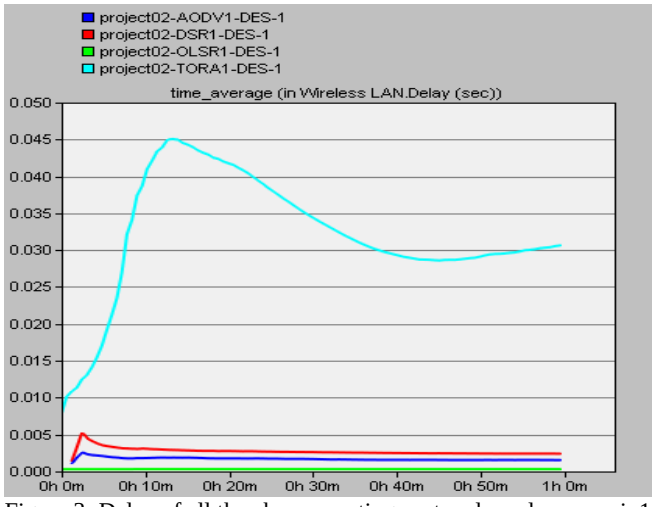


Figure 3: Delay of all the chosen routing protocols under scenario1

The performance of routing protocols AODV, DSR, OLSR and TORA in terms of delay under the speed of 2 m/s with pause time of 100 and 500 seconds is respectively shown in Figure 3 above and the Figure 4 below. The x-axis for both Figure 3 and Figure 4 represent the simulation time in hours while the y-axis represent the delays in seconds.

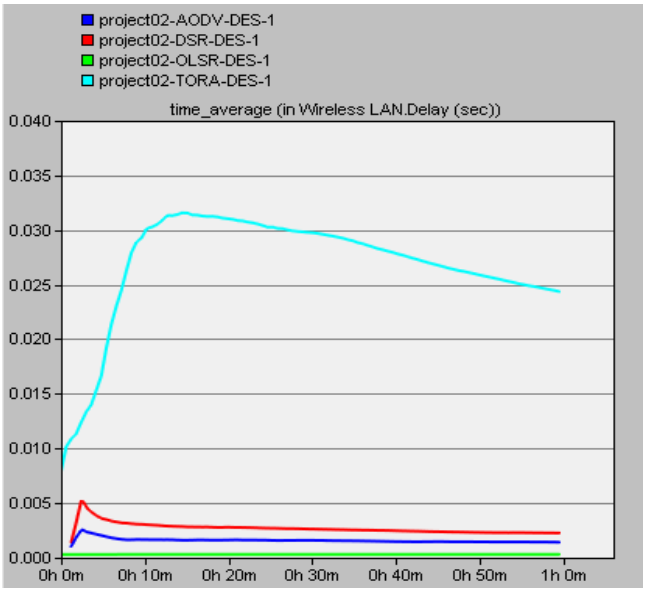


Figure 4: Delay of all the chosen routing protocols under scenario2

The Figure 3 and Figure 4 show that TORA experiences the longest delay, but the average delay of TORA in Figure 3 is more than the average in Figure 4. This illustrate that, TORA experiences the longest delay with the lowest speed and lowest pause compared to other scenarios; this is due to the fact that TORA route creation does not occur rapidly, primarily creating prospective long delays whereas waiting for discovery of new routes, but based on the pause time, TORA experiences a longest delay with a pause time of 100 seconds.

The Figure 3 and Figure 4 show that the DSR has the second longest delay with a pause time of 100 seconds compared to the pause time of 500 seconds. This is also caused by the route discovery mechanism performed by DSR. The AODV competes with OLSR in terms of short delay. However, the OLSR has slightly lower delay than the AODV. From the beginning of the simulation time, the AODV has a slightly longer delay for about 2 minutes, from there, AODV and DSR competes for the longest delay.

**Throughput comparison of Scenario 1 and Scenario 2**

The performance routing protocols AODV, DSR, OLSR and TORA in terms of throughput of the MANETs under the of speed of 2 m/s with a pause time of 100 seconds and 500 seconds is respectively shown in Figure 5 and Figure 6. The x-axis for both Figure 5 and Figure 6 represent the simulation times in hours while the y-axis represents the throughput in bit per seconds.

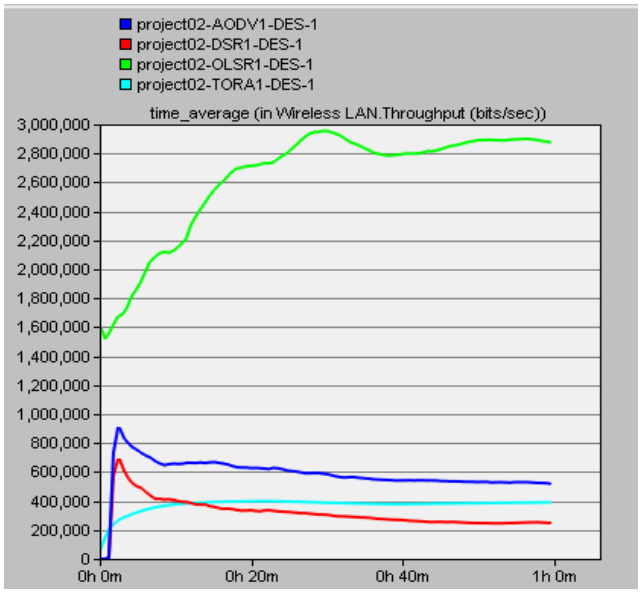


Figure 5: Throughput for all chosen routing protocols scenario1

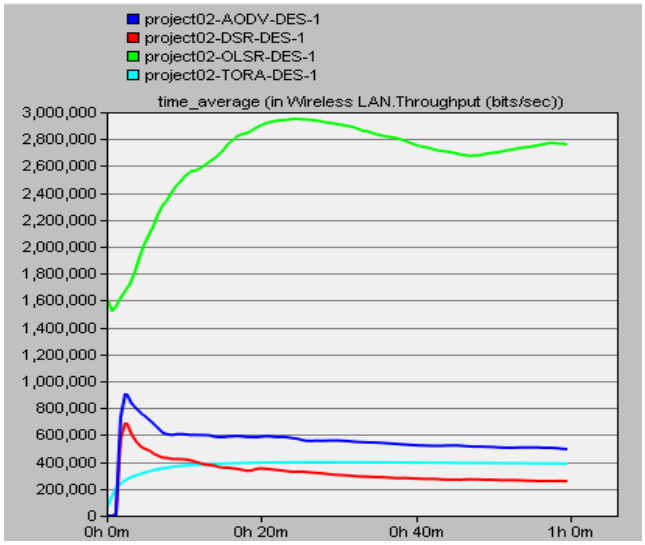


Figure 6: Throughput for all chosen routing protocols scenario2

In Figure 5 and Figure 6 indicates that all the reactive routing protocols AODV, DSR and TORA perform poorly in terms of throughput. In profile1, OSLR outperformed all the reactive routing protocols respectively under scenario 1 and scenario 2. This is due to the fact that all the paths are already available for the OLSR to transmit the node from the source node to the destination node. Also this implies that OLSR source node would be able to transmit more packets when the routing algorithm is applied on the nodes. The Figure 5 and Figure 6 also show that AODV has the second best throughput behind OLSR, but in terms of pause time, AODV performs better with a lower pause time, since the results of scenario 1 is better than the results of scenario 2; this due to the fact that AODV is a hop-by-hop routing protocol, while DSR exhibited worst throughput in both scenarios, this is due to the fact that DSR is a source routing protocol.

*B. Profile2: Speed of 10m/s.*

**Delay comparison of Scenario 3 and Scenario 4**

The Figure 7 and the Figure 8 represent the performance of AODV, DSR, OLSR and TORA in terms of delay under the speed of 10 m/s with a pause time of 100 seconds and 500 seconds respectively. The x-axis represent the simulation times in hours while the y-axis shows the delays in seconds

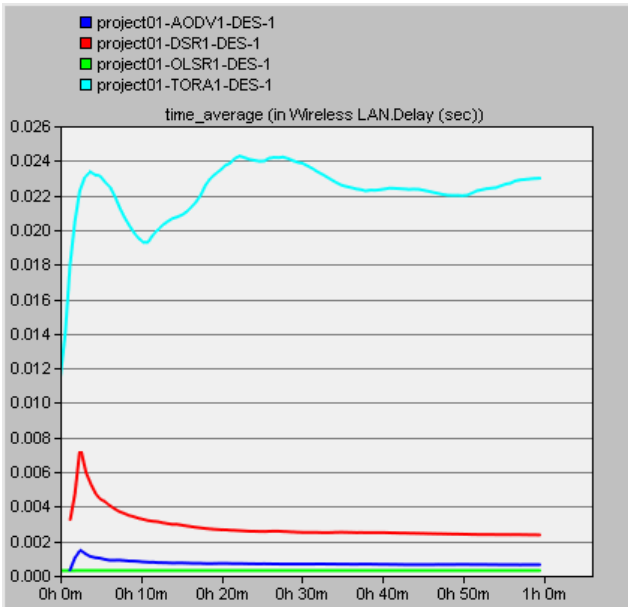


Figure 7: Delay of all the chosen routing protocols scenario3

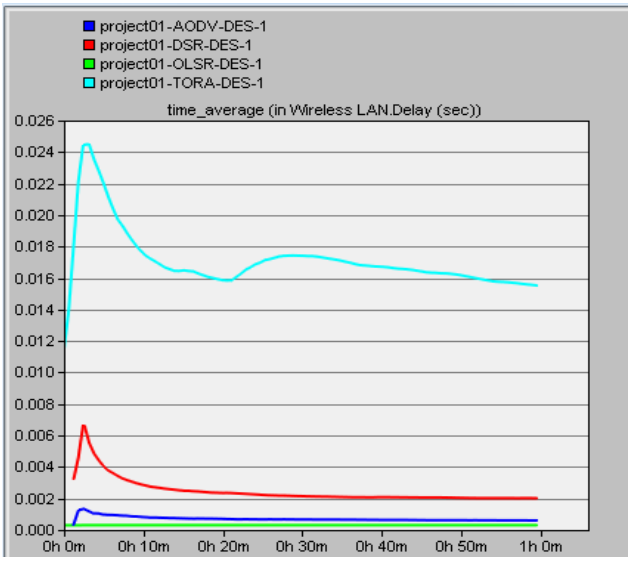


Figure 8: Delay of all the chosen routing protocols scenario4

The Figure 7 and Figure 8 show that TORA experiences the longest delay. This is due to the fact TORA route creation does not occur rapidly, primarily creating prospective long delays whereas waiting for discovery of new routes, but based on the pause time TORA experiences the longest delay with a pause time of 500 seconds. The Figure 7 and Figure 8 show that the DSR has the second longest delay under scenario 3 and scenario 4. This is also because of its route discovery mechanism. The AODV competes with OLSR in terms of short delay. However, the OLSR has a slightly lower delay than the AODV, but as the simulation time begins, the AODV has a slightly longer delay. The AODV as shown in Figure 8, has a shorter delay compared to other reactive routing protocols DSR and TORA. This is because of the hop-by-hop initiation process by AODV protocol on nodes.

**Throughput comparison of Scenario 3 and Scenario 4**

The Figure 9 and the Figure 10 represent the performance of AODV, DSR, OLSR and TORA in terms of delay under the speed of 10 m/s with a pause of 100 seconds and 10 m/s with a pause of 500 seconds respectively. The x-axis for both Figure 9 and Figure 10 represent the simulation times in hours while the y-axis represents the throughput in bit per seconds.

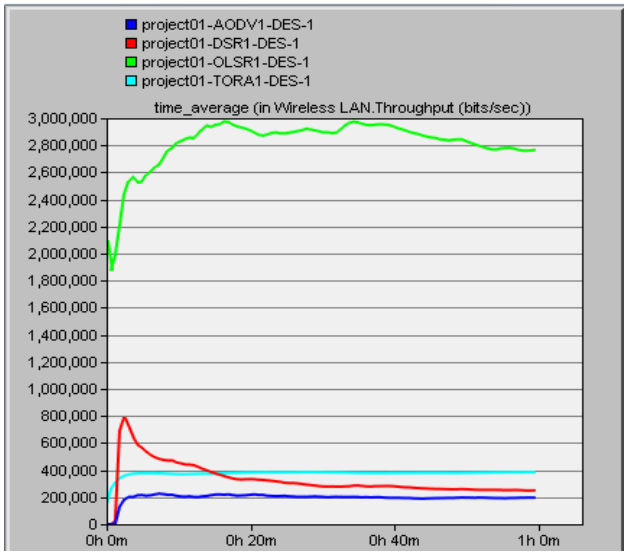


Figure 9: Throughput for all chosen routing protocols scenario3

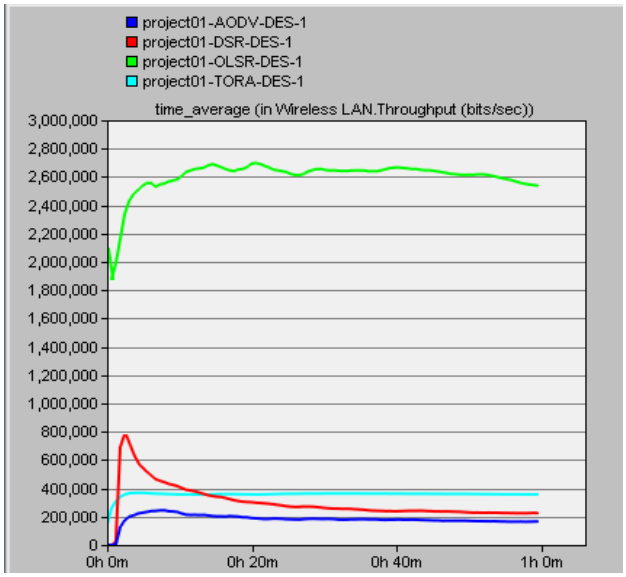


Figure 10: Throughput for all chosen routing protocols scenario4

In Figure 9 and Figure 10 show that the proactive routing protocol OSLR outperformed the reactive routing protocols AODV, DSR and TORA respectively under scenario 3 and scenario 4 in terms of throughput. This is due to the fact that all the paths are already available for the OLSR to transmit the node from the source node to the destination node. Also this implies that OLSR source node would be able to transmit more packets when the routing algorithm is applied on the nodes. However, in terms of pause time, the OLSR performs better with the lower pause time (100 seconds) because the average throughput in scenario 3 is better than in scenario 4.

The Figure 9 and Figure 10 also shows that DSR has the second best throughput behind OLSR at the beginning of the simulation times, after it then experiences lowest throughput behind TORA. This is due to the fact that DSR use source routing whereas TORA is a hop-by-hop routing protocol. AODV experiences the lowest throughput in scenario 3 and scenario 4, but AODV performs better in profile 1 than in profile 2. This shows that based on the speed, AODV experienced loss of packets with high speed. This is due to the fact that AODV uses only hello messages for neighbor detection.

## VI. CONCLUSION FUTURE WORK

In terms of delay, OLSR experienced the shortest delay but competed with AODV at an early stage of the simulation (simulation time). In terms of speed and pause time, OLSR outperformed the shortest delay with both minimum speed and pause time in all scenarios. DSR had the second longest delay behind TORA in all the scenarios. It was observed that TORA fluctuates and this is because TORA takes time to rebuild the route after a link failure. Therefore, it can be concluded that when MANETs are deployed in high mobile environments, OLSR and AODV are the first two protocols to be considered as they adapt very well to mobile nodes, compared to DSR and TORA which exhibits a poor performance in mobile environments.

In general, the results illustrated that in terms of throughput, the proactive routing protocol OLSR outperformed the reactive routing protocol AODV, DSR and TORA in all scenarios. AODV and DSR had the lowest throughput due to their route discovery process. Also, OLSR is the proactive routing protocol that transmits control messages to all the nodes, uses MPRs selector for each packet to be broadcasted and updates their routing information even if there is no actual routing request, hence, the routes are always up to date.

As future work, it will be interesting to use the same scenario used in this paper but with a much more complex routing protocol such as Gauss-Markov mobility model. It will also be interesting to see what will be effect of network load on the routing protocols simulated in this paper.

**ACKNOWLEDGMENT**

The authors would like to thank the Department of Computer Science North West University, Mafikeng Campus and the TELKOM CoE for their support.

**REFERENCES**

[1] A.K. Gupta, H. Sadawarti, A.K. Verma, "Performance analysis of AODV, DSR & TORA Routing Protocols." *International Journal of Engineering Technology*, Vol.2, No.2, pp. 226231 April 2010

[2] A. Goel, A. Sharma, "Performance Analysis of Mobile Ad-hoc Network using AODV Protocol." *International Journal of Computer Science and Security (IJCSS)*, Volume (3): issue (5), pp.334-343, 2009

[3] Mrs.M. Sivajothi and E.R. Naganathan, "Analysis of Reference Point Group Mobility Model in Mobile Ad hoc Network with an Ant Based Colony Protocol." *in*

*Proc. International MultiConference of Engineers and Computer Science 2009 Vol 1 IMEC 2009*, March 18-20, 2009, Hong Kong

[4] Nadir Shah, Depei Qian and Khalid Iqbal, "Performance evaluation of multiple routing protocols using multiple mobility models for mobile ad hoc networks." *Multitopic Conference, 2008 INMIC 2008 IEEE International*, pp. 243-248, IEEE 2008

[5] N.Gupta and R.Gupta "Routing Protocols in Mobile Ad-hoc Networks*." Proceedings of IEEE International Conference on the Emerging Trends in Robics and Communication Technologies*, 2012, pp. 173-177

[6] C. Perkins, "Ad hoc On-Demand Distance Vector (AODV) Routing," RFC, July 2003

[7] D.Vir, S.K.Agarwal, and S.A.Imam. "Simulation of energy consumption analysis of multi-hop 802.11 wireless ad-hoc networks on reactive routing protocols." *International Journal of Engineering Research and Application (IJERA)*, Vol. 3, Issue1, pp. 1213-1218, 2013.

[8] V. Park and S. Corson "Temporally-Ordered Routing Algorithm Version1 Functional Specification (TORA)", RFC 2026, July 2001

[9] Y.K. Hassan, M.H.A. El-Alziz and A.S.A. El-Radi. "Performance Evaluation of Mobility Speed over MANET Routing Protocols." *International Journal of Network Security*, Vol. 11, No.3, pp. 127-137, 2010.

[10] M.K. Jeya Kumar and R.S. Rajesh. "Performance Analysis of MANET Routing Protocols in Different Mobility Model." *International Journal of Computer Science and Networks Security (IJCSNS)* Vol.9, No.02, pp. 22-29, February 2009.

[11] C. Bettstetter, H. Hartenstein, X. Perez–Costa, "Stochastic Properties of the Random Waypoint Mobility Model," available at: http://xcosta.com/publications/bettstetter-rwp-winet-preprint.pdf

[12] B.Malarkodi, P.Gopal and B.Venkataramani, "Performance Evaluation of Ad-hoc Networks with different Multicast Routing Protocols and Mobility Models." *International Conference on Advances in Recent Technologies in Communication and Computing*, 2009 IEEE, pp. 81-84.

[13] OPNET 14.0 Documentation.

**Botshelo Thelma Nokane** received her BSc and BSc Honours in computer science respectively in 2012 and 2013 in North West University. She is currently an intern with VODACOM. Her research interest includes routing and performance evaluation of routing protocols in MANETs.

# MANAGEMENT

# Congestion Control in Multi-Serviced Heterogeneous Wireless Networks Using Dynamic Pricing (with Users' Willingness to Pay Incorporation)

Samson O. Orimolade and Olabisi Falowo *Senior Member IEEE*
Department of Electrical Engineering
University of Cape Town
Rondebosch 7701, Cape Town, South Africa
email: {orims, bisi}@crg.ee.uct.ac.za

*Abstract*— **Network congestion is still a major problem in multi-serviced heterogeneous wireless networks, despite the installation of more macro and micro base stations and deployment of more efficient radio access technologies (RATs) such as 3G and LTE. Firstly, congestion is caused by the increasing number of subscribers, ground breaking technologies in the smart-mobile world and advent of new user friendly applications. Secondly, as users utilize the resource constrained network, they crave to maximize their Quality of Service (QoS) and monetary utilities. The effect of users' willingness to pay on rate of calls has not gotten much attention recently. We attempt to solve this problem using an economic approach. The different behavioral context of users to changes in price can be used to control radio resource usage. Users' price sensitivity function inferred from their Willingness to pay (WTP) can be employed in determining optimal service price which will enhance providers' revenue and users' utility during congestion. Some existing pricing schemes do not consider subscribers' WTP in setting service prices, which reduces the efficient utilization of resources and revenue enhancement. Our proposed pricing scheme incorporates subscribers' WTP to determine optimal prices for the network services during congestion. Our results show that incorporating users' WTP in determining optimal price enhances users' utility and operators' revenue, while minimizing congestion.**

**Keywords—Heterogeneous wireless networks, optimal congestion price, Willingness To Pay (WTP), Radio access technology (RAT), Call Admission Control, common radio resource management (CRRM).**

## I. INTRODUCTION

Heterogeneous wireless networks (HWN) consists of multiple radio access technologies (RATs) coexisting in the same geographical location and supporting multiple services such as voice, video and data [1]. HWN suggests a combination of pool of resources provisioned by the coexisting and interworking of different Radio Access Technologies (RATs). In cooperative heterogeneous wireless networks (HWN), radio resources of the multiple RATs are jointly managed to enhance quality of service provisioning and improve radio resource utilization. The pool of resources is managed by the joint radio resource management (JRRM), sometimes referred to as the Common RRM (CRRM) schemes, aimed at achieving an efficient utilization of radio resources [2].

However, network congestion is still a major problem in HWN, despite the installation of more macro and micro base stations and deployment of more efficient RATs such as 3G and Long Term Evolution (LTE). The heterogeneity of service and RATs makes it imperative to devise an intelligent algorithm for allocating resources to users.

Congestion in radio access networks can be attributed to increase in data traffic volume resulting from the increase in the number of subscribers, advent of exciting new applications, availability of smart phones, and increasing demand for high-bandwidth consuming services such as video streaming, gaming and so on. Apparently, the limited radio resource is oversubscribed and individual RAT capacity is reached. Consequently, subscribers experience poor quality of service (QoS) and in the long run are dissatisfied.

Users are unaware of the network condition as they crave for more bandwidth consuming services and aim to maximize their utility [3]. In order to control this behavior, an economically efficient mechanism will be viable using a dynamic pricing approach. By dynamically pricing the network, an incentive is introduced to control the consumption of radio resources. The general assumption of network pricing is that users are price-sensitive [4]-users will reduce consumption when access price is high and vise versa.

Users' sensitivity to price creates the service provider an opportunity to control the network utilization, without necessarily installing any network equipment. As a result it is more cost-effective and higher revenues are generated. However, the service providers should aim to achieve a fair pricing mechanism that makes them *'subscribers' first choice'*, else congestion pricing can be viewed as promoting social unfairness [4].

The general approach of dynamic pricing is based on load condition only, with non-inclusion of users' preference in terms of their willingness to pay (WTP) in service price determination. Considering the increasing competitiveness of the telecommunication sector, the incorporation of users' WTP in evaluating service price will improve the price sensitivity of users, thereby making users to perceive a particular network provider to be 'budget friendly' than the

other. As a result more users are attracted to service provider that considers WTP, consequently providers' revenue and users' utility is enhanced.

Our pricing scheme proposes an efficient dynamic pricing scheme that will suit the increasing competitiveness of telecommunication network market and users craving for high bandwidth consuming services. A provider keen on maintaining the lead in the market should adopt efficient pricing schemes that will evolve with the current telecommunication market trend.

### *A. Related works*

Several literatures have considered dynamically pricing the network based on load only, with no cognizance to users' WTP inclusion in evaluating service price. The following are some of the literatures that did not consider users' WTP in setting congestion price.

In [5], B. Al-Manthari *et al* propose a call admission control-based dynamic pricing scheme that aims at preventing congestion and maximizing the utilization of resources in wireless access systems. The authors dynamically compute the price of units of bandwidth, forcing actual number of connection to optimal ones based on network load only.

In [6], Feng Chen *et al* presents a contract binded call admission control (CBCAC) pricing scheme which calculates optimal arrival rates and limits the call admitted.

Scheme [7], R. Piqueras *et al* proposes a dynamic pricing for decentralized RAT selection in heterogeneous scenario. They suggest that whenever the load in one RAT exceeds certain threshold, another substitutive RAT is made more attractive to users by means of the offered price. However, due to different user behavior, some subscriber can have a high WTP for the congested RAT, thereby increasing the operators' revenue.

The pricing schemes examined above did not consider incorporating users' WTP in determining service price. Noting that there exists a challenge in ascertaining a common ground between service price and WTP for the benefit of operators and users respectively, the contribution of this paper is to determine the optimal congestion price that will minimize congestion, while enhancing user's utility and operators' revenue in a HWN.

The result of our proposed scheme called the WTP incorporated (WTPI) dynamic pricing scheme, show that users' WTP inclusion in determining optimal congestion price enhance users' utility and providers' revenue.

The rest of the paper is organized as follows: In Section 2, a brief description and function of pricing scheme is presented. Section 3 examines congestion in heterogeneous wireless networks, effects of congestion and modeling of users' willingness to pay. The optimal congestion price analysis was also presented in section 3. Numerical simulations, performance comparisons and results are explained in Section 4. Section 5 draws the conclusion of this paper.

## II. PRICING SCHEMES

Pricing schemes are used by service providers (SPs) to provision QoS to subscribers [8] as well as to maintain profitability among competitors. An efficient pricing scheme is one able to set fair prices of service to subscribers and ensure the efficient utilization of radio resources. A broad categorization of pricing schemes are fixed and dynamic pricing schemes, with the later proven to be more profitable [8], due to the introduction of incentives. The pricing schemes are usually integrated with the conventional call admission control (CAC) schemes [5], to achieve a more efficient network. On the other hand, some pricing schemes suggest setting dynamic price only during peak period of the day. However, this reduces operators' opportunity of increasing revenue, because separate treatment of congestion as a function of time or network is not efficient in heterogeneous wireless networks. Pricing a network is usual based on policy adoption which differs with service providers.

Network service providers adopt price policies which enable them obtain revenue for services provided. Policies are set of outlined rules that states how a network should behave, in this context a price policy will determine how service prices should be evaluated. Pricing of the network services may differ, though offering same services. The pricing algorithms are implemented in the Mobile switching center of a GSM network, while in an EPC core network, the policy and charging rules function (PCRF) handles real-time charging rules and functions for each service in the network. The PCRF is a software component (node) where SPs implement a multitude of real-time charging rules and functions for each service in the network and IP multimedia services (IMS) or Evolved Packet Core (EPC) core networks [9].

## III. CONGESTION IN HETEROGENEOUS WIRELESS NETWORKS

Heterogeneous networks can simply be defined as the coexistence of large, small and tiny RATs, interworking to provide pool of resources. Each constituting RAT can have different capacities and physical characteristics [10]. The figure 1 below shows a Heterogeneous network setup consisting of 2G 3G, WiFi, LTE access networks. The RATs have different characteristics in terms of bandwidth, application, and coverage, ideally under the control of a single entity- the common radio resource management (CRRM) entity.

The advent of higher demand for bandwidth consuming services creates congestion on the heterogeneous networks, despite the installation of higher capacity equipments. The trend in the telecommunication over the years shows that demand always meets up with provisioned capacity, thereby posing a challenge of matching demand and capacity efficiently.

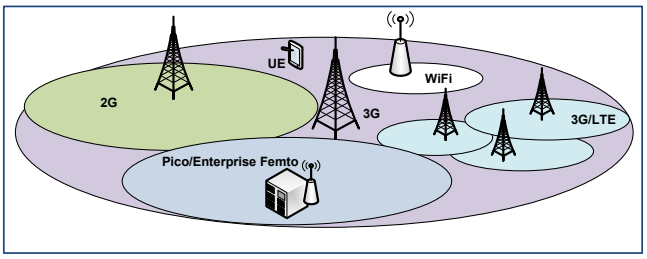



**Figure 1: Heterogeneous Wireless Network illustration**

Wireless networks heterogeneity birthed the emergence of diverse bandwidth consuming services. Subscribers still 'crave' for more bandwidth consuming services existing on the HWN like high quality voice and gaming, high definition video, real-time video and video streaming. Consequently, congestion problem may still be experienced. Therefore congestion control in HWN still remains an issue. HWN in this present convergent paradigm of telecommunications therefore requires a dynamic network management to minimize congestion issues. The primary aim of this work is how to minimize overall congestion in a HWN by determining suitable prices at different times.

*A. Network condition*

Heterogeneous networks are a combination of several RATs that jointly provide access to radio resources [11]. The individual RATs making up the heterogeneous network supports multiple services such as video, voice, data and so on. The total state space of the HWN comprises total participating RAT and services available on the RATs.

Consider a network having $m$ total number of RATs and s total number of services, the state space can be written as [11]:

$$\Omega = \{p_{nij}, q_{hij}, : i = 1,..m, j = 1,..s, b = b_j \forall j \in s\} \quad (1)$$

Let $C_{ms}$ be the total capacity of the HWN and $L_t$ be usable or admissible load capacity. $L_t$ is expressed in (2) below.

$$L_t = \sum_{j=1}^{s} b_j(p_{nij}) \le t_n \ \forall i \wedge b_j(q_{hij}) \le t_h \ \forall i \ \wedge \sum_{j=1}^{s} b_j(p_{nij} + q_{hij}) \le B_m \ \forall i \quad (2)$$

Where $m$ is the total number of RATs in the HWN $(i = 1,..m)$, $s$ is the total number of available services in the HWN $(j = 1,..s)$, $b_j$ is the basic bandwidth unit (bbu) needed to make a call of any class of service, $B_m$ is the total bbu of the service in the HWN. $p_{nij}$ is the number of ongoing newly accepted calls and $q_{hij}$ is ongoing handoff call, in RAT-$i$ with service class-$j$ respectively. The network provider ensures that $L_t < C_{ms}$ to avoid high call dropping and call blocking levels.

$$\sum_{i=1}^{m} \sum_{j=1}^{s} B_j(p_{nij} + q_{hij}) < C_{ms} \quad (3)$$

The provider develops a pricing scheme for the efficient operation of the network as in equation (3) above.

*B. Call blocking and call dropping*

New call blocking and handoff call dropping are measures of determining congestion in the network. Call blocking and dropping probabilities increases as network load increases and vise versa.

At high congestion levels, new calls will not be accommodated into the network as a result of limited network resources, apparently new calls will be blocked. Likewise, if the reserved bandwidth for the handoff calls is exceeded, subsequent handoff calls will be dropped. The ongoing calls are request that have being granted access to use the radio resources. The ongoing calls, both handoff and newly accepted calls determines the current network load condition. As different service class of calls are admitted into the network, with different basic bandwidth units, the probability of accepting a new call or handoff call request reduces gradually, until a call request is blocked or dropped.

Hence, the probability of accepting a new call or handoff call into a HWN is determined by the availability of resources. Recalling from equation (2), the bandwidth of handoff and new ongoing calls is given as $q_{hij}$ and $p_{nij}$ respectively. A new call is blocked if the condition in (4) is met and a handoff call is dropped if the condition in (5) is met [11].

$$\sum_{j=1}^{s} b_j p_{nij} > t_n \ \vee \sum_{j=1}^{s} b_j(p_{nij} + q_{hij}) > B_m \ \forall m \quad (4)$$

$$\sum_{j=1}^{s} b_j q_{nij} > t_h \ \vee \sum_{j=1}^{s} b_j(p_{nij} + q_{hij}) > B_m \ \forall m \quad (5)$$

Where $t_n$ is the threshold capacity of new calls and $t_h$ is the threshold capacity of handoff calls which cannot be exceeded, $B_m$ is the maximum combined threshold of calls a RAT can accommodate. The Erlang B loss system is used to model the new call blocking and handoff call dropping probabilities [12]. Using the traffic intensities of new calls and handoff calls given as $\tau_n = \frac{\lambda_n}{\mu_n}$, and $\tau_h = \frac{\lambda_h}{\mu_h}$.

Considering the scope of this research, the relevant metric to this study is the new call blocking probability, because in a dynamically priced network scenario, handoff calls are only charged at admission level, therefore handoff calls are not affected by dynamic prices since they have being charged from the cell where the call was initiated [4]. As a result we only examine congestion levels based on call blocking probability of newly initiated calls.

*C. Users' willingness to pay*

Users are price sensitive, therefore the service price of a network will determine users' preference among competing service providers. Users' willingness to pay comes into play when users perceive their utility is enhanced by a service provider than the other. Consequently, a provider that pays cognizance to users' willingness to pay in evaluating service price will achieve an enhancement of revenue.

Users will pay for a service only if the price of the service enhances their utility and will not pay otherwise. Therefore

our proposed scheme, WTPI will ensure an economically efficient network that incorporates users' WTP when evaluating the network congestion price. The primary aim is to obtain an optimal congestion price of services. Our proposed pricing scheme suggests an economically efficient and fair network. The figure 2 below gives the classification of the different methods for measuring WTP [13]. The adoption of any method depends on the nature of task, inherent limitations and accuracy of measurements that will be obtained.

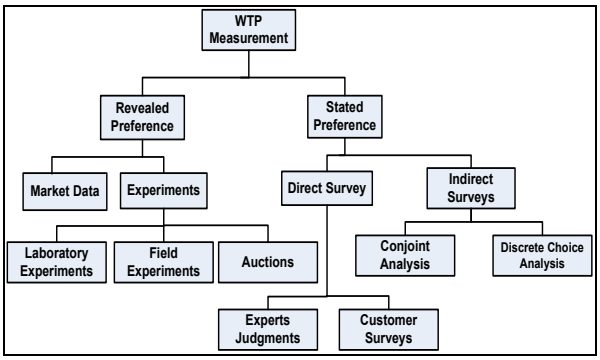



**Figure 2: Methods of measuring WTP**

For the HWN scenario, the method adopted in this paper is the revealed preference, obtained from historical market data of subscribers' response to service price updates.

On the other hand WTP values can vary with geographical location, individual budget, level of education, age and sex [14]. Some areas may have a high WTP than others, thereby creating better opportunity for the providers to maximize profit. Therefore service providers' knowledge of users' WTP can enhance their revenue.

*1) Modelling users' willingness to pay*

Modeling WTP in this scenario will be dependent on the historical responses of users to previously broadcasted or updated service price. More accurate results will be obtained because this will reveal the true preference of users during congestion periods. Obtaining WTP through surveys will be flawed in this scenario because of users' heterogeneity and diverse user preferences. Furthermore, if users reveal their true valuation of a service, it doe s not directly translate into real purchasing behavior [13]. The WTP differs according to users' importance attached to the class of service demanded. While some users can pay high amount for voice, other will rather pay higher for data services. It is assumed that there exists a price, $p$ which is above users' WTP [15]. Let the arrival rate for which WTP is below price $p$ be denoted as $\lambda^*$. Demand is a function of price, therefore it determines the arrival rate of subscribers to the network and call holding time. Arrival rate determined by a demand function is given as follows [4];

$$\lambda_{ij}(t) = f(p_{ij}) = \sigma_{ij}(t) e^{-\theta_{ij}(t) p_{ij}(t)} \qquad (6)$$

$\lambda_{ij}$ is a Poisson arrival process of a subscriber into RAT-$i$ $(\forall i \in m)$ and demanding service-$j$ $(\forall j \in s)$. Where $\sigma_{ij}(t)$, is the demand shift constant, and $\theta_{ij}(t)$, is the price elasticity of demand (the change in demand for a product or service due to changes in price) and $p_{ij}(t)$ is the unit price of bandwidth of call requested in that time instance. The equation (6) above gives the representation of users' WTP or preference for the services, considered in terms of users' call arrival rate, using a demand function. Due to the variance in the QoS of service classes and different user behavior, the price elasticity of demand, $\theta_{ij}(t)$ and demand shift constant, $\sigma_{ij}(t)$ can assume different values at anytime of the day [4]. The demand function above takes care of these changes. An important property of the demand function states that marginal revenue with respect to prices will always be negative. Fig. 3 shows the price sensitivity function of users to price variation, derived from equation (6). It shows that WTP can be inferred from users' price sensitivity.

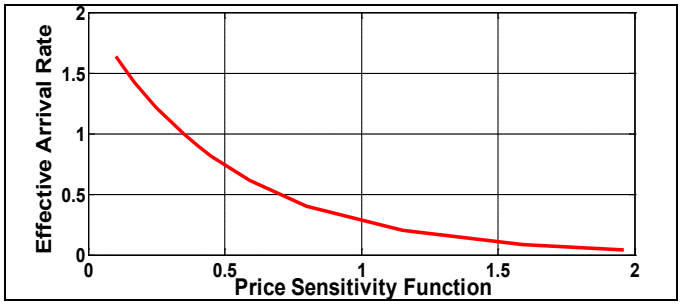



**Figure 3: Variation of Arrival rate inferred from Price sensitivity function**

*D. Optimal congestion price*

In order to obtain the optimal congestion price of services in a network, (considering both providers' and users' preference), some optimal measures need to be observed. The optimal price as obtained in [15] is modified to suit a HWN as given in (7) below. The operator will prefer to choose a price of services so as to maximize revenue over a period of time in a given RAT-$i$, which is the following objective;

$$P(i,t) = \frac{1}{u_i} \sum_{j=1}^{s} p_{ij}^* \max_{p(t)} \int_0^T \lambda(p,t)\, p(t)\, dt \qquad (7)$$

Subject to:

$$(p_{ij}^*)\, \lambda(p,t) - uq \le Q \qquad \{\forall q \in s\} \qquad (8)$$

Where $$\frac{dq^*}{dt} = (p_{ij}^*)\, \lambda(p,t) - uq$$

The table 1 below gives the definition of the variables in (7) and (8)

Table 1: Definitions of optimal congestion price equation variables

| Variables | Definition |
|---|---|
| $\frac{1}{u_i}$ | Service rate of RATs in the HWN |

| | |
|---|---|
| $P_{ij}^{*}$ | Optimal probability of acceptance of a call request in RAT-i, requesting service j |
| $0-T$ | Duration of a business cycle |
| $\lambda(p,t)$ | Call arrival rate at time $t$ , when price is $p$ |
| $p(t)$ | Unit price per bandwidth requested |
| $\frac{dq^{*}}{dt}$ | The optimal QoS requirement of ongoing calls in that instant of time that must not be violated |
| $q$ | QoS of expected number of calls in progress |
| $Q$ | Maximum QoS the available bandwidth can guarantee |
| $uq$ | It's the additional QoS of the departing or terminating calls |

Equation (8) is the constraint on QoS commitment to ongoing callers.

On the other hand the users aim to maximize their QoS utility and monetary utility [16]. Let $\lambda^{*}(p,t)$ be the optimal call arrival rate which depicts the average WTP of users. Therefore incorporating users preference into (7), yields an optimal congestion price, $P(i,t)$ given in (9) below;

$$P(i,t)=\frac{1}{u_i}\sum_{j=1}^{s}p_{ij}^{*}\max_{p(t)}\int_{0}^{T}\lambda^{*}(p,t)\,p(t)\,dt \qquad (9)$$

Subject to: (8) above and (10) below

$$p(t)\le WTP_{\max} \qquad (10)$$

Equation (9) introduces an objective function which is maximized by the average arrival rate of previously updated historical congestion price. All variables in (9) retain their meaning, (10) is the WTP constraint, $WTP_{\max}$ is the maximum willingness to pay boundary, a price above which users will not use the service no matter how good the QoS is [16].

Operators do have an expected target of revenue and profit envisaged for every fiscal and operational year. From the forgoing, the expected revenue is dependent on the number of successful calls admitted and the duration of each calls or data volume consumed by subscribers. The profit is obtained by deducting the capacity and running costs from the revenue in equation (7). Traffic intensity, $\tau$ - in terms of connection time, $t$ and arrival rate, $\lambda_{ij}$ is expressed as;

$$\tau=\lambda_{ij}\,t \qquad (11)$$

The expected number of connection, $E_k$ which relates to the usable load capacity of the system and acceptance probability, can be written as [17];

$$E_k=N\tau(p_{ij}) \qquad (12)$$

Therefore, total number of connection in the network is expressed as,

$$E_T=N(\tau_n(p_{kin})+\tau_h(p_{kih})) \qquad (13)$$

Expected revenue, $R_E$ is the product of total number of connections $E_T$ and unit price per bandwidth of the service class requested [15, 17].

## IV. NUMERICAL RESULTS

The performance of the proposed dynamic pricing scheme is investigated using MATLAB simulation tool. The HWN scenario considered is a two RATs mechanism, each having different capacities. The parameter for the HWN considered are as follows: C1= 20, C2 = 30, b1 = 2, b2 = 3, u1= u2= 0.6, tn1=14, tn2=22, th1=20, th2=30. The arrival rates, $\lambda$ into the individual RATs are obtained from historical prices, $p$ given by the demand function in (6). Unit price per bandwidth for voice, video and data are 0.2, 0.4 and 0.1 monetary values respectively.

Figure 4 below shows the effective arrival rate obtained by incorporating users' WTP in dynamic pricing as compared to the scheme in [5].

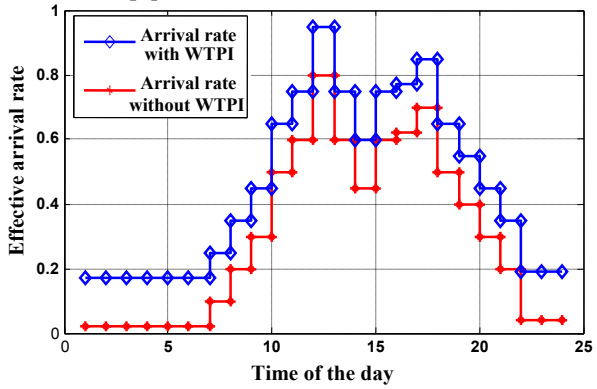



**Figure 4: Effective arrival rate obtained by incorporating WTP in dynamic pricing**

Figure 5 below shows reduction in new call blocking probabilities with the incorporation of WTP in the dynamic pricing scheme. A significant reduction in the blocking probability levels was achieved, illustrating the congestion control feature of the proposed dynamic pricing scheme as compared with other schemes. Results show that call blocking probabilities was reduced considerably with an optimal arrival rate.

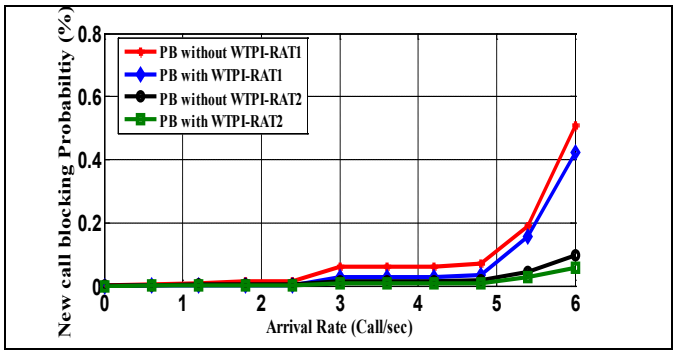



**Figure 5: Reduced new call blocking probability**

Figure 6 below shows a more efficient utilization of radio resource achieved by users' WTP inclusion. Meaning more users gained access into the network, with significantly low blocking and dropping probabilities.

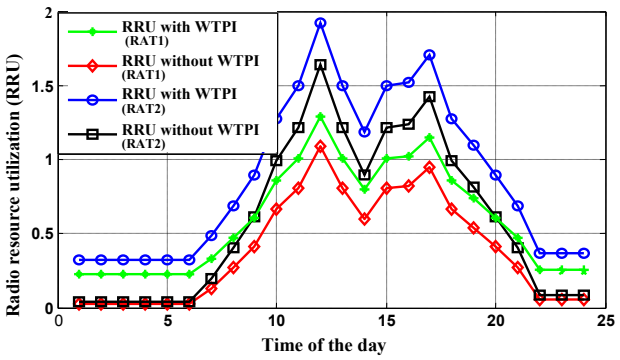



**Figure 6: Radio resource utilization enhancement achieved by proposed pricing scheme**

Figure 7 and 8 below shows results of users' utility and providers' revenue enhancement, effectively during congestion periods.

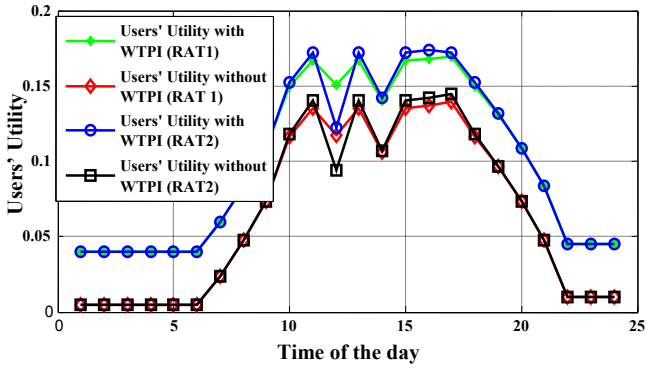



**Figure 7: Users' monetary utility enhancement**

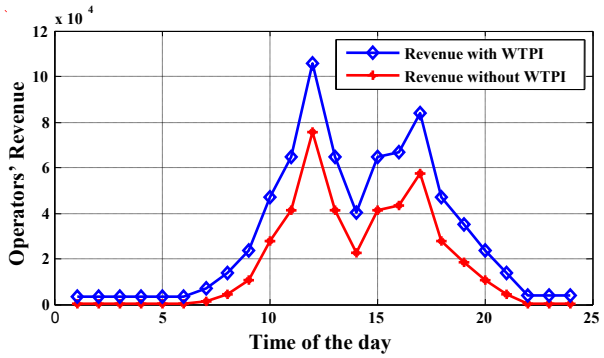



**Figure 8: Operators' revenue enhancement on voice calls**

## V. CONCLUSION

In this paper, the effect of incorporating WTP of users' in evaluating congestion price in HWN was examined. Users' WTP data was historically obtained from arrival rates of previous updated service prices. This paper has developed an analytical model of a dynamic pricing scheme (WTPI), which incorporates users' WTP in evaluating service price. An improved congestion service price was achieved compared to other dynamic pricing schemes. Numerical simulation results show that congestion was minimized while enhancing users' utility and operators' revenue.

*ACKNOWLEDGEMENTS*

*This research is supported by Telkom South Africa, Jasco /TeleSciences and the Department of Trade and Industry / National Research Foundation/Technology and Human Resources Programme (DTI/NRF/THRIP).*

## REFERENCES

[1] O. E. Falowo and H. A. Chan, "Effect of call dynamics of a multiservice multimode terminal on RAT selection in heterogeneous wireless networks," in *Global Communications Conference (GLOBECOM), 2012 IEEE,* 2012, pp. 5249-5253.

[2] *Radio Resource Management in Heterogeneous Networks*. Available: http://researchwebshelf.com/uploads/289_T02.pdf.

[3] M. Manaffar, H. Bakhshi and M. Pilevari, "A new dynamic pricing scheme with call admission control to reduce network congestion," in *Advanced Information Networking and Applications - Workshops, 2008. AINAW 2008. 22nd International Conference on,* 2008, pp. 347-352.

[4] B. Al-Manthari, N. Nasser and H. Hassanein, "Congestion Pricing in Wireless Cellular Networks," *Communications Surveys & Tutorials, IEEE,* vol. 13, pp. 358-371, 2011.

[5] B. Al-Manthari, N. Nasser, N. A. Ali and H. Hassanein, "Congestion prevention in broadband wireless access systems: An economic approach," in *Computers and Communications, 2009. ISCC 2009. IEEE Symposium on,* 2009, pp. 606-611.

[6] Feng Chen, Rui Ni, Xiaowei Qin and Guo Wei, "A novel CAC-based network pricing scheme for wireless access networks," in *Wireless Information Technology and Systems (ICWITS), 2010 IEEE International Conference on,* 2010, pp. 1-4.

[7] R. Piqueras, J. Perez-Romero, O. Salient and R. Agustí, "Dynamic pricing for decentralised rat selection in heterogeneous scenarios," in *Personal, Indoor and Mobile Radio Communications, 2006 IEEE 17th International Symposium on,* 2006, pp. 1-5.

[8] C. A. Gizelis and D. D. Vergados, "A Survey of Pricing Schemes in Wireless Networks," *Communications Surveys & Tutorials, IEEE,* vol. 13, pp. 126-145, 2011.

[9] *Testing policy and charging functions PCRF PCEF*. Available: http://www.tekcomms.com/success-stories/testing-policy-and-charging-functions-pcrf-pcef?gclid=CJGMrdjY3roCFSrJtAodwRQA6A. November 11th, 2013.

[10] Telecoms academy, "Heterogeneous networks," Africacom conference 2013, Cape Town South Africa, 2013.

[11] O. E. Falowo and H. A. Chan, "Join call admission control algorithm to enhance connection-level QoS in heterogeneous cellular networks," in *AFRICON 2007,* 2007, pp. 1-7.

[12] V. Shakhov, "Simple approximation for erlang B formula," in *Computational Technologies in Electrical and Electronics Engineering (SIBIRCON), 2010 IEEE Region 8 International Conference on,* 2010, pp. 220-222.

[13] C. Breitdert, M. Hahsler, and T. Reutterer, "A review of methods for measuring willingness-to-pay," vol. vol. 2, Innovative marketing, 2006.

[14] M. O. Oduh, M. Ogechi, "Determinants of Willingness to Pay for Mobile Telecommunications Services in Nigeria", Journal of Information Engineering and Applications, ISSN 2224-5782 (print) ISSN 2225-0506 (online), Vol. 2, No 6, 2012.

[15] Q Wang, J M Peha, M A. Sirbu, "Optimal price for Integrated-Services Networks with Guaranteed Quality of Service ," Internet Economics, McKnight, L. and Bailey, J. eds., MIT Press, Cambridge, Mass., paper 453, 1997.

[16] T. O. Kamoto and T. Hayashi, "Analysis of service provider's profit by modeling customer's willingness to pay for IP QoS," in *Global Telecommunications Conference, 2002. GLOBECOM '02. IEEE,* 2002, pp. 1549-1553 vol.2.

[17] N. J. Keon and G. Anandalingam, "Optimal pricing for multiple services in telecommunications networks offering quality-of-service guarantees," *Networking, IEEE/ACM Transactions on,* vol. 11, pp. 66-80, 2003.

**Samson Orimolade** received his BSc. degree in Electrical and Computer Engineering, from Ahmadu Bello University, Zaria Nigeria in 2010. He is presently studying towards his MSc. at University of Cape Town, South Africa. His research interests include Radio Resource Management and Planning, Pricing Schemes, and Optimization Algorithms.

# Freight tracking cost analysis to improve logistics management operations

A de Coning and AJ Hoffman
School of Electrical, Electronic and Computer Engineering
North-West University, Potchefstroom Campus, South Africa
Tel: +27 18 299 1971, Fax: 086 521 2569
Email: {20270186, alwyn.hoffman}@nwu.ac.za

***Abstract*—The cross-border freight industry suffers from inefficiencies that can be addressed by implementing technology based systems. Inefficiencies cause time delays at amongst others border posts and weigh stations that in turn increase turnaround times. Higher turnaround times affect the business bottom line as it reduces the turnover generated on capital assets and therefore decrease profit levels. Efficient logistics management will also ensure increased trade in Sub Saharan Africa which will ensure economic growth in the region. GPS tracking systems deployed on fleets of freight vehicles are currently focused on vehicle theft recovery and communication with the units is typically suspended once a national border is crossed to avoid roaming costs. This article proposes a mind shift towards the increased use of cross-border communication with freight vehicles in order to use GPS tracking and other telemetry data available from vehicles to enable improved logistics management. The article will analyse GSM and satellite communication costs and compare these costs with the potential financial benefits to be gained from the use of more frequent communications. Our analysis has shown that the use of telemetry data for logistics management offers the potential to increase profits by up to R400,000 per vehicle over its lifetime. Careful selection of the most suitable networks is however essential to achieve a positive cost-benefit ratio.**

***Index Terms*—Freight logistics, Freight, Road freight, Cross-border, Logistics management**

## I. Introduction

Supply chain efficiency can greatly increase or decrease the financial turn over and profitability of a freight business [1]. This efficiency can be increased by incorporating proactive supply chain management systems into business operations [2]. Implementing intelligent transport systems (ITS) that have a focus of logistics can result in higher efficiency and productivity in the supply chain [3]. These systems are highly dependent on information from the daily supply chain operation. This in turn requires real time information sharing within the supply chain.

The scale of supply chain operations, specifically freight transport which is the focus point in this study, is on the increase in Sub Saharan Africa [4]. The use of technology within this industry for tracking purposes is not a new concept but has till present been focused on security and vehicle recovery [5]. One of the most common features implemented is global positioning system (GPS) tracking of the freight vehicle. These tracking systems are a necessity for insurance purposes. Unfortunately these services mainly focus on security tracking with a minimal effort to improve the day to day efficiency of the logistics operation. This has resulted in a situation where tracking companies configure their systems to implement services at minimum costs causing communication with the vehicle to be suspended once it travels across a national border until the vehicle returns to within the local cellular network. Due to the pricing models applicable to roaming under the global system for mobile (GSM) communication, which is the communication option used by most tracking systems, costs increase exponentially when the unit crosses a border and has to switch over to a different network.

The purpose of this paper is to demonstrate that the benefit of utilizing GPS tracking and other types of telemetry data with a focus on logistics management will in most cases outweigh the cost implication associated with cross-border communication. Firstly the transporter, consignor and consignee can benefit from the use of tracking data by generating alarms as soon as an exception has occurred (e.g. when the doors were opened at an unauthorized location) and by sending regular updates about the current status of the consignment (e.g. advanced delivery notifications sent to client sites to allow early preparation at the goods receive depot). In the process unnecessary delays can be eliminated, resulting in decreased delivery times, and decreased turnaround times which in turn will increase revenue. Secondly governmental agencies responsible for inspecting vehicles at overload control stations and border posts can use GPS and other telemetry data received from freight vehicles to reduce the risk levels associated with such vehicles, resulting in reduced inspection frequencies; this in turn will result in reduced border crossing times and weigh station delays [4].

A number of different communication options are available for vehicle telematics units. Several GSM network operators provide coverage into most of Wouthern African Development Community (SADC) and neighbouring countries; the size of data packets that are handled by the different networks are however not identical, and there are significant discrepancies between the rates offered. In the more remote areas GSM coverage is either absent or unreliable, resulting in many transporters opting for satellite communication or a combination of GSM and satellite. The satellite based service are in general more expensive per data packet or monthly subscription, but

covers the area of roaming more reliably; it could hence be expected that a management system using satellite communications into Africa will achieve higher levels of uptime compared to a system that relies on GSM communication only.

The rest of the paper is organized as follows: section II provides a description of the research objectives; section III analyses GSM and section IV satellite communication costs; section V describes a practical case study that compares both GSM and satellite communication costs with the likely benefits from improved logistics operations; section VI concludes with a summary and a description of future work in this field.

## II. Research objectives

The following objectives were identified for the research work reported in this paper:

1) Identification of potential cost savings to be achieved by use of real time communications in cross-border transport.
2) Determining specific communication needs for optimal management of such operations (how often to communicate and how much data to communicate at a time).
3) Identification of communication options and the pros and cons of each approach (GSM roaming vs satellite).
4) Calculation of realistic costs (capital outlay and running costs) for each communication option.
5) Performing an estimated cost-benefit analysis for each communication option and calculating the internal rate of return (IRR) and net present value (NPV) of a telematics system.
6) Making recommendations about which option will be superior under which conditions.

## III. GSM cost

The GSM data cost analysis is completed by making some key assumptions regarding the nature of GSM services. These assumptions are made to allow a generic cost benefit analysis without going into the details of the features of the specific tracking units (or equivalent systems) that may be installed. The assumptions are as follows:

- The GSM service providers analysed operate on a pre-paid basis given that data bundles cannot be used to roam internationally. This will minimise the risk of unexpected high billing costs when roaming takes place.
- It is assumed that the tracking device can automatically switch to another network of choice without human intervention.
- When another network selection takes place, it is assumed that the cheapest service provider partner is selected. This will ensure the best case scenario is considered for the study.

The analysis is completed for multiple service providers in Southern Africa, given that the analysis is done for vehicles travelling from South Africa into Sub Saharan Africa. These providers include Vodacom, MTN, Cell C, and Telkom mobile. The data cost vary for each service provider and for each country. The specific destination countries to which vehicle travel will therefore be of utmost importance.


Fig. 1. Freight corridors in Sub Saharan Africa[6]

### A. Destination countries

Figure 1 above depicts the geographical representation of the freight corridors in Sub Saharan Africa. The destination countries included in this study are:

- South Africa
- Namibia
- Mozambique
- Zimbabwe
- Botswana
- Zambia
- Malawi
- Democratic Republic of Congo

### B. Data packets

The data packet size of each transmission will also greatly affect the data roaming costs. The standard packet of data will include the required information with a GSM header. A typical GPS coordinate data packet will vary between 60 to 70 bytes with a header file of 300 bytes making a total of 370 bytes for one information packet transmission. This will however differ for each service provider as they have a minimum data size that is billed while data roaming. Currently some of the tracking

units can string multiple different data types by making use of one GSM header. This may typically be a GPS coordinate, temperature reading, fuel level, etc.

### C. Service provider data cost

The minimum roaming data packets sizes and cost vary significantly between different service providers. These costs will be tabulated for each individual service provider with the cheapest roaming cost taken into account for case study calculations. Vodacom, MTN and Cell C will offer data roaming on pre- paid packages. Telkom mobile on the other hand will only grant data roaming for contract SIM cards.

The minimum data packet size, which will greatly influence the monthly data costs, varies for each service provider. Vodacom has a 10 kB minimum data packet with the message cost as seen in Table I. MTN, Cell C and Telkom each has a minimum data packet of 25 kB. These data cost can be seen in Table 2, Table 3, and Table 4.

TABLE I
VODACOM DATA COST PER COUNTRY [7]

| Countries | Vodacom per MB data | Per message cost |
|---|---|---|
| **Namibia** | R 51.10 | R 0.50 |
| **Mozambique** | R 51.10 | R 0.50 |
| **Zimbabwe** | R 51.10 | R 0.50 |
| **Botswana** | R 51.10 | R 0.50 |
| **Zambia** | R 102.40 | R 1.00 |
| **Malawi** | R 102.40 | R 1.00 |

TABLE II
MTN DATA COST PER COUNTRY [8]

| Countries | MTN per MB data | Per message cost |
|---|---|---|
| **Namibia** | R 106.50 | R 2.60 |
| **Mozambique** | R 106.50 | R 2.60 |
| **Zimbabwe** | R 106.50 | R 2.60 |
| **Botswana** | R 106.50 | R 2.60 |
| **Zambia** | R 106.50 | R 2.60 |
| **Malawi** | R 106.50 | R 2.60 |

TABLE III
CELL C DATA COST PER COUNTRY [9]

| Countries | Cell C per MB data | Per message cost |
|---|---|---|
| **Namibia** | R 21.38 | R 0.52 |
| **Mozambique** | R 104.69 | R 2.56 |
| **Zimbabwe** | R 156.04 | R 3.81 |
| **Botswana** | R 149.49 | R 3.65 |
| **Zambia** | R 20.40 | R 0.50 |
| **Malawi** | R 83.58 | R 2.04 |

The cost per message is calculated by multiplying the cost per Mb with the size of each data packet. Vodacom and Cell C offer the cheapest data cost of R0.50 per message, while MTN and Telkom mobile has a message cost of R2.60 and R2.94 respectively.

TABLE IV
TELKOM MOBILE DATA COST PER COUNTRY [10]

| Countries | Telkom Mobile per MB data | Per message cost |
|---|---|---|
| **Namibia** | R 120.27 | R 2.94 |
| **Mozambique** | R 120.27 | R 2.94 |
| **Zimbabwe** | R 120.27 | R 2.94 |
| **Botswana** | R 120.27 | R 2.94 |
| **Zambia** | R 120.27 | R 2.94 |
| **Malawi** | R 120.27 | R 2.94 |

## IV. SATELLITE COST

The costs for satellite tracking services will be analysed in the same manner as for the GSM data costs. These costs will be considered for services provided by the two main satellite communication tracking services: Iridium and Inmarsat.

These service providers support communication via satellite for most of the African continent. A redundant GSM communication module is usually in place in a tracking unit for lower cost communication when travelling within South Africa; the unit will automatically switch over to satellite communications once cross-border roaming takes place.

The main difference between the service providers are the satellite constellation configurations that they use. These constellations are either geostationary or in a geosynchronous orbit and this difference in constellation will practically change the uptime of communication. A geostationary orbit will ensure high uptime of the satellite, as it will appear that the satellite is always at the same location [11]. A satellite using a geosynchronous orbit will vary its location as it orbits the earth and may practically have communication loss for several hours with the tracking unit.

### A. Service providers

*1) Iridium:* Iridium makes use of geosynchronous orbiting satellites [11]. This will cause brief delays in communication with the satellites travelling in their orbits. Practically this can result in a communication delay of between a few minutes to about two hours. Table 5 below shows a standard monthly cost for Iridium service. The minimum data bundle used is 500 messages at a cost of R377 per month.

TABLE V
IRIDIUM DATA COSTS

| Description | Cost (USD) | Cost (ZAR) |
|---|---|---|
| 500 messages (bundle) | $ 35.71 | R 376.38 |
| Local sim (10 MB data) | | R 20.00 |
| 17 messages (out of bundle) | $ 1.07 | R 11.28 |
| | | |
| Cost per message (in bundle) | | R 0.75 |
| Cost per message (out of bundle) | | R 0.66 |
| | | |
| USD/Zar exchange | R 10.54 | 2014-04-22 |

*2) Inmarsat:* Inmarsat makes use of geostationary orbiting satellites [12]. This will ensure a higher communication uptime during tracking but at a higher cost than Iridium. The 500

message bundle will cost R625 per month as stated in Table 6 below.

TABLE VI
INMARSAT DATA COST

| **Description** | |
|---|---|
| 500 messages (bundle) | R 625.00 |
| 10 MB data | (included) |
| | |
| cost per message (in bundle) | R 1.25 |
| cost per message (out of bundle) | R 1.60 |

## V. CASE STUDY

A typical trip across national borders is from Johannesburg (South Africa) to Lusaka (Zambia). The route is shown in Figure 2 and Figure 3. The total distance travelled one way is 1780 km with 457 km in South Africa, 754 km in Botswana, and 569 km in Zambia. Inefficiencies in cross-border trips typically cause 50% of the travel time being lost during a 14 day trip [4].

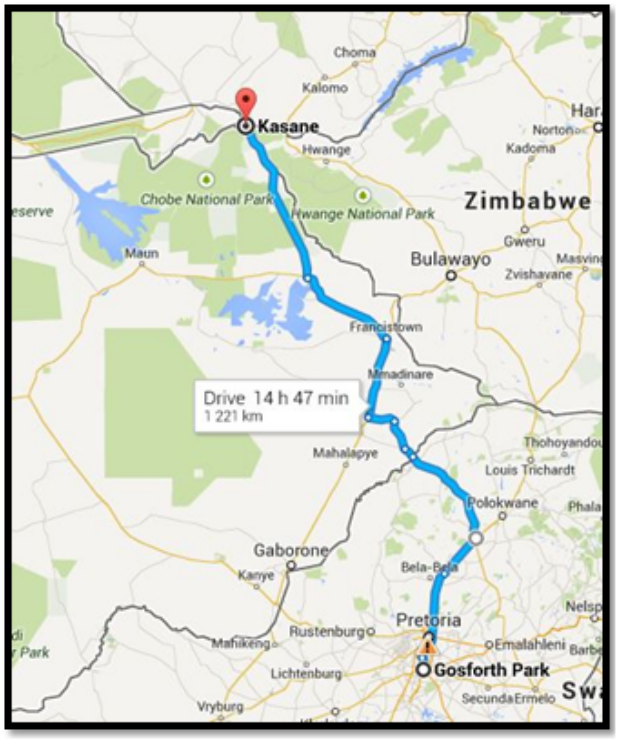



Fig. 2. Case study route part 1 of 2

As stated before various delays during the course of a trip (including border crossing delays, weigh station delays and delays at the destination) can be minimised by making use of regular data transmissions. With lower turn around time, a large monthly profit can be expected due higher utilisation of a trucks trips. If correctly implemented this will ensure a lower turnaround time per trip with an estimated increased

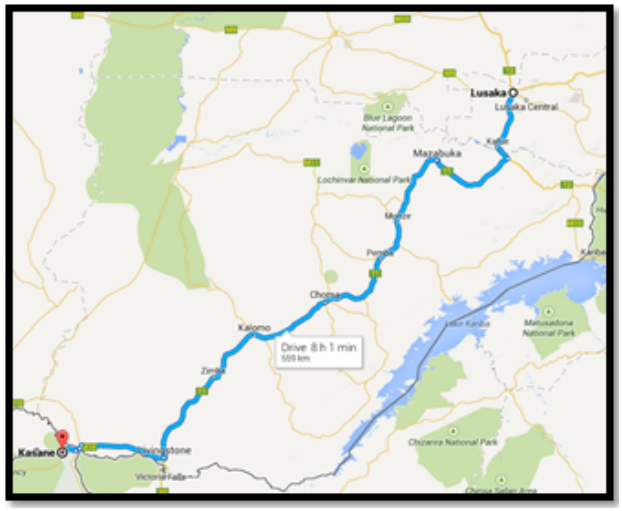



Fig. 3. Case study route part 2 of 2

revenue per month on a truck of at least 10% as previous work stated by Hoffman [4]. For the purpose of the case study we consider 5 scenarios with reductions in turnaround time ranging from 20% (optimistic scenario) to 2.5% (pessimistic scenario). An additional benefit is the increase in exports and imports from South Africa to Sub Saharan Africa [13]; this benefit is however not included in the case study as it is difficult to quantify accurately.

As GSM communication tends to be unreliable in many parts of the region we furthermore use a probability of 50% that GSM communications will be successfully established each time that the device tries to communicate with its home base the expected improvement in turnaround time is moderated by this figure to provide for non-ideal implementation of the relevant logistics management concepts due to non-ideal communications. For satellite systems a figure of 90% is used as the satellite networks cover the subcontinent more effectively compared to cellular networks.

A GSM interval of 5 minutes was chosen for the study. As the satellite communication uptime results in some delay in communication an interval of 10 minutes is used for satellite communication. By combining this with the average cost per message and assuming that a truck is on average travelling 12 hours per day the total messaging cost for each communication option could be calculated.

We assume that the telematics service will be offered at a rate of R300/truck/month, over and above the direct communication cost, that the cost to install the equipment on a truck is R10,000 and that maintenance costs amount to 20% of capital outlay per annum.

The revenue generated by a truck per annum is calculated by using an average trip length of 14 days with R70,000 revenue per trip to arrive at an annual revenue per truck (without using telematics) of R1.46 million. The gross profit margin is taken as 30%. Using the percentage improvements for each of the 5 scenarios we can calculate the increase in gross profits;

TABLE VII
VODACOM IRR AND NPV PROJECT ANALYSIS

| **Potential % decrease in turnaround time with ideal comms** | **20.0%** | **15.0%** | **10.0%** | **5.0%** | **2.5%** |
|---|---|---|---|---|---|
| **Initial capital cost** | R -10 000.00 | R -10 000.00 | R -10 000.00 | R -10 000.00 | R -10 000.00 |
| **Year 1 potential profit increase (Vodacom)** | R 31 575.20 | R 18 422.04 | R 5 961.16 | R -5 860.70 | R -11 547.16 |
| **Year 2 potential profit increase (Vodacom)** | R 31 575.20 | R 18 422.04 | R 5 961.16 | R -5 860.70 | R -11 547.16 |
| **Year 3 potential profit increase (Vodacom)** | R 31 575.20 | R 18 422.04 | R 5 961.16 | R -5 860.70 | R -11 547.16 |
| **Year 4 potential profit increase (Vodacom)** | R 31 575.20 | R 18 422.04 | R 5 961.16 | R -5 860.70 | R -11 547.16 |
| **Year 5 potential profit increase (Vodacom)** | R 31 575.20 | R 18 422.04 | R 5 961.16 | R -5 860.70 | R -11 547.16 |
| **IRR (Vodacom)** | 315% | 183% | 52% | 0% | 0% |
| **Required rate of return** | 25% | 25% | 25% | 25% | 25% |
| **NPV (Vodacom)** | R 59 931.64 | R 31 633.63 | R 4 824.99 | R -20 608.85 | R -32 842.85 |

TABLE VIII
IRIDIUN IRR AND NPV PROJECT ANALYSIS

| **Potential % decrease in turnaround time with ideal comms** | **20.0%** | **15.0%** | **10.0%** | **5.0%** | **2.5%** |
|---|---|---|---|---|---|
| **Initial capital cost** | R -10 000.00 | R -10 000.00 | R -10 000.00 | R -10 000.00 | R -10 000.00 |
| **Year 1 potential profit increase (Iridium)** | R 81 633.26 | R 53 845.30 | R 28 805.60 | R 6 125.66 | R -4 431.24 |
| **Year 2 potential profit increase (Iridium)** | R 81 633.26 | R 53 845.30 | R 28 805.60 | R 6 125.66 | R -4 431.24 |
| **Year 3 potential profit increase (Iridium)** | R 81 633.26 | R 53 845.30 | R 28 805.60 | R 6 125.66 | R -4 431.24 |
| **Year 4 potential profit increase (Iridium)** | R 81 633.26 | R 53 845.30 | R 28 805.60 | R 6 125.66 | R -4 431.24 |
| **Year 5 potential profit increase (Iridium)** | R 81 633.26 | R 53 845.30 | R 28 805.60 | R 6 125.66 | R -4 431.24 |
| **IRR (Iridium)** | 816% | 538% | 288% | 54% | 0% |
| **Required rate of return** | 25% | 25% | 25% | 25% | 25% |
| **NPV (Iridium)** | R 167 627.75 | R 107 844.06 | R 53 973.05 | R 5 178.89 | R -17 533.48 |

subtracting the overall communication and other system costs per truck per annum allows us to arrive at the change in net profits. The economics used for the calculation can be seen in table IX at the end of the paper.

The feasibility of a project can be determined by making use of internal rate of return (IRR) and the net present value (NPV) of the project over the intended lifetime. These calculations will take into account the initial capital expense and the net annual change in profits. The calculation is performed over the five (5) year life cycle of a truck. A 25% required rate of return is used for a project to be viable for implementation. Inflation will not be considered into the calculations in this article. The IRR and NPV were calculated for each communication options and for each improvement scenario; the best potential options for GSM and satellite will be displayed in tables VII and VIII.

Amongst the cellular services the most attractive figures were obtained for Vodacom and Cell C. Assuming a 20% decrease in truck turnaround time a telematics system using the Vodacom GSM option can potentially have an IRR of 315% and a NPV per truck of just below R60,000. As Cell C currently matches the data cost of Vodacom it will have identical IRR and NPV figures. The GSM communication will still be viable if a 10% decrease in turnaround is realised: A 10% decrease will ensure a IRR of 52% and a NPV just below R5,000. The other GSM operators will not achieve a positive NPV for any of the scenarios using their respective packet costs.

The Iridium satellite communication appear to be more attractive to implement for this application compared to the Inmarsat. A 20% decrease in turnaround time will have a IRR of 816% with a NPV of about R167,000 per truck. The implementation of the Iridium satellite communication will still be viable if the turnaround time can be decreased by just 5%. This will result in a IRR of 54% and a NPV of just above R5,000, whereas the Inmarsat option will result in a negative NPV for this scenario. An additional scenario was considered, but not displayed due to results, for the average GSM cost across the different countries which resulted in a IRR of 0% and a negative NPV.

## VI. CONCLUSION

The research has shown that a shift in focus from vehicle security to improved logistics management can significantly improve the turnover and profitability of a cross-border freight operation. Additional costs incurred from implementing a telematics system aimed at a reduction in truck turnaround times can be cost effective when considering the expected profit increases resulting from a reduction in inefficiencies.

The current calculation show that a telematics system using GSM communication based on the Vodacom or Cell C pricing options can produce an IRR of 315% and NPV per truck of about R60,000 if a 20% decrease in travel times can be achieved. Using Iridium satellite communication can potentially result in an IRR of 816% and NPV per truck of R167,000 for the same decrease in truck turnaround times. The Iridium satellite communication implementation will have a lower risk of implementation as it will still be viable if the turnaround time is only decreased by 5%. Implementation of either system can potentially in- crease gross profits of between R240,000 and R400,000 over the lifetime of a vehicle for the 20% improvement scenario. For a small cross-border fleet that includes typically about 25 vehicles this will ensure an increased net profit of between R750,000 and R1,500,000 per annum.

Future work will be conducted to confirm the communication cost across border by performing practical in field experiments on cross-border vehicles. Currently 6 tracking

units are being tested for this purpose, using a combination of GSM and satellite communication. The tests will also determine what the actual communication interval should be for GSM and satellite. Another aspect to be investigated in more detail is the actual efficiency improvements that will be achieved by a telematics systems that uses effective cross-border communications.

**Arno de Coning** received his undergraduate engineering degree in 2010 from the North-West University Potchefstroom Campus and his Masters of engineering degree in 2013. He is presently studying towards his PhD degree at the same institution. His research interests include socio economic impact, freight logistics, and operational improvement. **Alwyn J. Hoffman** received his B. Eng. degree in 1985, his M. Eng. degree in 1987 and his Ph.D. degree in 1991 from the University of Pretoria in South Africa. He has been professor in the School of Electrical, Electronic and Computer Engineering at North-West University since 1994. His current research interests include artificial intelligence, RFID and business intelligence.

## REFERENCES

[1] S. Wadhwa and A. Saxena, "Decision knowledge sharing: flexible supply chains in km context," *Production Planning & Control*, vol. 18, no. 5, pp. 436–452, 2007.
[2] F. Bodendorf and R. Zimmermann, "Proactive supply-chain event management with agent technology," *International Journal of Electronic Commerce*, vol. 9, no. 4, pp. 58–89, 2003.
[3] D. Thomas, "Expanding infrastructure: the its option." South African Transport Conference, 2001.
[4] A. J. Hoffman, K. Lusanga, and E. Bhero, "A combined gps/rfid system for improved cross-border management of freight consignments," 2013.
[5] U. N. Economic, S. C. for Asia, and the Pacific, "Secure cross border transport model," Tech. Rep., 2012.
[6] M. Fitzmaurice, "Ns corridor routes and costing," Compiled for TradeMark Southern Africa (TMSA), Tech. Rep., December 2013, compiled for TradeMark Southern Africa (TMSA).
[7] Data roaming. [Online]. Available: http://www.vodacom.co.za/personal/services/ roamingproductsandservices/dataroaming
[8] Using mtn when overseas. [Online]. Available: https://www.mtn.co.za/everyday˙services/RoamingInternational/Pages/UsingPhonesAbroad.aspx
[9] International roaming. [Online]. Available: http://www.cellc.co.za/international-roaming
[10] International roaming. [Online]. Available: http://www.telkommobile.co.za/coverage/ internationalroaming/l/116
[11] L. K. Wee and G. H. Goh, "A geostationary earth orbit satellite model using easy java simulation," *Physics Education*, vol. 48, no. 1, p. 72, 2013.
[12] Satellite constellation. [Online]. Available: http://www.iridium.com/About/IridiumGlobalNetwork/ SatelliteConstellation.aspx
[13] Our satellites. [Online]. Available: http://www.inmarsat.com/about-us/our-satellites/

TABLE IX
ECONOMIC OF TRIP MANAGEMENT

| | |
|---|---|
| **Length of trip (days)** | 14 |
| **% time travelling** | 50.00% |
| **Travel time (days)** | 7 |
| **Transmit period GSM (min)** | 5 |
| **Transmit period Satellite (min)** | 10 |
| **Total no of transmissions per month** | 1008 |
| **Cost per transmission Vodacom** | R 0.50 |
| **Cost per transmission Iridium** | R 0.75 |
| **Total vehicle turnover per trip** | R 70 000.00 |
| **Potential % decrease in turnaround time** | 10% |
| **Potential increase in revenue per trip** | R 7 000.00 |
| **Cost to install tracking equipment (once off)** | R 10 000.00 |
| **Life expectancy of equipment (years)** | 5 |
| **Fraction of year truck active** | 0.8 |
| **Number of trips per annum** | 20.9 |
| **Number of trips over lifetime** | 104.3 |
| **Total comms cost over lifetime (Vodacom)** | R 52 457.34 |
| **Total comms cost over lifetime (Iridium)** | R 39 565.42 |
| **Maintenance cost p.a. as fraction of capital** | 20% |
| **Total maintenance cost over lifetime** | R 10 000.00 |
| **Total operational cost of system per truck over lifetime (Vodacom)** | R 72 457.34 |
| **Total operational cost of system per truck over lifetime (Iridium)** | R 59 565.42 |
| **Increase in revenue over lifetime** | R 730 000.00 |
| **Gross Profit margin** | 30% |
| **Increase in gross profits** | R 219 000.00 |

# CoBI: A Collective Biosignal-Based Identification Model

Dustin van der Haar and Sebastiaan von Solms
Academy of Computer Science and Software Engineering
University of Johannesburg, Auckland Park Campus, South Africa
Tel: +27 11 559 3657, Fax: +27 11 559 2138
email: {dvanderhaar, basievs}@uj.ac.za

**Abstract-The rise of new portable sensors that monitor physiological sensors in the human body has allowed quality of life and medical diagnostic applications to be taken directly to user, without the constraints of physical space or inconvenience. The potential of these sensors in the domain of authentication and identification is becoming more feasible each day and current research in these biometric systems shows a great deal of promise. The paper builds on the proliferation of these sensors and proposes an interoperable model called CoBI, which allows multi-factor identification to take place. The model provides a platform for any viable biological signal that can be used for the purposes of identification and authentication, by providing pluggable sensor components and converting them into a common format (a feature vector consisting of autoregressive (AR) coefficients). Once they are in a common format they can then be merged together to form a consolidated feature vector. This consolidated feature vector can then be persisted during enrolment or passed further classification in order to achieve matching. The results have shown that cardiac and neurological components (from an electrocardiogram (ECG) and electroencephalogram (EEG), respectively) can be consolidated using the CoBI model successfully. By utilising the correct model order during feature estimation for the cardiac and neurological components, along with the appropriate classifier for matching, the biometric system yields nominal results for authentication and identification.**

**Index Terms— Biometrics, Biological Signals, Authentication, Wearable Computing, Frameworks**

## I. INTRODUCTION

The age of mobility is upon us. Recent smart phones and devices have crept into our daily lives in the form of quality of life products and advanced medical diagnostics. The abundance of sensors in these devices has led to the proliferation of applications that leverage these sensors in unique ways, such as the measuring of physical activity [1] or for the detection of Parkinson's disease tremors [2]. However, one especially interesting application is the use of these sensors to identify or authenticate individuals for the task of access control. A sensor, such as a fingerprint scanner, captures a biological metric or biometric and by analyzing the unique sample every user provides, it is used to determine whether the user should gain access.

However, the use of biometric attributes for the task of identification and authentication is not new. They have been used for physical access control [3] and to confirm banking transactions [4] for over a decade. Fingerprint and face recognition have attained mainstream acceptance and continue to grow in various industries.

Esoteric biometric systems that utilize cutting edge sensors that capture user attributes, such as heart rate or brain waves, are also seeing developments. The methods for capturing, processing and interpreting these biological signals or biosignals have also improved, making them an attractive option in biometric systems. They have grown from utilizing large immovable sensors that have limited mobility to using sensors that are highly portable and that can be worn by the user.

The abundance of these sensors is a welcome change in the biometric community, but with it comes the problem of interoperability. The rich varieties of biological sensors or biosensors do not conform to biometric standards and require a separate implementation for each sensor. In addition to this problem, the specialised nature of these formed biometric systems make a change of hardware or processing methods difficult. More so, if a multi-factor or multi-modal system should be required, the lack of interoperable components would result in a significant increase in resource costs.

The paper attempts to address this problem, by providing a model that can deal with multiple sensors, deriving an interoperable format and using pluggable processing components to achieve identification and authentication (based on previous work done in [5]). It begins by providing a discussion on the problem domain and the related methods in current work, followed by a discussion on how biosignal consolidation can be achieved for multi-modal systems. The model is then discussed and the results presented, along with a critical analysis and support. The paper then ends with a conclusion.

## II. PROBLEM BACKGROUND

The use of token or knowledge-based authenticators in access control is slowly diminishing as limitations, such as theft or cracking become more of a concern. Although approaches such as the use of location-based authenticators [6] or more encryption attempt to resolve these limitations, they are still subject to problems, such as an indirect relationship with the user and reverse engineering, as seen in

[7]. An authenticator that has a stronger relationship with the user and that is more difficult to steal needs to be used.

The use of biometric attributes for a user, such as their fingerprints or face, which encapsulate the user, has shown to be a good alternative authenticator in access control systems [3]. Unfortunately as they become more popular, the limitations of these attributes (such as biometric spoofing [8]) become a more relevant threat. Alternative biometric attributes are currently being researched to address these and other threats. One promising category of biometric attributes currently being researched is that of biosignal-based biometric systems. These biometric attributes are discussed in the next subsection.

### A. *Biosignal-Based Biometric Systems*

Human locomotion, respiration and cardiac rhythm have one thing in common. They require signals that transmit chemical or electric-based messages in order to coordinate physiological systems in the human body. These signals are also known as biological signals or biosignals. In the medical field these biosignals are monitored to aid patient diagnostics. Medical practitioners use sensors, such as an electrocardiogram (ECG) or electroencephalogram (EEG), which provide electric-based representations of the recorded signals found in the heart and brain, respectively.

Due to the unique physiology each human has, these biosignals are also unique and serve as a good biometric attribute. Capturing these unique attributes, removing any present noise, extracting features and classifying them can be used for the task of identification and authentication.

In EEG-based biometric systems, there are a variety of approaches that can be applied in order to achieve identification. Poulos et al. achieves identification by isolating the alpha band (8-12 Hz), estimating autoregressive (AR) coefficients and classifies them using learning vector quantization (LVQ) neural networks [9]. However, this approach requires a significant amount of training time and resources for codebook vector generation, along with their subsequent updates. Palaniappan and Mandic achieve identification by applying different methods to EEG signals [10]. Selected channels of the EEG are preprocessed by mapping the signal to a smaller range and applying an Elliptic filter, followed by the energy being calculated and then used as features. An Elman neural network is then trained with resilient backpropagation (RB) using the extracted energy features. However, their approach also incurs a great deal of training time. Marcel and Millán take a more probabilistic approach to identification using EEG signals [11]. Preprocessing is achieved using a surface Laplacian (SL); power spectral density-based features are derived and fed into a Gaussian Mixture Model (GMM). However, their approach is not stable over time and results in high intra-class variability, which increases the false rejection rate (FRR).

In ECG-based biometric systems the methods used in them fall under two types of approaches: fiducial and non-fiducial-based approaches [12]. As part of preprocessing, fidicual methods search for parts of the sinus rhythm waveform, which represent the depolarisation and repolarisation of the human heart, using a technique such as [13]. Non-fiducial methods do not search for these waveforms and the ECG signal is passed straight to feature extraction. Biel searches for the complete waveform and derives features from parts of the waveform such as P duration and QRS duration and classifies them using soft independent modeling of class analogy (SIMCA) and principal component analysis (PCA) [14]. However, this approach includes redundant features, which cause unnecessary computational overhead. Singh perform QRS fiducial detection, followed by the extraction of simple features such as amplitude and angle from the found fiducials [15]. Bayesian decision theory is then used to achieve authentication. However, these features are not stable enough over time and also increase intra-class variability. Balli shows that non-fiducial methods that utilising spatial methods such as principal component analysis (PCA) or linear discriminant analysis (LDA) can be used directly to achieve good results [16].

However, with this variety of methods comes the problem of interoperability, where modules cannot be interchanged among systems for scalability or reuse. Furthermore, by consolidating EEG and ECG components to attain a multi-modal system, it becomes an even more difficult task. The next subsection discusses the approaches for potentially fusing these components.

### B. *Biosignal Consolidation*

When fusing multiple biometric attributes to form a multi-modal biometric system there are three approaches that can be used: feature level fusion, match level fusion and decision level fusion [17]. Each of these fusion approaches depend on what stage of the authentication process fusion takes place, where feature level is the earliest and decision level takes place in the latest stage.

Greene et al. proposes fusion approaches at a feature level and matcher level for seizure detection [18]. Although they propose that later fusion methods are better for real world purposes, it comes at the cost of performance. Bandeira proposes a feature and decision level fusion approach by using self-organising maps (SOM) to help with sport shooting accuracy [19]. However, their approach comes at the cost of training time and interoperability issues at a feature level. A fusion scheme needs to be found that has improved performance, but that does not come at the cost of interoperability in the biosignal system. The next section attempts to resolve these issues by proposing a model called the CoBI model.

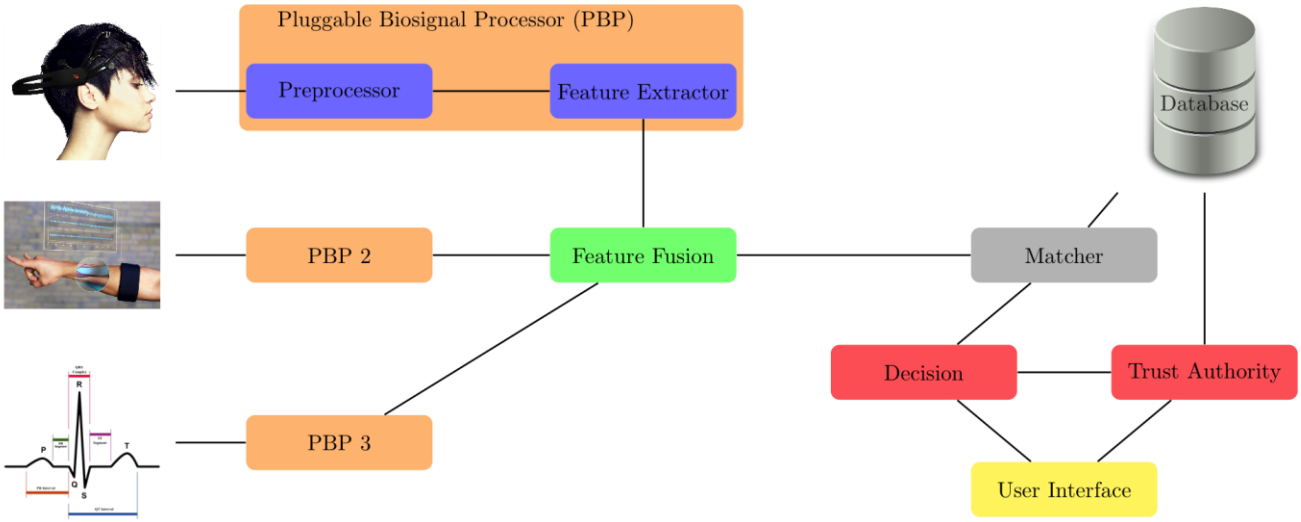



**Figure 1: The Collective Biosignal-Based Biometric Identification (CoBI) Model**

## III. Model

In order to address the issues of interoperability and biosignal consolidation in multiple biosignal-based biometric systems, the authors have proposed a model called CoBI or the collective biosignal-based identification model. The approach is cognizant of the multiple biosignal sources, which can be used for the task of identification and integrates methods for the task of multi-modal identification. The structure of the model can be seen in Figure 1 and is discussed in the next subsections.

### *A. Pluggable Biosignal Processor*

In order to accommodate multiple biosignal sources, the CoBI model allows multiple sensors to be integrated into the system by coupling it with a component the author calls a pluggable biosignal processor or PGP. For a sensor to be compatible with the CoBI model, it simply needs an implementation of an appropriate PGP component.

The PGP takes a captured biosignal, performs preprocessing on the biosignal and estimates autoregressive (AR) coefficients, which are passed further in the system. A typical preprocessor module removes noise or artifacts by applying a bandpass or Butterworth filter that restricts the frequency band that features are derived from. In ECG signals this range is 5-15Hz and in EEG signals it is typically restricted to the Gamma band (30-50 Hz). For each preprocessed signal, the next module (the feature extractor module) then estimates common features: AR coefficients. These coefficients are such that:

$$x[n] = -\sum_{k=0}^{p} a_k \, x[k-1] + \, e[n]$$

where $x[n]$ represents the sampled point $n$, $p$ the model order, $a_k$ the real-valued AR coefficients (which will be used as features) and $e[n]$ the error term that is independent of past samples. The model order can be fixed or derived each time using the Akaike Information Criterion (AIC) method. However, the author has determined that a fixed model order (such as 6 for ECG and 12 for EEG) per biosignal component is more computationally efficient. The coefficients can then be successfully derived using a variation of the Burg method [20]. Finally the estimated AR coefficients can then be passed to the next component for fusion.

### *B. Feature Fusion*

The feature fusion component takes the AR coefficients derived for each biosignal source and consolidates or fuses them into one feature vector. Due to the common AR coefficients derived for each biosignal source, fusion is a simple concatenation operation (with optional redundancy filtering). The fused feature vector then becomes the biometric attribute that will be interpreted by subsequent components. The feature fusion component then sends the whole feature vector to the matcher component.

### *C. Matcher*

Once the matcher component receives the feature vector, it compares it with persisted forms in the database, which were enrolled by authorized users. A classifier (such as k nearest neighbor) is trained with the persisted feature vectors retrieved from the database (one identity for the case of authentication and many for the case of identification). The currently captured feature vector can then be tested against the trained classifier to gain a confidence measure. The confidence measure is then sent to the decision component for further interpretation.

### *D. Decision, Trust Authority and User Interface*

The decision component then takes the confidence measure and determines whether it is a match or not and sends the outcome to the user interface, which executes the outcome. The trust authority component is responsible for the enrolment of authorized identities and is responsible for regulating the decision component's outcome. If the outcome is a match, the user interface retrieves the extent of access through the trust authority and issues it to the user. If the outcome is a non-match, the user is denied access and the trust authority is informed for subsequent stricter matching.

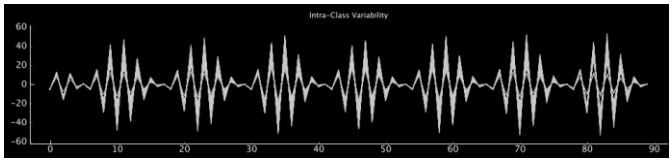



**Figure 2: An intra-class variability plot for multiple biometric attribute presentations for one user.**

## IV. Results

In order to validate the effectiveness of the CoBI model, two of the most common biosignals were selected (EEG and ECG) and a multi-modal biometric system was implemented using the CoBI model. The control subjects in the CAP sleep database [21] were used as a secondary data set to validate the model. The results of this evaluation are presented in the next subsection.

### A. Performance Benchmark

By implementing the components in the CoBI model with the suggested parameters for each of the modules (such as an AR model order of 6 for ECG and 12 for EEG, along with a K nearest neighbor classifier), the performance of the system can be gauged. As seen in Figure 2, the consolidated feature vector containing all the estimate AR features, remains stable across multiple presentations. Furthermore, the derived feature vector is sufficiently unique to differentiate users from one another, as seen in the receiver operating characteristic (ROC) curve seen in figure 3. The implementation also achieved a false rejection rate (FRR) of 6.85% and false acceptance rate (FAR) of 2.55% making error sufficiently low enough to be a viable alternative biometric attribute. However, as seen in the next subsection, a critical analysis reveals there are considerations that can be applied to improve the implementation.

### B. Critique

Although the implemented model provides a viable approach for achieving multi-modal identification and authentication, it is not without its limitations. There are implementation considerations that can be applied to improve future deployments of the CoBI model.

Upon analyzing the resource contribution of the various components in the implemented CoBI model, one component stood out, the PBP. Further inspection revealed that the AR estimation process found in the PBP consumed relatively more resources than the other components. The cost of the biosignal diversity comes at the cost of resources utilization. However, one approach that may help lower resource utilization would be to derive common power spectrum density (PSD) in a similar approach to [11].

When utilizing a low amount of biosignal sources, a k-nearest neighbor classifier may be adequate, but when the biosignal sources are substantially more, a different classifier may be required. At a higher end, the resources allocated for the classifier may be better spent on another method such as a LVQ neural network-based approach similar to [9].

### C. Support

There are also benefits of the CoBI implementation that

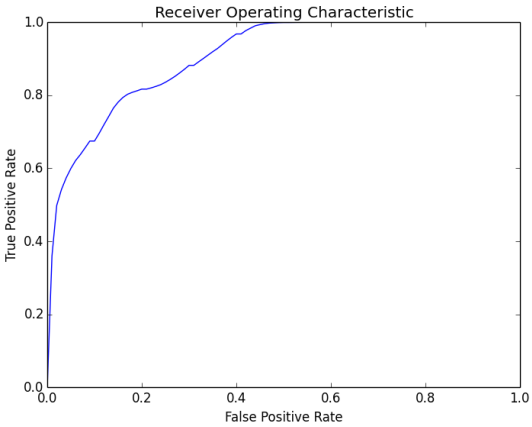



**Figure 3: The receiver operating characteristic (ROC) curve, which maps the false positive rate against the true positive rate.**

justify the use of the model for multi-modal identification. The resource utilization and timing for the matcher component is quite efficient when compared to other multilayer perceptron (MLP) neural network-based approaches. The reduced training time also lowers the overall computational footprint of the system, thereby lowering resource requirements for the system.

The pluggable nature and familiarity of the various components in the system make the system more flexible and scalable. The components in the system can be effectively being replaced with newer or more efficient approaches in order to deal with varying populations and demographics. The common derived features also make it possible for inter-vendor compatibility; thereby further increasing interoperability for future developments.

Another interesting property certain biosignals hold is that minor variability found within AR features of a biosignal (from the baseline signal) can be used to form gesture sequences. A physical or mental task manifests itself as minor variability from the baseline biosignal. These differences can then be used to form cognitive gestures similar to [22]. A collection or sequence of these gestures can then form a cognitive signature or password that can be revoked and replaced with another signature. By persisting labelled consolidated feature vectors found in the CoBI, cognitive verification schemes can be realized.

## V. Conclusions

The abundance of biosensors in modern society is a welcome addition to research that achieves integrated applications for fitness or health purposes, but their potential in security is growing. The same sensors can be used within a biometric context for the purposes of identification and authentication of users. However, the variety of these sensors should not be a deciding factor when it comes to implementing biometric systems that leverage these sensors.

The paper introduced the CoBI model that attempts to resolve this interoperability problem by taking in multiple biosignal sources and deriving common features for each of them, through the use of pluggable components. These

common features (AR coefficients) can then be consolidated and classified according to persisted feature vectors to gain a confidence measure. The confidence measure can then be used to gain an outcome for identification or authentication.

The results have shown that the proposed approach yields nominal results and is a viable alternative biometric attribute. The approach portrays low intra-class variability and a sufficiently high inter-class variability. Upon performing further benchmarks, the system is shown to yield a FRR of 6.85% and a FAR of 2.55%, which shows that it is low enough for the task if identification and authentication.

Although the feature estimation incurs a minor resource cost, the benefits gained outweigh the cost. By deriving common AR coefficients, it may resolve interoperability issues, but it comes at the cost of additional resources. The pluggable nature of the components in CoBI makes it flexible and scalable, which make it possible to accommodate changing environments and populations.

Lastly, the structure of CoBI also makes it possible to cater for future cognitive biometric verification schemes. The common derivations of AR coefficients make it possible to encapsulate the manifestation of physical or mental gestures. These cognitive-based identification schemes and biosignal-independent systems may bring us one step closer to more secure access control mechanisms.

## REFERENCES


[1] R. Troiano, D. Berrigan, K. Dodd, L. Mâsse, T. Tilbert and M. McDowell, "Physical activity in the United States measured by accelerometer," *Medicine and science in sports and exercise*, vol. 40, p. 181, 2008.

[2] R. LeMoyne, T. Mastroianni, M. Cozza, C. Coroian and W. Grundfest, "Implementation of an iPhone for characterizing Parkinson's disease tremor through a wireless accelerometer application," *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pp. 4954-4958, Aug. 2010.

[3] A. Cavoukian, Consumer biometric applications: a discussion paper, Information and Privacy Commissioner/Ontario, 1999.

[4] L. Coventry, A. De Angeli and G. Johnson, "Usability and biometric verification at the ATM interface," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2003.

[5] D. T. van der Haar, "Collective Human Biological Signal-Based Identification and Authentication in Access Control Environments," PhD Thesis, Academy of Computer Science and Software Engineering, University of Johannesburg, 2014.

[6] N. Sastry, U. Shankar and D. Wagner, "Secure verification of location claims," in *Proceedings of the 2nd ACM workshop on WIreless security*, 2003.

[7] K. Nohl, D. Evans, S. Starbug and H. Plotz, "Reverse-Engineering a Cryptographic RFID tag.," in *USENIX Security Symposium*, 2008.

[8] S. Schuckers, "Spoofing and anti-spoofing measures," *Information Security technical report*, vol. 7, no. 4, pp. 56-62, 2002.

[9] M. Poulos, M. Rangoussi, N. Alexandris and A. Evangelou, "Person identification from the EEG using nonlinear signal classification," *Methods of information in Medicine*, vol. 41, no. 1, pp. 64-75, 2002.

[10] R. Palaniappan and D. Mandic, "EEG Based Biometric Framework for Automatic Identity Verification," *The Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology*, vol. 49, no. 2, pp. 243-250, 2007.

[11] S. Marce and J. Millán, "Person authentication using brainwaves (EEG) and maximum a posteriori model adaptation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 4, pp. 743-752, 2007.

[12] F. Agrafioti, J. Gao and D. Hatzinakos, "Heart biometrics: theory, methods and applications," *Biomoetrics: Book*, vol. 3, pp. 199-216, 2011.

[13] A. Lourenço, H. Silva, P. Leite, R. Lourenc¸o and A. Fred, "Real Time Electrocardiogram segmentation for finger-based ECG Biomoetrics," in *BIOSIGNALS*, 2012.

[14] L. Biel, O. Pettersson, L. Philipson and P. Wide, "ECG Analysis: A New Approach in Human Identification," *Intrumentation and Measurement, IEEE TRansactions on*, vol. 50, no. 3, pp. 808-812, 2001.

[15] Y. Singh and S. Singh, "Evaluation of Electrocardiogram for Biometric Authentication," *Journal of Information Security*, vol. 3, no. 1, pp. 39-48, 2012.

[16] T. Balli and R. Palanoappan, "Classification of biological signals using linear and nonlinear features," *Physiological measurement*, vol. 31, no. 7, p. 903, 2010.

[17] A. Ross and A. Jain, "Multimodal biometrics: An overview," in *Proceedings of the 12th European Signal Processing Conference*, 2004.

[18] B. Greene, G. Boylan, R. Reilly, P. de Chazal and S. Connolly, "Combination of EEG and ECG for improved automatic neonatal seizure detection," *Clinical neurophysiology*, vol. 118, no. 6, pp. 1348-1359, 2007.

[19] N. Bandeira, V. Lobo and F. Moura-Pires, "EEG/ECG data fusion using Self-Organising Maps," *Proceedings of EuroFusion99*, 1999.

[20] P. Brockwell, R. Dahlhaus and A. Trindade, "Modified Burg algorithms for multivariate subset autoregression," *Statistica Sinica*, vol. 15, no. 1, pp. 197-213, 2005.

[21] M. Terzano, L. Parrino, A. Sherieri, R. Chervin, S. Chokroverty, C. Guilleminault, M. Hirshkowitz, M. Mahowald, H. Moldofsky and A. Rosa, "Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep," *Sleep Medicine*, vol. 2, no. 6, pp. 537-553, 2001.

[22] K. Revett, F. Deravi and K. Sirlantzis, "Biosignals for user authentication-towards cognitive biometrics?," *Emerging Security Technologies (EST), 2010 International Conference on*, pp. 71-76, 2010.

**Dustin van der Haar** is currently completing his PhD degree in Computer Science at the University of Johannesburg. He also lectures design patterns at an undergraduate level and biometrics at a postgraduate level in the Academy of Computer Science and Software Engineering at the University of Johannesburg. His research interests include biometrics, wearable computing and information security.

**Prof SH (Basie) von Solms** is a Research Professor in the Academy for Computer Science and Software Engineering at the University of Johannesburg in Johannesburg, South Africa. He is presently the Director of the ITU-UJ Centre for Cyber Security of the University of Johannesburg.

# Perishable Produce Temperature Profiling Using Intelligent Telematics

Christian C. Emenike and Alwyn J. Hoffman
Department of Electrical and Electronic Engineering
North West University, Private Bag X6001, Potchefstroom 2520
Tel: +27 18 2991963, Fax: +27 18 2991977
email: {23718528, Alwyn.Hoffman}@nwu.ac.za

**Abstract- Controlling, monitoring and maintaining the temperature of agricultural products in the cold chain is critical to avert the ills associated with this industry. The various segments of the cold chain not only require visibility to prompt the necessary decisions and actions in times of negative events but continuous amelioration of systems and technologies employed for safe, fresh and edible farm produce.**

**This article aims at providing an improved concept for temperature monitoring by addressing operational issues experienced in Southern Africa cold chains. It is demonstrated how detailed profiling of actual temperatures experienced by cargo can be implemented using a system based on a combination of local wireless communications between sensors and a controller, and long range communication between the controller and a central office via a GSM tracking unit.**

**Index Terms— perishable food, temperature monitoring, cold chain, GSM**

## I. INTRODUCTION

The cold chain involves the transportation of temperature sensitive products by means of refrigerated trucks, commonly called reefers, along a supply chain through thermally controlled and refrigerated packaging methods. The transportation of chilled and/ or agricultural products in reefer containers has grown to become a large and steadily growing business in Southern Africa and the world at large[1].

The South African fruit industry is a significant employment generator; it employs approximately 460 000 people who have two million dependents[2]. The industry accounts for 50% of all agricultural exports in South Africa [3], with an annual export value of approximately R12 billion [2]. Unfortunately, a huge amount of this profit and commodities are lost due to poor quality of these products before they reach their target destinations.

The internal biological and chemical process of fresh produce, such as respiration, continues after harvesting. This implies that the product absorbs oxygen and releases carbon dioxide and ethylene. This results in the liberation of heat energy. Lowering the temperature reduces the respiration and consequently the heat considerably, hence avoiding deterioration due to high concentrations that may be caused by these latent activities.

The delivery of these cargo types in good conditions from point of production to point of distribution or consumption has been an issue for all players (the growers or producers, the logistic service provider and other transport companies, and the final consumer) involved the supply chain. Efficient monitoring of the temperature of these cargoes at a reasonable cost is the cry-for-help of these stake holders. Approximately 35% of fruits and vegetables are lost in the cold chain [4], partly due to the lack of cost-effective alternatives.

Refrigeration is basically removing heat by evaporation. Farm produce in the cold chain are refrigerated for the sole purpose of prolonging their shelf life [5] , state and quality thereby avoiding cold chain ruptures. Maintaining the required transportation temperature and humidity is of key importance in actualizing this goal. The required temperature in cold chain mainly depends on the cargo type. Fresh fruits and vegetables are usually transported between 0°C to 8°C, Meat and cold chilled products at a temperature below -18°C, dairy products like margarine, butter usually between -8°C to 7°C, frozen foods and ice cream are usually transported at -24°C to -18°C while chocolate at -8°C to -18°C and pharmaceutical products usually between 2°C to 8°C.

## II. RESEARCH OBJECTIVES

The following research objectives have been defined:

- Firstly the extent of the cold chain management (CCM) problem for typical cold chain operations in Southern Africa must be quantified. As more detailed monitoring implies higher system costs it is necessary to determine what will constitute an optimal level of monitoring to prevent losses while keeping costs at a reasonable level. For this purpose data will be gathered from cargo owners, research groups and organizations, and service providers involved within the industry.
- Secondly a monitoring methodology must be designed to characterize cold chain operations with sufficient accuracy to pinpoint problem areas.
- Various configurations of off the shelf instrumentation will be used in monitoring temperatures during transportation, including standalone temperature sensors with built-in data loggers (that require manual downloading of data)

and RFID based temperature loggers (that can support wireless downloading of data).
Experiments will be set up by placing the data loggers in various configurations inside reefer containers (including on all the sides of the reefer container and the doors except the floor) as well as inside the cargo and by configuring them at the required transit temperature. Temperature loggings from the truck will also be collected.

- Using the above experimental setups data will be collected to characterize actual cold chain operations for a representative set of actual trips, including different types of cargo and trips to different destinations.
- Temperature data of the cargo will be collected from the different tiers of the container where the loggers are installed.
- Lastly an optimal approach will be designed to conduct cold chain monitoring on an on-going basis as part of standard operations, finding a balance between sufficient accuracy of monitoring and the cost of the monitoring system.
- This will result in providing cold chain monitoring solutions customized to the end-user's problems, needs and budget.

## III. LITERATURE STUDY

Temperature monitoring studies aimed at the needs of the cold supply chain industry have been conducted and are still been conducted around the globe. All perishable products have a finite lifespan and are in a state of decline from the moment of harvest. According to [6], temperature is the greatest determinant of fresh- produce deterioration rates and potential market life. Studies have shown changes of 5ºC or more occurring within perishable cargo which far exceeds the required accuracy to optimize the shelf-life of such products [7]. A study by the University of KwaZulu-Natal [8] found that the quality of avocados is severely affected by breaking the cold chain. The results of this study estimated that 80% of fruit stored at the appropriate temperature without a break in the cold chain maintained its quality, whilst only 31% – 60% of fruit that had a cold chain break achieved the same quality levels [9] . Deviations of only a few degrees have led to losses amounting to thousands of Rands in [7], [10], [11]. A recent study shows that refrigerated shipments rise above the optimum temperature in 30% of trips from the supplier to the distribution center, and in 15% of trips from the distribution center to the stores [7] . Roy *et al.* analysed the supply of fresh tomatoes in Japan and quantified product losses of 5% during transportation and distribution [7] .
Thermal variations during transoceanic shipments have also been studied [7], [12–14]. The results showed that there was a significant temperature variability both spatially across the width of the container as well as temporally along the trip, and that cargo temperatures were out of the specification more than 30% of the time.

These studies provide evident proof that there is a clear need for accurate monitoring of perishable cargo in transit in order to provide early warning when such cargo is approaching the thresholds of the approved temperature range for storage. Most of these studies however involved complicated and expensive experimental setups to allow the temperatures of cargo in transit to be accurately monitored. Given the nature of a typical logistics operation and continuous cost pressures to remain globally competitive the need therefore exists for a cost-effective methodology, firstly to characterize the temperature behaviour of various types of perishable cargo in transit and secondly to implement continuous monitoring as part of standard operations to ensure accurate management of cold supply chains.

Literature studies conducted on the state of the art in cold chains [10], [14–25] identified the following approaches to improve cold chain monitoring:

- Data loggers based monitoring,
- RFID based monitoring,
- Wireless sensor network monitoring,
- Real time web based monitoring,

Due to the manpower intensive nature of monitoring based on standalone data loggers the current trend is towards monitoring using wireless communications with the sensing device. This implies the need for suitable wireless communication protocols between the temperature sensor, a local controller and the office from where operations are controlled. The table below lists the communication protocols that may be employed in such a system.

| **Wireless Network Type** | **Implementation** |
|---|---|
| Wireless Wide Area Network ( WWAN) | Global System for Mobile-Communication (GSM)<br>Code Division Multiple Access (CDMA)<br>General Packet Radio Service (GPRS)<br>Universal Mobile Telecommunications System {UMTS(3G)} |
| Wireless Local Area Network (WLAN) | IEEE 802.11.x<br>HyperLAN<br>Home RF |
| Wireless Sensor Networks (WSN) | Bluetooth<br>ZigBee |
| RFID | Active RFID<br>Passive RFID<br>Battery-Assisted Passive (BAP) |

Table 1: Communication Protocols

Existing solutions for cold chain visibility are still limited due to the following:

- On-board instrumentation of reefer containers does not provide real time status data on status of cargo.
- In the case of temperature breach or abuse, on-board instrumentation possess no alert notification in other to salvage cargo in transit
- In cases where vehicle tracking systems do include some element of temperature sensing, the sensing data is usually not accessible once the consignment crosses a national border due to the high costs associated with GSM roaming.

- In case of cargo damage there is normally no record of exactly when or where it occurred and which party was responsible for the cargo at that point in time, making it difficult to enforce such aspects of service level agreements.

A need therefore exists for methods and technologies that will address the issues mentioned above in order to enable the more effective management of cold supply chains.

## IV. MATERIALS AND METHODS

Experiments were carried out to determine temperature profile distributions at different tiers in a reefer container. For these tests 10 data loggers were installed at different positions along the inside periphery of a 15.3 meters long refrigerated truck and 4 data loggers were placed inside boxes in between fruits.

The positions of the loggers within the trucks were as follows:

1. Left side of the reefer container – three (3) data loggers.
2. Right side of the reefer container – three (3) data loggers.
3. Roof top of container – three (3) data loggers.
4. Right Door – at the center (1) data logger.

The loggers were configured to collate a maximum of 8000 data points every 5minutes and to trigger alarm at temperatures below 0ºC and above 2ºC.

The trips of the truck were as follows:

1. 24 pallets ( 1200 boxes) of yeast to Zimbabwe at a set point temperature of 2ºC
2. 4 pallets of apples and mixed vegetables ( carrots, garlic and lettuce) to Lusaka , Zambia at a set point temperature of 2ºC
3. Mixed vegetables cargo ( carrots, garlic and lettuce) from Mozambique at a set point temperature of 2ºC

In a second field test, 6 pallets of apples and 17 pallets of potatoes and onions were transported to Zambia from South Africa via Botswana and back to South Africa at a set temperature of 2ºC. 10 temperature loggers including humidity loggers were used and recovered from the experiment. They positions are shown in table 2 below.

| 0.32m from vent | 1st tier 3.32m from vent | 2nd tier 6.32m from vent | 3rd tier 9.32 m from vent | 4th tier 12.32m from vent | 5th tier 15.32m from vent |
|---|---|---|---|---|---|
| T19 RTCT | T15 RTCT | T13 RTCT | HT 22 RTCT | T18 RTCT | HT 21 RTCT |
| | | T4 LHST | T3 LHST | T2 LHST | |
| | | | | T11 RHST | |

Table 2: Table 2: Loggers Positioning (RTCT: Roof top center of tier; LHST: left hand side on trailer; RHST: right hand side on trailer)

## V. EXPERIMENTAL RESULTS

The data collated for the first test route of South Africa to Zambia via Zimbabwe show temperature variations at different tiers in the container as reported in figure 1 with lowest temperature reading of 1.3ºC, highest temperature reading of 50.0ºC, an average reading ranging from 12.5ºC to 14.9ºC and a standard deviation ranging from 7.9ºC to 8.7 ºC for loggers installed in the 1st tier and door side of the trailer in the 5th tier. Average Temperature and humidity recordings of data loggers installed at the door and roof top on the side of the container were 11.8ºC to 13.2ºC at 74.0%RH to 76.3% RH. Also analysis of the graphs shows that sensors in the 1st tiers revealed low temperature reading while those in the 5th tier high temperature reading. Sensors placed in the same tier but at different sides exhibited temperature variations within the same range.

Figure 2 shows results from the second experiment for the South Africa – Zambia route via Botswana and back to South Africa. Lowest average temperature of 13.8º at 45.0 % RH and maximum average reading of 23.5ºC at 47.1% RH was realized from all sensors mounted in the trailer according to the set-up in table 2. Overall average deviations ranged from 5.8ºC to 9.3ºC. Analysis of the graph indicates same temperature variation for sensors at the roof tops at the center of each tiers. Upon retrieval, data loggers were on trigger mode which indicated deviations from set point temperatures.

These data depicts serious deviations from expected transportation set points of 2ºC. The deviations could be attributed to any of the following:

- Non-compliance to fresh produce transportation temperatures in the SADC region since it is in the habit of drivers to open trailers either to smuggle goods or any other reasons for personal gains.
- Trailer been powered off and on in other to save fuel which hence results in sharp rises in temperature.
- Temperature profile describing the data captured at the door side which will indicate opening of doors during transportation.
- Trailers transport dry cargoes on their way back from destination countries

### *A. On-going and future experiments*

Experiments are currently been carried out on apples, mixed fruits and strawberries in order to collect a representative data set relating different set points and destinations in the Southern African Development Community (SADC) region. In order to collect even more accurate temperature profiles in some cases 20 data loggers are installed in a trailer, 6 on each side of the containers (left side, right side, roof top), 3 meters apart and diagonally and 1 logger on each door.

Further set of experiments will also be conducted with RFID based temperature loggers to enable real time wireless downloading of data. This will allow a comparison of the accuracy and cost associated with each approached so that the best technology options can be recommended to be used for different cargo types, destinations and clients.

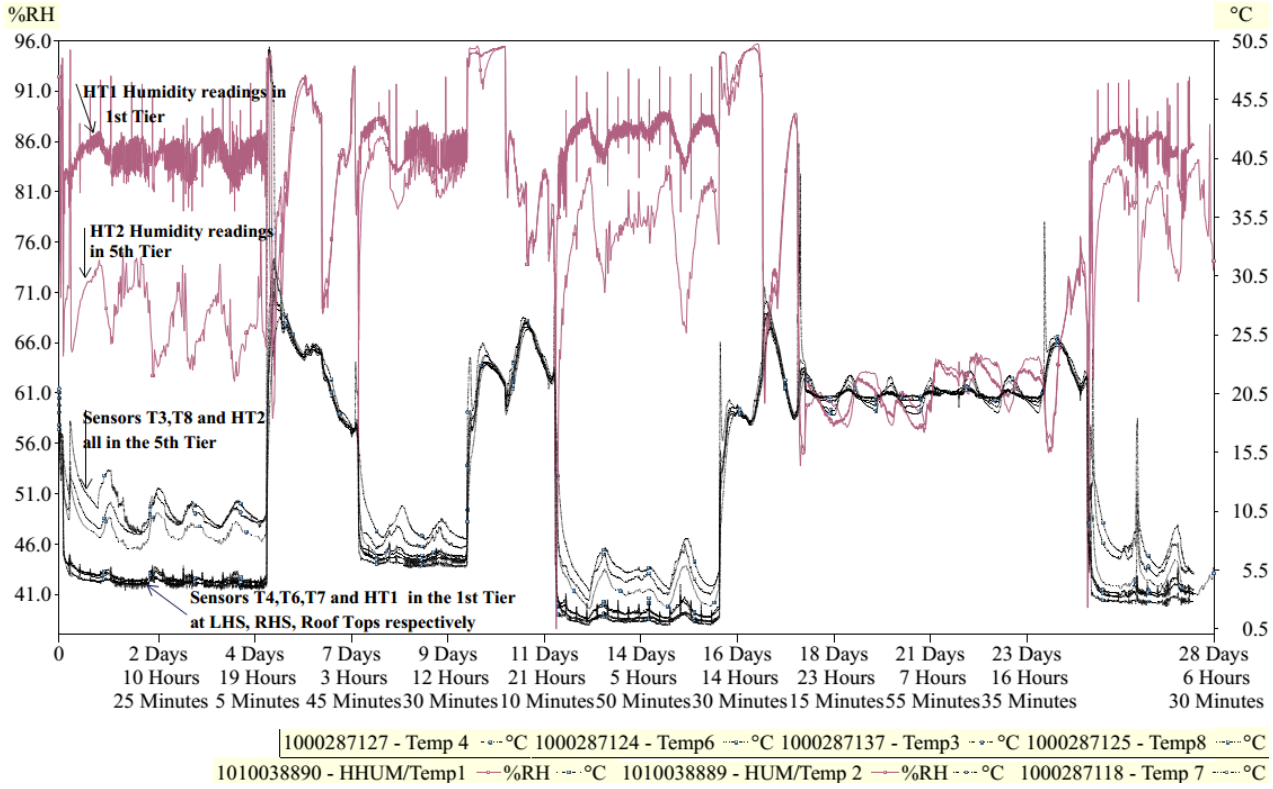



Figure 1: Test 1 Temperature and Humidity Variations at 1st and 5th Tier, roof and door sides of the container

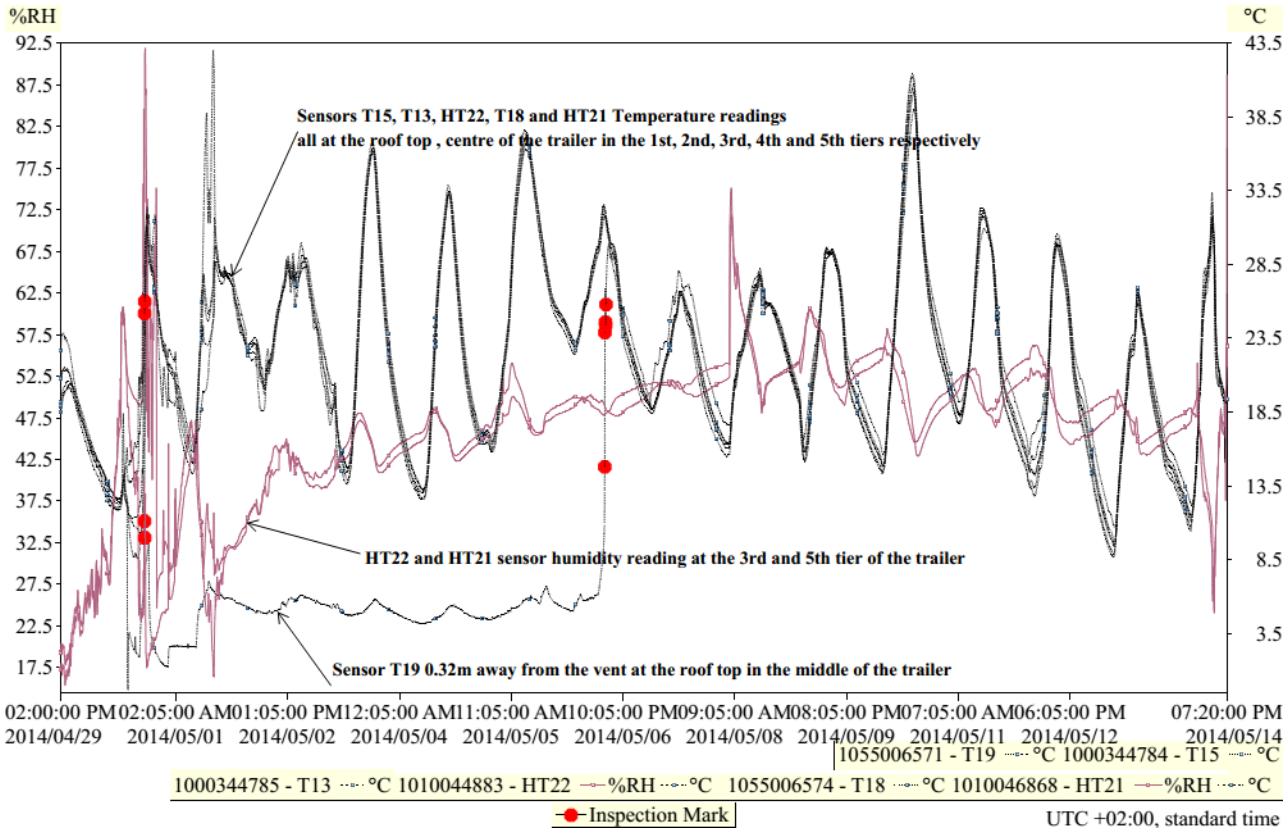



Figure 2: Test 2 Roof tops Temperature and Humidity Variations at 1st, 2nd, 3rd, 4th and 5th Tier of the trailer

## VI. VIABLE APPROACH FOR CONTINUOUS MONITORING OF COLD CHAIN OPERATIONS

The discussion and practical results displayed in previous sections provides proof that there is a definite need to use technology to provide visibility in cold chain operations. The technology options that can be used separately or in combination includes standalone data loggers, RFID based monitoring, wireless sensor network monitoring and real time web based monitoring. . The implications of each of these options are briefly discussed below:

- Standalone data logger based monitoring requires physical installation into the trailer and consignment, and manual data retrieval. These loggers are cheaper or have an equivalent cost as a RFID tag that offers similar sensing but better communication functionality. Data loggers therefore require more man-power and are susceptible to data loss as the loggers must be successfully retrieved after the trip to retrieve the data. This technique can however provide better characterization of temperature profiles based on the superior memory capacity of these loggers.
- RFID monitoring requires installing the technology in the trailer and environment to be monitored. Data are logged to tags' internal memory and later retrieved when the tags are in range of reader communicating at a regulated frequency band. Data retrieval can be manual or automatic depending on how the reader and tags have been configured and installed. Data can only be accessed when tags are within in range of a reader; there is however the possibility to retrieve the data without necessarily retrieving the tags or before the consignment has been offloaded. RFID tends to be cheaper than the alternative of wireless sensor networks.
- Wireless sensor networks based on protocols like Zigbee can also monitor transit cargo temperature effectively and can provide the user with threshold alarms when there is breach. Little man power is needed once the networks have been deployed. Unfortunately these devices cost 4 to 5 times as much to monitor the same amount of pallets than RFID tags.
- Real time web based monitoring can be supported by a combination of RFID for tag-reader communication and GSM for communication for reader to the outside world. This solution is ideal as it provides real time monitoring but is somewhat more expensive than a standalone RFID system – based on a technology survey that was conducted it will add approximately 20% to overall system cost. While this option is more costly it is also more efficient as the risk of losing data is very slim.

The above discussion shows that cold chain monitoring can be achieved effectively when the needs of and cost implication to the customer are correctly balanced. Based on the available evidence it would seem that in the long term, real time web based monitoring supported by a combination of RFID and GSM communications will be the better option to be used for temperature compliance in the cold chain. The optimal design of such a system in terms of the number of sensors required per consignment and the frequency of reporting however requires more accurate temperature profiling of a representative set of consignment types of trips; this is the primary focus of on-going research in this field.

## VII. CONCLUSIONS AND FUTURE WORK

This paper has demonstrated that there is a clear need for more accurate monitoring of cold chain operations in the supply chains of Southern Africa. Our experimental results have shown that serious deviations from standard best cold chain practices are occurring that contribute to the high levels of losses to perishable cargo.

We identified and evaluated a number of viable technology options and strategies to support cold chain monitoring, each associated with different levels of cost and efficiency. The optimal choice requires careful consideration of the needs of the various stakeholders and the nature of current operations.

Future work will aim to make the following contributions towards knowledge in the field of cold supply chain management:

1. Accurate spatial temperature profiles and cargo conditions within a reefer container loaded with different kinds of perishable cargo will be generated based on which it can be determined how many monitoring points are actually required.
2. A detailed comparison will be performed of the different options for low cost temperature monitoring methodologies and an optimal approach will be proposed.
3. A detailed databank will be developed containing a representative set of data for different kinds of cold chain operations. Such a databank can be used to design more effective cold chain management processes in order to protect the quality of the goods, and to create cold chain performance benchmarks.
4. A method will be proposed and evaluated to perform temperature monitoring as part of on-going operations based on wireless communications in order to support the real time management of cold supply chains in a cost-effective manner.

**Christian C. Emenike** received his undergraduate degree in Electrical and Computer Engineering in 2009 from the Federal University of Technology Minna, Nigeria and is presently studying towards his Master of Engineering degree at the North West University Potchefstroom Campus, South Africa. His research interests include wireless communications and telemetry.

**Alwyn J. Hoffman** received his B. Eng. degree in 1985, his M. Eng. degree in 1987 and his Ph.D. degree in 1991 from the University of Pretoria in South Africa. He has been professor in the School of Electrical, Electronic and Computer Engineering at North-West University since 1994. His current research interests include artificial intelligence, RFID and business intelligence.

## VIII. REFERENCES


[1] J. Bekker and M. Mostert, "Simulation of fruit pallet movement in the port of Durban : A case study," vol. 21, no. 1, pp. 63–75, 2005.

[2] Economic Research Division, "Estimated Impact of The Transnet Strike Action on Traders of Agricultural Products and Seafood," 2010.

[3] S. Barrientos, "South African horticulture : opportunities and challenges for economic and social upgrading in value Working Paper 12," no. September. pp. 0–43, 2012.

[4] H. Vega, "Journal of Air Transport Management Air cargo , trade and transportation costs of perishables and exotics from South America," *Journal of Air Transport Management*, vol. 14, no. 6, pp. 324–328, 2008.

[5] L. Ruiz-garcia, P. Barreiro, J. I. Robla, and L. Lunadei, "Testing ZigBee Motes for Monitoring Refrigerated Vegetable Transportation under Real Conditions," pp. 4968–4982, 2010.

[6] D. W. J. Thompson, J. J. Kennedy, J. M. Wallace, and P. D. Jones, "A large discontinuity in the mid-twentieth century in observed global-mean surface temperature.," *Nature*, vol. 453, no. 7195, pp. 646–9, May 2008.

[7] L. Ruiz-garcia, P. Barreiro, J. I. Robla, and L. Lunadei, "Testing ZigBee Motes for Monitoring Refrigerated Vegetable Transportation under Real Conditions," pp. 4968–4982, 2010.

[8] R. J. Blakey and J. P. Bower, "The Importance of Maintaining the Cold Chain for Avocado Ripening Quality," in *SOUTH AFRICAN AVOCADO GROWERS' ASSOCIATION YEARBOOK 32, 2009*, 2009, pp. 48–52.

[9] I. B. and S. Z. Tesfay, "Avocado sugars – key to postharvest shelf life?," 2010.

[10] L. Ruiz-garcia, L. Lunadei, P. Barreiro, and J. I. Robla, "A Review of Wireless Sensor Technologies and Applications in Agriculture and Food Industry: State of the Art and Current Trends," pp. 4728–4750, 2009.

[11] P. Barreiro and J. I. Robla, "Thermal study of a transport container," *Journal of Food Engineering*, vol. 80, pp. 517–527, 2007.

[12] J. Palafox-albarrán, R. Jederman, and W. Lang, "Prediction Of Temperature Inside A Refrigerated Container In The Presence Of Perishable Goods ." 2010.

[13] R. Jedermann, L. Ruiz-garcia, and W. Lang, "Spatial temperature profiling by semi-passive RFID loggers," *Computers and Electronics in Agriculture*, vol. 5, no. 65, pp. 145–154, 2009.

[14] P. Barreiro and J. I. Robla, "Thermal study of a transport container," *Journal of Food Engineering*, vol. 80, pp. 517–527, 2007.

[15] P. Barreiro, J. I. Robla, L. Ruiz-Garcia, and J. Rodriguez-Bermejo, "Review . Monitoring the intermodal , refrigerated transport of fruit using sensor networks," *Spanish Journal of Agricultural Research*, vol. 5, no. 2, pp. 142–156, 2007.

[16] H. L. C.M. Yeoh, B.L. Chai, T.H. Kwon, K.O. Yi, C.S. Lee, G.H. Kwark, "Ubiquitous Containerized Cargo Monitoring System Development basd on wireless Sensor Network Technology," *International Journal of Computers, Communications & Control*, vol. Vol. VI, no. No. 4 ( December), pp. 782–796, 2011.

[17] C. Turcu, Ed., "Sustainable Radio Frequency Identification Solutions," Feb. 2010.

[18] M. Babazadeh, "Plausibility check and energy management in a semi-autonomous sensor network using a model-based approach Plausibility check and energy management in a semi-autonomous sensor network using a model-based approach," University of Brumen, 2010.

[19] R. H. Bishara, "Cold Chain Management - An Essential Component of the Global Pharmaceutical Supply," *American Pharmaceutical Review*, no. February, pp. 1–4, 2006.

[20] Carlos Seminario; Emmanuel Marks, "Using Real-Time Truck Transportation Information to Predict Customer Rejections and Refrigeration-System Fuel Efficiency in Packaged Salad Distribution Master of Engineering in Logistics Using Real-Time Truck Transportation Information to Predict Customer," Massachusetts Institute of Technology, 2011.

[21] K. B. A. Busboom, "Container Surveillance System And Related Method," U.S. Patent US20040233041A12004.

[22] J. Frith, "Temperature Prediction Software for Refrigerated Container Cargoes." 2004.

[23] I. W. C. Chain-management, R. Kajetan, and D. Gibis, "Packaging Design For Chilled Product," In *Packaging Design For Chilled Product*, 2010, pp. 3–10.

[24] R. Jedermann, C. Behrens, R. Laur, and W. Lang, "Intelligent containers and sensor networks Approaches to apply autonomous cooperation on systems with limited resources," pp. 1–18.

[25] Z. S. Jan Havenga, L. E. J, and G. Leila, "Extending Freight Flow Modelling To Sub-Saharan Africa To Inform Infrastructure Investments – Trade Data ISSUES," *Journal of Transportand Supply Chain Management*, 2012.

# Using mobile networks for effective cold chain management

Bernardus P. van Eyk and Second Alwyn J. Hoffman
School of Computer and Electronic Engineering
Northwest University, Private Bag X6001, Potchefstroom 2520
Tel: +27 798515700
email: bpvaneyk@gmail.com

Abstract- **Research has shown that substantial in transit losses in perishable cargo are sustained due to malpractices in cold chain logistics. Such losses are even more prevalent in developing countries where supply chain systems are not yet as advanced as those found within developed economies. This situation can be improved by using real time monitoring of refrigerated cargo, which implies the use of mobile communication technologies to sustain monitoring through the entire supply chain. This paper investigates the application of remote sensing technologies, including RFID, wireless sensor networks, GSM and satellite communications, to provide cost-effective solutions to the needs of the refrigerated supply chain industry. A system architecture is proposed that can build upon existing vehicle tracking and supply chain management systems in order to maximize the utilization of existing infrastructure.**

**Index Terms—RFID, GSM, GPRS, WSN, LSP**

## I. INTRODUCTION

The effective control of freight logistics is a critical part of moving goods from the producer to the end user. Depending on the items being transported the rules applied during transit will vastly differ. For perishable goods one of the most important rules that must be adhered to is the temperature thresholds to be maintained during all transportation processes. If the required thresholds are not maintained it could lead to reduced shelf life or the loss of the entire consignment. In fresh produce logistics alone losses of 35% globally and 40% in developing countries have been documented [1]. The temperature at a specific moment during transit is not the only concern, but how the consignment was handled over its entire life cycle is of particular importance and therefore has to be monitored.

### *A. Key fundamentals of cold chain management ( CCM )*

The fundamentals applied in the cold chain are based on the standard model that is applied in other supply chains. A balance must always be maintained between the safe transportation of goods and the cost-effectiveness of the logistic process. In the supply chain there are seven fundamentals that are applied to maintain this necessary balance [2] :

1. The segmentation of customers based on services needed.
2. Customization of the logistic network to fit these needs.
3. Evaluation of market demand and planning according to the results obtained.
4. Dynamically deciding which product to stockpile and so speed up conversion across the supply chain.
5. Use resources strategically.
6. Develop a supply chain-wide technology strategy.
7. Adopt channel-spanning performance measures.

By applying these fundamentals to the cold chain a company can maximize its efficiency whilst being competitive in its individual markets. In the application of these fundamentals the standards that must be adhered to, to ensure goods are delivered at its optimal quality to the consumer, should not be neglected. A direct correlation between the deterioration of perishable goods and their relative temperatures can be established. As temperature increases the tempo of the natural degradation of these goods will also increase. A few examples of the influence of temperature on some perishable goods are given in Table 1.

Table 1: Estimated shelf life of perishable goods under different temperatures

| ***Item*** | ***Storage potential*** | | | |
|---|---|---|---|---|
| *Type of Goods* | *Optimal temp (OT)* | *OT +10°C* | *OT +20°C* | *OT +30°C* |
| ***Fish*** | *0°C* | *10°C* | *20°C* | *30°C* |
| ***Shelf life*** | *10days* | *4-5 days* | *1-2 days* | *Few hours* |
| ***Mangoes*** | *13°C* | *23°C* | *33°C* | *43°C* |
| ***Shelf life*** | *2-3weeks* | *1 week* | *4 days* | *2 days* |
| ***Green Vegetables*** | *0°C* | *10°C* | *20°C* | *30°C* |
| ***Shelf life*** | *1 month* | *2 weeks* | *1 week* | *<2days* |
| ***Apples*** | *-1°C* | *10°C* | *20°C* | *30°C* |
| ***Shelf life*** | *3-6months* | *<2 months* | *<1 month* | *2 weeks* |

As the table indicates the influence of temperature on shelf life varies according to the item being stored. While temperatures higher than the optimal temperature substantially reduces the shelf life even greater losses are suffered for goods being stored at temperatures below the optimal temperature. This may lead to chilling damages which can render a whole consignment unfit for distribution. Most first world countries have adopted cold chain control mechanisms and as a result service providers have deployed temperature sensor networks to manage cold chains to ensure the longest possible life for goods being stored and transported from producer to consumer. These systems can unfortunately be very costly and thus is not being implemented in a similar fashion in developing countries.

### *B. Cold chain management in South Africa*

In typical supply chains operated in Southern Africa cold chain integrity is maintained for certain sections during transit; full traceability of the cold chain from producer to consumer is however not yet a reality. This at least partly results from the fact that the independent entities in the value

chain, including the producer, logistics service providers (LSPs), the retailer and the consumer each have different priorities when it comes to the delivery of goods.

Producers and LSP's need to maintain their profit margins and would therefore try to limit additional costs such as refrigerated transportation to the distribution center (DC); this may lead to retailers receiving goods that will not have the shelf life promised to consumers. LSPs however increasingly realize the importance of cold chain management over the entire process to guarantee the delivery of goods of the required quality, and the value of real time monitoring to ensure traceability and accountability.

### *C. Problems encountered in current structures*

The implementation of technology for the purpose of cold chain management is unfortunately not done without encountering many hurdles. A great many of the solutions available today is designed for the technological excellence of the equipment and not for the effectiveness of solving the problem on ground level to meet the specific needs of a client. As suggested the solutions on offer have some disadvantages that limit their practical deployment:

- Complexity of solution beyond the scope of the operator.
- Equipment is labor intensive to use.
- The equipment can get lost very easily in transit.
- The solution is prone to malfunction and damage.
- The solution is perceived to be too expensive relative to the benefit it provides.
- The solutions do not integrate with the current systems customers are already using.
- The solution does not fit into existing internal operational processes.

Due to these disadvantages many potential end-users are hesitant to pursue existing solutions to manage their cold chains more effectively.

### *D. Research objectives presented in this paper*

CCM is of crucial importance to improve the quality of perishable goods being transported. In order to make a contribution to existing knowledge in this field it is firstly important to understand the existing operational processes followed in typical cold supply chains and to quantify the need for continuous monitoring of cargo. From this study we can determine what combination of functional needs must be satisfied to provide a solution that will make it possible to eliminate existing problem areas.

Secondly we need to understand the capabilities and limitations of existing technology that should be considered in designing an effective solution. This will include sensors with built-in communication capabilities, wide-area communication networks and the required levels of intelligence at different layers in the system. In this respect we need to consider not only the functionality of the technology options but also cost and interoperability. The third aspect to be researched focus on the design of a system architecture that can integrate all of the functional capabilities into a system that will provide the required monitoring needs,existing infrastructure use and that can be integrated with existing legacy systems used in the supply chain.

Section 2 defines the end-user needs. Section 3 considers interoperability issues and network standards used by individual technologies. Section 4 discusses existing technologies that should be considered in the design of a solution, their benefits and implementation requirements. Section 5 proposes an architecture to aid in the integration and combined use of different systems. The paper concludes with a summary of the results.

## II. DEFINING END-USER NEEDS

The complexity of monitoring perishable cargo in the supply chain derives from the fact that, due to the biological processes within the cargo itself, temperatures cannot be expected to be uniform throughout an entire cargo consignment. Ideally the temperature of the cargo itself should be monitored, rather than only the temperatures on the inside of the container in which it is transported. This implies the need to communicate wirelessly with sensors that have been embedded inside the cargo. An important research question is the required number of points to be monitored within each type of perishable cargo, how often measurements should be taken and how frequently this data needs to be communicated to the outside world to prevent losses.

Continuous wireless communication with such embedded sensors may present a problem due to the difficulties associated with RF propagation through substances containing high levels of water, as is typical of fresh produce. In order to retrieve accurate audit trails it is however necessary to continuously record the temperatures that were experienced by the cargo; this implies the need for local data storage capabilities on the sensor to allow the historical audit trail to be retrieved as soon as communication with the sensor can be re-established.

Though it may not be possible to sustain continuous communications with all deployed sensors it will be of importance to at least have real time visibility of the remote monitoring system and of a limited set of sensor readings. If the temperatures on the periphery of the container, where a combination of wired and wireless communication should be possible, could be available in real time, this information could be combined with information regarding the general status of the vehicle (GPS location, whether it is moving or not, if the doors have been opened, etc.) to identify potential high risk situations that may lead to loss of cargo.

The above description implies that intelligence regarding the logic of the supply chain operation must be designed into the system. Some of the intelligence must reside at sensor level (e.g. to record data locally if communication with the outside world cannot be established, as well as recording the time at which defined thresholds have been exceeded). At the transport vehicle level further intelligence is required, amongst others to implement an optimal balance between remote communication costs and the risk of losing cargo (e.g. to generate alarms and communicate alarms when specific rules have been breached in terms of temperature thresholds or where and for how long doors were opened). The central monitoring function must apply system-wide intelligence by interpreting information received from remote monitoring stations within the context of the entire logistical operation. An example will be to interpret cargo

temperatures data based on the GPS location of the consignment (e.g. when the vehicle has reached a DC and goods are off-loaded some temporary temperature fluctuations should be expected).

## III. Network Standards Utilised

The limited extent to which systems have been practically deployed is partly due to the fact that most available systems are based on proprietary protocols used for communication between different system components. This exposes end-users to the risk of being locked into a single source of supply. Interoperability between tags, readers and remote tracking units will help to open up the market and to allow end-users to use the same technology infrastructure for multiple purposes.

### *A. Interoperability of systems*

Due to the use of wireless communication at different levels within the CCM system the possibility of communication interference cannot be overlooked [8] – GSM and UHF RFID for example both use communication in bands close to 900 MHz. The implementation of standards based communication will however ensure that such interference is prevented; standards based protocols will furthermore provide the required level of anti-collision to ensure that large numbers of data packages from RFID tags will be successfully collected by readers even when many devices are operated in close proximity. This will prevent a system from becoming unstable or dysfunctional when its functionality is scaled up to accommodate the needs of a large supply chain operation. Without such standards, providing the ability to integrate different devices into a single large system, effective CCM will not be possible.

As most of the options for remote communications use existing GSM or satellite communication networks. This aspect of the systems tend to be standardized based on the protocols that form part of those networks. The bigger challenge is to ensure interoperability between the different kinds of RFID tags and the devices that need to connect them to the outside world.

### *B. Network standards used by the systems to enable effective communication.*

The air interface for RFID is managed by the ISO 18000 network standard [12]. The ISO 18000-1 standard defines the generic parameters for global air interfaces and the ISO 18000-2 to 18000-7 standards define the unique parameters according to the frequency range for device communication. These standards provide guidelines for the use of RFID tags in an integrated sensor network.

While most commercially available passive UHF RFID tags follow the ISO18000-6C standard, this is not true in the case of active RFID, where only a small fraction of available solutions are based on the ISO18000-7 standard. Table 2 indicates what network standard is used with different kinds of RFID tag.

Table 2: RFID networks standards used [7]

| Interface Type | Frequency | Range | Standard |
|---|---|---|---|
| LF | 125 kHz | 30cm | ISO 18000-6A |
| HF | 13.56 MHz | 1m | ISO 18000-3 |
| Active RF | 433MHz | 1m+ | ISO 18000-7 |
| UHF | 850-950 MHz | 10m + | ISO 18000-6C |
| Microwave | 2.4 - 2.45 GHz | 100m + | ISO 18000-4 |

The standards provide a margin of flexibility to enable the interoperability of multi vendor devices. ISO 18000-6C for example enables the use of sensors with the network devices and requires that a read operation be completed within a defined period so enabling multiple tags to be read by the reader and operation to take place with a broad range of variability [7]. Standards based RFID is limited to communication ranges up to hundreds of meters; to add the ability to communicate over long ranges, the use of GSM or satellite communications is required. The digital nature of GSM allows for synchronous and asynchronous transmission of data to and from ISDN terminals. GSM operates on the 900MHz and 1.8GHz frequency band, providing a 64Kbit/s signal for data transmission over long distances to the appropriate party for analysis [11].

## IV. Technological solutions

The implementation of technology for CCM has been applied with varying success in many application areas to enhance the visibility and environmental conditions of consignments. The type of technology used in a container directly influences the data that can be acquired and the level of control over the monitoring operation during transit.

### *A. The Influence of mobile technological solutions on cold chain management*

The growth in size and complexity of the logistics industry has led to a growing need to progress from traditional methods of temperature monitoring and data logging. This is where RFID (Radio frequency identification) and GSM technologies has become a vital part of any sophisticated logistic chain. These technologies have the ability to store thousands of data samples during transit [3] and either relay this data in real-time to a web application or store the information for retrieval once an RF connection has been established.

This shift to mobile technology has proven that with little to no human participation enhanced traceability of a consignment and the retrieval of large quantities of data from many sensors are a possibility [4]. To truly understand what the different technologies are capable of an analysis of the functionality of RFID and hybrid GSM/RFID solutions will now be discussed.

### *B. RFID technologies*

RFID is a relatively new technology in cold chain management that can add much value in terms of detailed visibility but that is also relatively complex to deploy. In order to justify its use within a specific scenario it is essential to first determine the expected ROI (return on investment) in order to decide if the technology is worthwhile to implement [5].

RFID sensors can fall into one of three possible categories: passive, semi-passive and active. Each category has essential characteristics that make it ideal for specific applications. The different types are compared in Table 3

Table 3: Comparison of RFID tag categories [6]

| Tag Type | Passive | Semi-Passive | Active |
|---|---|---|---|
| Power Source | Harvesting RF energy | Partly Battery | Battery |
| Communication mode | Response only | Response only | Respond and initiate |
| Max Range | 10m | >10m | >100m |
| Relative Cost | Least Expensive | More Expensive | Most Expensive |
| Example Supply Chain Applications | Spotting items at defined points | Continuous monitoring of status of tagged items | Continuous monitoring of status of tagged item |

The primary differences between the types of RFID tags are the ability (or not) to continuously record the current status of the tagged item, as well as the read range over which information can be retrieved from the tag. While passive RFID is used to only spot the presence of a tagged item passing a point where a reader has been installed, semi-passive and active RFID can record data continuously and communicate over longer ranges. RFID sensors come in many shapes and sizes and operate uniquely to the application it has been developed for. Several factors can influence the functionality that is achieved within specific applications; these include the physical environment and the physical properties of the material the sensor is placed on [7].

Wireless sensor networks (WSNs) can also contribute towards cost and energy effective cold chain monitoring, as they can provide sensors with the ability to relay data via other sensor nodes in the network. The combination of the individual sensors within a WSN allows the environmental status to be monitored with a higher degree of accuracy than what would be possible if all sensor nodes had to communicate directly with a reader node or hub. The ISO18000-7 protocol for active RFID provides for the options of communication between tags in order to effectively implement a WSN. The application of interpolation techniques that use available temperature data to calculate temperature values in other sectors in the consignment were no sensors are placed can provide improved scalability and portability to sensor networks as it can reduce the unnecessary cost of additional sensors[8].

In order to extend the capabilities of WSNs to roaming networks that deliver real-time data to a data center, the WSN must be connected to a communication device capable of roaming. The most effective way to enable roaming sensor networks to deliver real-time data is its integration with cellular technologies to enable long range communication of the WSNs.

### *C. Hybrid solutions*

The integration of cellular technology such as GSM and GPRS with RFID into a WSN provides an efficient method to enable real-time monitoring of a consignment in transit [9]. The technology is however limited to the cellular network coverage in the country where the consignment is done. In Africa this solution would work effectively in Nigeria, Egypt and South Africa with varying functionality in the other 51 countries [10]; in other African countries this method for remote communications can become very expensive due to roaming charges amongst different network operators.

A hybrid system consists of the combination of RFID sensor tags with a RFID reader that has cellular communication capabilities. The data is read from the tag's internal memory and then encapsulated into a data package according to the protocol being used and send to a data server. The layers of encapsulation is then removed by the middleware and stored into the database, from where the user can access the data through a GUI designed for the system. This system topology is shown in Figure 1.

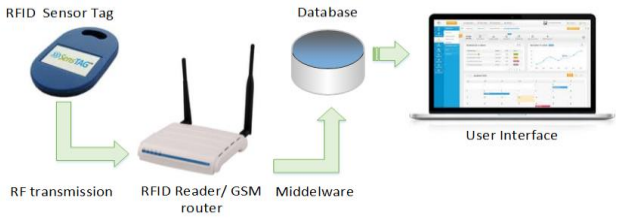



Figure 1: Hybrid system topology

The real-time monitoring benefits of such hybrid systems based on the added visibility to the cold chain are clear. The user, most likely LSP is provided with the ability to check if the internal cargo temperature of a consignment is maintaining the cold chain requirements for the specific product. Besides the real time monitoring capabilities the system can also aid in establishing accountability for damage to goods in case of to failure to maintain the cold chain standards, by recording at which point the cold chain thresholds were exceeded. This can protect a service provider from illegitimate claims for compensation if a consignment is indeed lost.

### *D. Solution integration and environmental analysis*

Each supply chain has its own obstacles that must be confronted to achieve effective temperature tracking. In our technology survey [11] a group of different technological solutions available for CCM where evaluated to identify the limitations and abilities of each option. In this section the application of the different systems in different scenarios are investigated.

The first system consists of standalone data loggers placed on the interior of the container wall and within the consignment. These data loggers have the ability to measure the immediate ambient temperature with an accuracy of ±0.2°C; there is also the possibility for external probes added to the data logger for internal temperature measurement. The loggers can typically store between 1000 and 16000 data samples on its internal memory. The sampling rate can be defined using a custom software application. The primary limitation of this option is the lack of any remote communication and the need for manual data retrieval, which is based on a USB port or through a magnetic reader.

An alternative to the standalone data loggers are GSM sensor units. The devices generally offer two options to monitor temperature: they are either equipped with an internal sensor or they can support external wired and/or RF

sensors to monitor the temperature/humidity of a container. The unit is typically installed inside the container wall in a position where the airflow in the container converges and where the most accurate measurement can be taken. Real-time monitoring of air inflow temperature and optional position tracking is possible. The unit generally requires an external antenna to connect to the mobile cellular or satellite network. This may require modifications to be made to the container to place the antenna in a position with the lowest signal attenuation. The units can store between 4K and 16K of data on its internal memory. To limit the roaming cost of these devices the system must be configured to either send data at pre-determined intervals, on request by the user or when a specific set of conditions is satisfied (e.g. when a threshold is exceeded).

The next option is the use of RFID tags with internal data logging capabilities that are embedded within the cargo and read by readers installed at the depots. The tags have up to 16000 sample internal storage capability and log data on a FIFO basis. The communication range depends on the type of tag/reader combination and the amount of tags read simultaneously by the network standard used. The data read by the readers are relayed to a web server where it is analyzed by a software agent. The reliable retrieval of the RFID tags from a consignment is a significant practical obstacle that can lead to equipment and data losses.

A further alternative is for a reader to be installed inside the container to retrieve data in real time. The collected data is stored on the internal memory of the reader and is sent to a central server when the container is in range of the selected wide area network. This can either be a Wifi network (which will limit the retrieval of data to take place when the vehicle enters a supplier or customer depot) or it can be a cellular network that will enable real time data retrieval when required. In this setup the tags can either store data internally or relay the measured data to the reader inside the container.

When the tags are not equipped with internal memory and must communicate continuously with the container reader data may be lost in case of a temporary break in tag-reader communications due to the screening effect of the cargo itself. Tags without internal memory therefore represent a less effective solution to manage the cold chain. GSM systems further have vulnerabilities due to coverage differing along each route. In the blackout areas real time communication is not a possibility and data is only forwarded when the network coverage is re-established with sufficient signal strength.

### *E. Cost analysis of solutions*

A survey [11] has been done to evaluate the different systems aimed at effective management of the cold chain based on cost and functionality. In the survey 12 technology options were evaluated, all of which provide RF communication capability and is relatively low cost for the functionality they provide. A cost estimation for the systems was done using the base scenario where 1000 pallets were monitored by an RFID system or 100 pallets by GSM sensor systems for a minimum duration of 1 year. The GSM sensor data is read when needed in transit while the RFID data is read at critical points along the route. The technology options, with their required software are shown in Table 4.

Table 4: Mobile cold chain solutions

| Scenario | Supplier | GSM | Rf-Enabled | Intern. Storage | Comm. Standard |
|---|---|---|---|---|---|
| RFID Systems | | | | | |
| 1 | Wireless Links | X | X | | ISO 18000-7 |
| 2 | YRless International | | X | X | Proprietary |
| 3a | YRless International | | X | X | Proprietary |
| 3b | YRless International | | X | X | GSM |
| 4 | GOARFID | | X | X | ISO 18000-7 |
| 5 | GOARFID | X | X | X | ISO 18000-6C |
| 6 | CAENRFID | X | X | X | ISO 18000-6C |
| 7 | Sensmaster | X | X | X | Proprietary |
| 8 | MicroDAQ | X | X | X | Proprietary |
| GSM Sensor systems | | | | | |
| 9 | Aptifirts | X | | X | GSM |
| 10 | HW group | X | | X | GSM |
| Standalone Data loggers | | | | | |
| 11 | Gemini Data loggers | | X | X | ISO 18000-6A |
| 12 | Elpro | | X | X | USB |

The cost of each system for the parameters of the scenario has been compared; the resulting data is depicted in Table 5.

Table 5: System Cost Comparison

| System | Reader Cost (R) | Sensor Cost (R) | Total System Cost(R) | Software Cost(R) |
|---|---|---|---|---|
| 1 | 12640 | 765000 | 777640 | 143880 |
| 2 | 21200 | 590000 | 611200 | 12000 |
| 3a | 21200 | 531000 | 552200 | 12000 |
| 3b | 319800 | 0 | 319800 | 60000 |
| 3 | 341000 | 531000 | 872000 | 72000 |
| 4 | 12980 | 638000 | 650980 | 0 |
| 5 | 40700 | 935000 | 975700 | 0 |
| 6 | 17136 | 456000 | 473136 | 0 |
| 7 | 11252.4 | 454900 | 466152 | 0 |
| 8 | 13618 | 1094500 | 1108118 | 0 |
| 9 | 711000 | 0 | 711000 | 0 |
| 10 | 1082000 | 0 | 1082000 | 0 |
| 11 | 2300 | 925000 | 927300 | 1814 |
| 12 | 0 | 1850000 | 1850000 | 0 |

### *F. Summary of mobile technologies*

The technological solutions evaluated,each have the ability to enhance the visibility of the cold chain. While the systems work well independently of each other, interoperability between most of these options will be difficult due to the use of proprietary standards. The systems 2, 3a, 7 and 8 each uses a proprietary communication standard and the systems 1 and 4 the ISO 18000-7 standard that is also used in combination with proprietary protocol for operation with the

specific readers it was designed for. The systems 5 and 6 using the ISO 18000-6C standard offer the best options to be used in an interoperable system. The UHF tags can be read by any reader using the standard, thus accommodating the possibility to integrate into an intelligent architecture for supply chain management.

The cost comparison indicates that RFID systems cost less to implement for a similar application than GSM sensor networks and standalone data loggers. The cost intensive component in RFID system is the reader and generally offers data retrieval only at depots and not in transit as the GSM sensors. This can be overcome by installing a reader with GSM capability on the container as is the case for system 1.

## II. Generic architecture to promote CCM for LSP's

In order to proficiently utilize the functionality of the systems and technologies as discussed above an architecture must be in place that is scalable to accommodate the complexities of the LSPs' application environment as well as the inherent nature of the goods that are transported. The proposed interoperable architecture is depicted in Figure 2

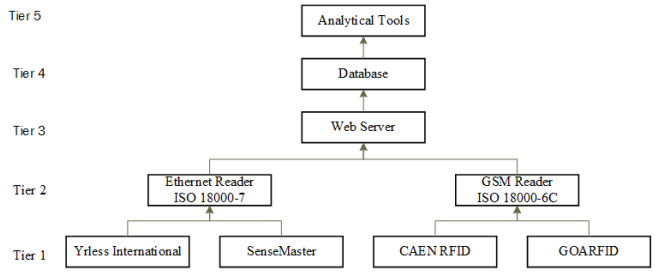



Figure 2: Proposed Architecture

The architecture is built on the principle that a single reader can be used to retrieve information from multiple tags from various vendors using generic network standards for data communication between tiers 1 and 2. At tier 1 sensors will retrieve and store data values awaiting configured transmission interval. When readers at tier 2 receives transmitted data it will then be forwarded using the current network infrastructure until it reaches tier 5. In order to limit the operational costs, data will only be transmitted in interval packages to the web server and on occurrence of temperature threshold deviations as set up by the user using analytical tools to determine the requirements for a specific consignment. At tier 5 the data will be processed using other analytical tools and interpolation techniques to provide the user with relevant information as configured. The architecture thus proposes to utilize ISO network standards to accommodate the use of products from different vendors under a single operation platform

## III. Conclusion and future work

Cold chain management is an important aspect of all supply chains involving temperature sensitive goods. The incorporation of technological systems including RFID sensors, mobile readers and GSM or satellite units into a roaming telemetry system provides the additional tools to help optimize CCM and decrease losses that may occur due to operator negligence regarding cold chain practices. Due to the complexity of the supply chain and of the technologies involved the proper utilization of these systems can only be achieved when implemented based on a well-designed architecture. This paper described how a combination of technologies can be deployed to achieve the objective of improved CCM to deliver better quality product with reduced in transit losses. Future work will focus on the practical deployment of some of the proposed system architectures and the evaluation of such prototype systems in order to verify their performance in the field against user requirements.

## IV. References


[1] L. Kitinoja, "Use of cold chains for reducing food losses in developing countries," ,PEF White Paper, no. 13, pp. 1–16, 2013.

[2] D. L. Anderson, F. F. Britt, and D. J. Favre, "The Seven Principles of Supply Chain Management," vol. 185647, pp. 1–15, 2005.

[3] A. Guillen, "RFID and Cold Chain Management,", International pharmaceutical industry, vol. 4, no. 4, 2012.

[4] E. Abad, F. Palacio, M. Nuin, a. G. De Zárate, a. Juarros, J. M. Gómez, and S. Marco, "RFID smart tag for traceability and cold chain monitoring of foods: Demonstration in an intercontinental fresh fish logistic chain," *J. Food Eng.*, vol. 93, no. 4, pp. 394–399, Aug. 2009.

[5] T. Blecker and G. Huang, "RFID in Operations and Supply Chain Management.", E.S. Verlag GmbH & Co, 2008

[6] S. A. Weis, "RFID (Radio Frequency Identification): Principles and Applications", MIT CSAIL, 2006

[7] A. Characterization, C. Test, and T. Parties, "Advanced RFID Measurements : Basic Theory to Protocol Conformance Test," pp. 1–10, 2013.

[8] A. Mitrokotsa and C. Douligeris, "Integrated RFID and Sensor Networks : Architectures and Applications,"RFID and Sensor Networks, pp. 511–536, 2009.

[9] A. Higgins, S. Reidy, F. Barrett, S. Klein, and N. Johanning, "Real-Time Cold Chain Mapping,", White Paper, 2010.

[10] T. Phillips and P. Lyons, "African Mobile Observatory 2011," GSM Asociation, 2011.

[11] "THRIP Project : Intelligent Freight Logistics Technology Survey : Cold chain management," Internal Research Report, North-West University, 2014.

[12] D. Reigelsperger and J. Harkins, "Keeping Pace With RFID", Lowry Computer Products Webinar



**Bernardus P. van Eyk** received his undergraduate degree in 2013 from the Northwest University, Potchefstroom and is presently studying towards his Master of Engineering degree at the same institution. His research interests include RF communication, Networking protocols and standards and Systems Integration and optimization.

**Alwyn J. Hoffman** received his B. Eng. degree in 1985, his M. Eng. degree in 1987 and his Ph.D. degree in 1991 from the University of Pretoria in South Africa. He has been professor in the School of Electrical, Electronic and Computer Engineering at North-West University since 1994. His current research interests include artificial intelligence, RFID and business intelligence.

# On the optimal Artificial Neural Network architecture for forecasting TCP/IP network traffic trends.

Vusumuzi Moyo and Khulumani Sibanda
Department of Computer Science and Information Systems
University of Fort Hare
P. O. Box X1314, Alice 5700, RSA
Tel: +27 40 6022464, Fax: +27 40 6022464
email: {vmoyo, ksibanda}@ufh.ac.za

**Abstract- Artificial Neural Networks (ANNs) have attracted increasing attention from researchers in many fields. They have been successful in solving a wide range of real world problems in various domains. A particular area in which ANNs have featured prominently is in the forecasting of TCP/IP network traffic trends. One of the most difficult and least understood tasks in the design of Backpropagation ANN models is the selection of the most appropriate network architecture. Although some guidance in the form of heuristics is available for the choice of this parameter, none have been universally accepted. To date there is no particular proven method or approach to determine the best neural network structure that we are aware of. In this paper we investigate various ANN architectures with the aim of determining the optimum network architecture, which is one of the most important attributes of an ANN. We used experimental method which is a proven method for testing and exploring cause and effect relationships. MATLAB version 7.4.0.287's Neural Network toolbox version 5.0.2 (R2007a) was used for our experiments. Our study found out that in contrast to Occam's razor principle for a single hidden layer, an increase in number of neurons produces a corresponding increase in generalization ability of a Neural Network. Furthermore considering network architectures, those with between 40 to 50 hidden neurons performed comparatively similar to the 60 hidden neurons. Perhaps, the most important conclusion derived from this study is that large networks do not always improve the generalization ability of ANNs.**

**Index Terms— Generalization ability, Artificial Neural Networks and Artificial Neural Network architecture**

## I. INTRODUCTION

Artificial Neural Networks (ANNs) have been used in many fields for a variety of applications, and proven to be reliable. Inspired by biological systems, particularly the observation that biological learning systems are built of very complex webs of interconnected neurons, ANNs are able to learn and adapt from experience. They have demonstrated to be one of the most powerful tools in the domain of forecasting and analysis of various time series [1]. Time Series Forecasting (TSF) deals with the prediction of a chronologically ordered variable and one of the most important application areas of TSF is in the domain of network engineering. As more applications vital to today's society migrate to TCP/IP networks it is essential to develop techniques that better understand and predict the behaviour of these systems. TCP/IP network traffic forecasting is vital for the day to day running of large/medium scale organizations. By improving upon this task, network providers can optimize resources (e.g. adaptive congestion control and proactive network management), allowing an overall better Quality of Service (QoS). TCP/IP forecasting also helps to detect anomalies in the network. Security attacks like Denial-of-Service (DoS) or even an irregular amount of SPAM can be detected by comparing the real traffic with the values predicted by forecasting algorithms, resulting in economic gains from better resource management.

Owing to the importance of TSF, several TSF methods have been proposed such as ARIMA, Box Jenkins model, Holt Winters and ANNs. Two of the most recent discoveries of the statistics of internet traffic over the last ten years are that internet traffic exhibits self-similarity and non-linearity behaviour. This nature of network traffic makes high accurate prediction difficult. Literature from various authors has indicated that unlike other methods, ANNs can approximate almost any function regardless of its degree of nonlinearity and without prior knowledge of its functional form [2]. Literature also shows their performance and success in predicting TCP/IP network traffic trends increasing in accuracy from 64.3% in earlier approaches to 70% and better [3, 4]. While ANNs provide a great deal of promise, they also embody much uncertainty. This is mostly due to the fact that ANNs are a relatively recent innovation thus they are generally not well understood. Current and potential users tend to treat them as 'black box' models. One major issue that limits the applicability of ANN models in forecasting tasks is the selection of the optimal network architecture. Network architecture means the set of input neurons, hidden neurons and output neurons together with the connections between neurons and the neuron groupings which combine to form a network of computing elements [5].

The network architecture is of great importance as it dictates the number of free parameters available during training, learning time, but most importantly it affects the generalization capabilities of the network. Generalization is a measure that tells us how well the network performs on the actual problem once training is complete. Once the ANN can generalize well, it means that it is capable of dealing with new situations such as a new additional problem or a new point on the curve or surface. In the absence of an exact paradigm to estimate an optimally or near-optimally performing ANN architecture, literature has been inundated with several strategies and heuristics to estimate the optimum number of hidden layers/neurons. However, none of these methods has the theoretical or practical rigor of revealing optimal or at least near-optimal solutions. In this

paper the effect of different network architectures on the performance of ANNs is investigated. Although the results presented in this paper are for a particular case study, they provide a valuable guide for network engineers and scientists who are currently using, or intend to use ANNs for the prediction or forecasting of TCP/IP network traffic.

## II. ARTIFICIAL NEURAL NETWORKS

Haykin [6] defines ANNs as "*physical systems which can acquire, store and utilize experimental knowledge*". The basic unit of an ANN is a neuron. An artificial neuron acts in the same way as a biological neuron; each has a set of inputs and produces an output based on the inputs. A biological neuron produces an output by comparing the sum of each input to a threshold value. Based on that comparison it produces an output. In addition, it is able to differently weigh each input according to the priority of the input. The inputs and outputs of a biological neuron are called synapses and these synapses may act as inputs to other neurons or as outputs such as muscles. Thus it creates an interconnected network of neurons which combined produce an output based on a number of weights, sums and comparisons. One motivation for ANN systems is to capture this kind of highly parallel computation based on distributed representations.

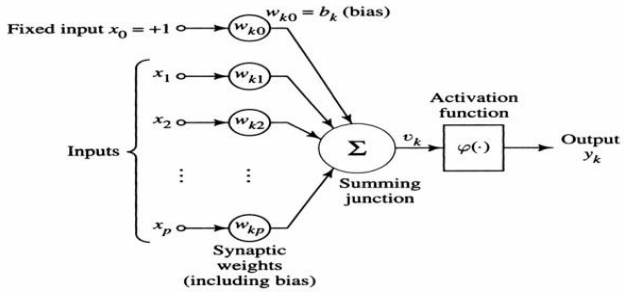



**Figure 1:** An artificial neuron (adapted from [7])

Figure 1 shows the typical structure of an artificial neuron, the inputs are denoted by $x_1, x_2 \ldots x_p$ and weights are denoted by $w_{ko}, w_{k1}, w_{k2} \ldots w_{kp}$. The neuron calculates the weighted sum $w_k, x$ as:

$$\boldsymbol{w_k, x = \sum_{i=1}^{p} w_{ki} x_i} \qquad (1)$$

The output of the neuron is governed by the activation function, which acts as a threshold. The output is given by:

$$\boldsymbol{y_k = f(\sum_{i=1}^{p} w_{ki} x_i + b_k)} \qquad (2)$$

Where *f is* the activation function, $(b_k)$ is the bias and $y_k$ is the output signal.

Among the various types of ANN models, Multilayer perceptron (MLP) is the most extensively applied to a variety of problems. MLPs are formed by several neurons arranged in groups called layers. The most popular and the simplest MLP consist of three layers, an input layer, a hidden layer, and an output layer. The network thus has a simple interpretation as a form of input-output model, with the weights and thresholds (biases) being the free parameters of the model. The sliding time window approach is the most common MLP model for forecasting. It takes as inputs the time lags used to build a forecast and it is given by the overall formula:

$$\boldsymbol{X_{p,t} = w_{o,0} + \sum_{i=I+H}^{I+H} f \sum_{s=1}^{k} \sum_{r=1}^{w_s} X_{st-L_{sr}} w_{i,j}} \qquad (3)$$

Where $w_{i,j}$ is the weight of the connection from node $j$ to $i$ (if $j$=0 then it is a bias connection), $o$ denotes the output node and $f$ is the Logistic sigmoidal activation function.

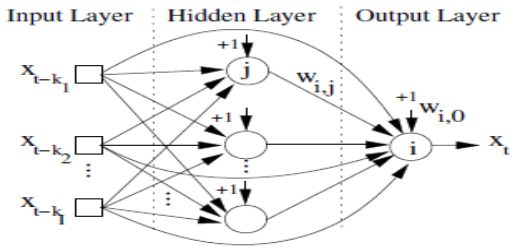



**Figure 2:** Sliding time window MLP (adapted from [7])

In the vast majority of papers that deal with the prediction and forecasting of TCP/IP traffic, Feedforward networks optimized with the aid of the Backpropagation (BP) algorithm have been used. According to Gately [8], "*this is because BP is easy to implement and fast and efficient to operate*". The BP process is commenced by presenting the first example of the desired relationship to the network. The input signal flows through the network, producing an output signal. The output signal produced is then compared with the desired output signal and the errors propagated backwards in the network. In this work we have adopted the BP sliding time window approach for our ANN models.

### A. *Importance of network architecture*

In ANNs the network architecture mainly refers to the hidden structure in a network. As the size of the input layer corresponds to the number of lagged observations used to discover the underlying pattern in a time series and the output layer corresponds to the forecasting horizon, the adjustable part in the ANN is the number and composition of the hidden layer. The number of hidden neurons and layers play very important roles in ANNs. It is the hidden neurons in the hidden layers that allow ANNs to detect the feature, to capture the pattern in the data, to perform complicated nonlinear mapping between input and output variables, to delineate underlying relationships and structures inherent in a particular dataset and more importantly, it is the hidden neurons in the hidden layers that determine the generalization ability of ANNs [9, 10]. For a given data set there may be an infinite number of network architectures relevant to learn the characteristics of the data. It has been experimentally proven that, a neural network with too small a hidden layer is unable to learn the input-output mapping, however, if too many hidden neurons are used, overfitting can occur. The choice of an appropriate number of hidden layer neurons is not a straightforward task, to date there has been no simple clear-cut method for determination of this parameter. Several heuristics to determine the optimal architecture have been suggested by various authors [11, 12]. However none of these heuristics have been universally accepted and ultimately, the selection of the most appropriate architecture comes down to trial and error.

A common sentiment shared by many ANN practioners is that the number of neurons in the hidden layer(s) should be large enough for the correct representation of the problem, but low enough to have adequate generalization capabilities. Most of these authors tend to apply some form of Occam's razor, the essence of which is that all things being equal, the

simplest solution tends to work the best. This approach although favored by some in the ANN research community, is emphatically rejected by others. For instance authors such as Gorman [13] and Mather [14], report a decrease in the general performance of their networks with an increase in the number of hidden neurons. Equally important is the number of hidden layers. Reports coming from Hornik [15] indicate that a single hidden layer is generally sufficient for most ANNs to achieve good generalization ability, particularly for forecasting tasks, however this is disputed by Cybenko [16] who insist that a second hidden layer does offer some added performance benefits. In one of their experiments Siestma and Dow [17] use 2 hidden layers in their network, they claim this resulted in a more compact architecture, which achieved a higher generalization and accuracy performance than the single hidden layer network which they had used earlier on. Zhang [18] found that networks with 2 hidden layers can model the underlying data better than 1 hidden layer networks for a particle time series. They also tried networks with more than 2 hidden layers but did not find any significant improvement. These results support the conclusion made by Chester [19] that a network never needs more than two hidden layers to solve most problems including forecasting. Other authors such as Baum and Haussler [20] have even gone on to suggest that the size of an ANN has relatively negligible effects on its generalization ability. Literature quite clearly indicates contrasting opinions in as far as the optimal architecture of ANNs is concerned. It is critical that this issue is further explored in a much more empirical and systematic manner. This is an area if fully investigated has potential for making significant strides in as far as the overall performance, particularly the generalization ability of ANNs is concerned.

## III. MATERIALS AND METHODS

In our approach for the study we used experimental method which is a proven method for testing and exploring cause and effect relationships. The benefit of using this method is that it allows the control of variables thereby enabling the isolation of a particular variable to observe effects on other variables [21]. This further acts as a light in the tunnel that guides us towards making conclusions about cause and effects of variables. The software used for the purposes of this study is Matlab Version 7.4.0.287 (R2007a). Matlab is an application software and programming language with interfaces to Java, C/C++ and FORTRAN. In this study, Matlab provides an environment for creating programs with built-in functions for performance metrics and forecasting using its Neural Networks toolbox Version 5.0.2 (R2007a). The toolbox provides comprehensive support for many proven paradigms, as well as Graphical User Interfaces (GUIs) that enable one to design and manage ANNs. The computer used to conduct this study is an Intel(R) Core(TM) 2CPU6300@1.86GHz. The data for this study was collected from the South African Tertiary Institutions Network (TENET) website (www.TENET.ac.za).

We analysed network traffic data which comprised inbound traffic in (bits/ sec) from the University of Fort Hare VC Alice Boardroom 101 – Fa 0/1 router. The data spanned from the 1st of March 2010 from 02:00 hours to the 21st of September 2013 02:00 hours in daily intervals, equating to 700 observations. As in all practical applications the data suffered from several deficiencies that needed to be remedied before use for ANN training. Preprocessing was done which included interpolation to fill in missing or null values, which amounted to 7 such observations. Matlab Neural Network toolbox has a built-in function, *mapminmax* which scales the data down before training so that it has 0 mean and unity standard deviation and then scales it up again after training, so as to produce outputs with 0 mean and unity standard deviation. We partitioned the data into train and test sets. The first 530 samples were allocated to the training set and the remaining 200 samples to the test set. On visual inspection of the time series a sliding time window of size 1 was arbitrarily chosen as input into the ANNs, with the pattern of outputs being the network targets. For maximum compatibility of results in all experiments an input layer of 1 neuron corresponding to the length of the sliding time window and an output layer of 1 neuron were used for the ANNs, since were interested in predicting 1 day ahead. Also the logistic sigmoid (logsig) and linear activation (purelin) functions were used in the hidden and output layers of the ANNs respectively. The BP Levenberg–Marquardt algorithm (trainlm) was the standard training algorithm throughout the duration of our investigations. Maier and Dandy [22] carried out a similar case study, a learning rate of 0.1 in conjunction with a momentum value of 0.6 was found to give good results and hence was used for all their models. In our study we adopted a similar approach and a learning rate of 0.1 and momentum of 0.6 was used in all our models. Training was conducted iteratively and in all the experiments weights were randomly initialized between [-0.5, 0.5]. Training was stopped after 1000 epochs and the performance of the networks tested by presenting the test set to the networks and calculating the Root Mean Squared Error (RMSE) between the actual and predicted values. RMSE is a dimensionless value calculated to compare ANN performance. The RMSE of the training set ($MSE_{tr}$) and the RMSE of the testing set ($MSE_{te}$) were calculated using the following equations:

$$\boldsymbol{RMSE_{tr} = \sum_{p=1}^{P_{tr}} (d_p - o_p)^2} \quad \textbf{(4)}$$

$$\boldsymbol{RMSE_{te} = \sum_{p=1}^{P_{te}} (d_p - o_p)^2} \quad \textbf{(5)}$$

where $d_p$ is the desired output for each input pattern and $o_p$ is the actual output produced by the ANN. Inorder to minimize the effect of the initial weights on training, for each experiment conducted, 10 training-testing runs were made and the results averaged. The size of the training set and test set were fixed during the entire investigations. We also ensured that all other variables that could potentially affect the quality of results remain constant. Hence throughout the duration of our investigations the learning rate, momentum and training algorithm also remained fixed. The primary evaluation criteria chosen for this study is the generalization ability (Gen). Generalization is a major indicator of a network performance as it shows how well a network can perform on data never encountered before. Therefore, the goal of the analysis was to evaluate all the network architectures and select the architecture that can model the data with optimal generalization performance. The most generalized architecture would have low generalization errors i.e. it should have close train and test error values [23]. To evaluate the generalization abilities of the architectures, the difference between RMSE values for training and testing were compared as follows:

$$\boldsymbol{Gen = mRMSE_{test} - mRMSE_{train}} \quad \textbf{(6)}$$

In addition to that, 2 other performance evaluation criteria were used. The correlation statistic (R) selected to measure the linear correlation between the actual and the predicted traffic. The optimal R value is unity and a value smaller than 0.8 is assumed to be problematic. To estimate the efficiency of the fit, the Coefficient of determination also known as the $R^2$ criterion is used. The optimum $R^2$ value is unity and a value smaller than 0.7 corresponds to a very poor fit. Architectures were labeled as p:q:r:s or p:q:s depending on the number of hidden layers, where q and r are the number of neurons in the hidden layers and p and s are the number of neurons in the input and output layers respectively.

## IV. THE EXPERIMENTS

### A. *Number of neurons in a single hidden layer.*

The first set of experiments examined the effects of using a single hidden layer on ANN performance. We trained the network using various numbers of hidden neurons in a single hidden layer. The number of hidden neurons examined was 10, 20, 30, 40, 50 and 60. The results for the experiments are shown in Figure 3.

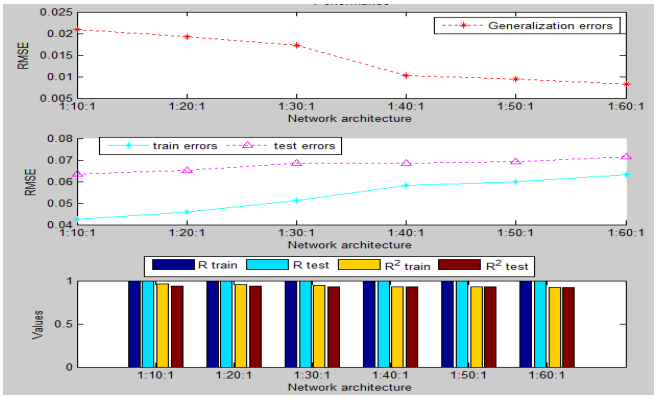



**Figure 3**: Results for various architectures.

### B. *Different ratios of first to second hidden layers neurons.*

In this series of tests, the effect of using different ratios of first to second hidden layer neurons on the performance of ANNs was assessed. The number of neurons in the first hidden layer was kept constant at 45 neurons, whilst the number of neurons used in the second hidden layer were varied according to 5, 10, 15, 20, 25 and 30. Figure 4 shows the results of the experiments.

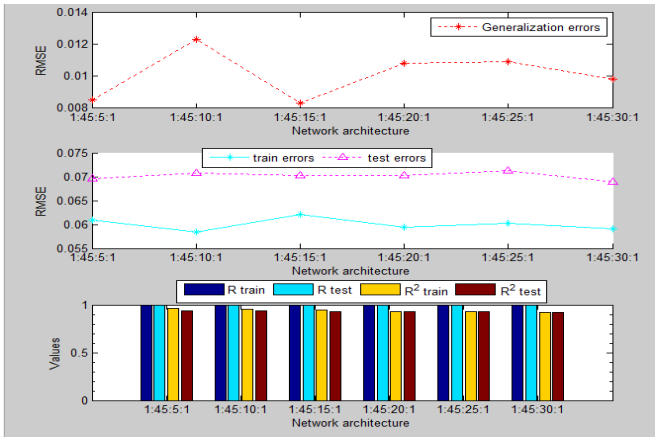



**Figure 4:** Results for various architectures**.**

### C. *Number of neurons in the first and second hidden layers.*

Kudrycki [24], found empirically that the optimum ratio of first to second hidden layer neurons is 3:1, even for high dimensional inputs, following such convention we maintained a ratio of first to second hidden neurons of 3:1. We sought to explore how this ratio affects the performance of ANNs. The number of neurons in the first hidden layer was varied according to 15, 30, 45, 60 and 75. Consequently the number of neurons in the second hidden layer was varied according to 5, 10, 15, 20 and 25. The results for the experiments are shown in Figure 5.

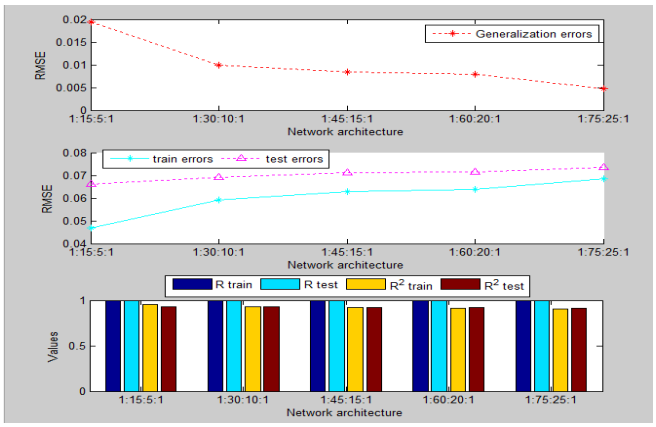



**Figure 5:** Results for various architectures.

### D. *Three hidden layers.*

Although literature does not indicate any practical or theoretical basis for a need of a third hidden layer, out of curiosity we examined how a third hidden layer could possibly impact on the performance of an ANN trained to predict trends in TCP/IP network traffic. We explored various 3 hidden layer architectures, we kept the same number of neurons in each of the hidden layers. The number of neurons examined was varied according to 15, 30, 45, 60 and 75. Figure 6 shows the results for the various experiments conducted.

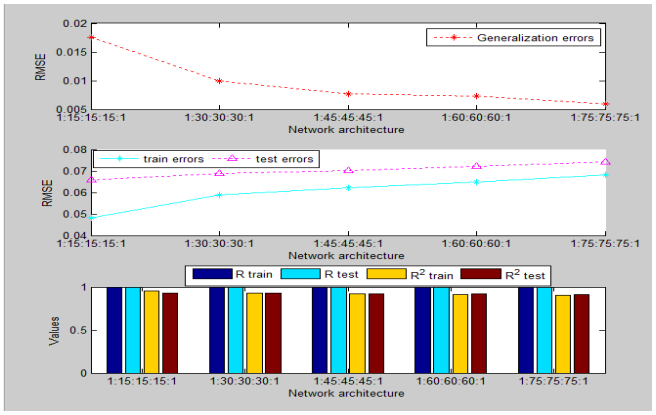



**Figure 6:** Results for various architectures**.**

### E. *1, 2 and 3 hidden layers.*

To give an overall picture of how the configuration of the network architecture affects the performance of ANNs we compared the performance of 1, 2 and 3 hidden layer

network architectures. Figure 7 shows the results from the experiments.

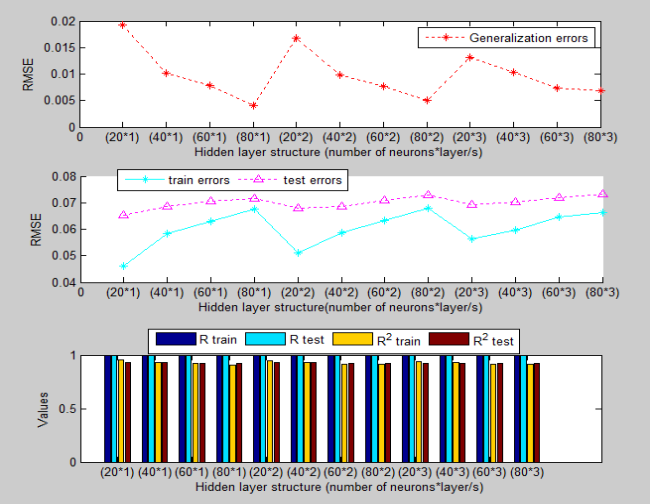



**Figure 7:** Results for various architectures.

## V. RESULTS AND DISCUSSIONS

To assess the generalization performance of different ANN architectures in the task of forecasting a TCP/IP network traffic time series, we conducted several experiments varying the numbers of hidden neurons and/or layers. We begin by examining the results exhibited by a single hidden layer. From Figure 3, note that with an increase in number of hidden neurons there is a corresponding increase in generalization ability of the network. Quite interestingly, this is in sharp contrast to Occam's razor principle which would have us believe that the network with 10 hidden neurons, since it's the smallest is bound to have better generalization ability than the rest. That seems not to be the case as the network with 60 hidden neurons, which is the largest, acheived the lowest generalization errors. However a network with architectures of between 40 to 50 hidden neurons performed comparatively similar to the 60 hidden neurons. The comparable levels of performance are even made more striking by examination of the train and test errors. From 40 hidden neurons the networks exhibit virtually no change in either train or test errors. Analysis of the statistical measures, show relatively good performance for all the network architectures examined. Values of R and ($R^2$) above 0.8 on both train and test sets indicate a good linear correlation and good fit between the network activations (predicted values) and network targets (actual values).

Turning to the results produced by 2 hidden layers, Figure 4 shows that using different ratios of first to second hidden neurons does not seem to have any significant impact on the generalization ability of the networks. No specific trend can be visualized as the generalization errors between the various architectures are largely fluctuating. This is inspite of the minimum generalization error recorded when 15 neurons were used in the second hidden layer. This corresponds to a ratio of 3:1 between the first and second hidden layer neurons which affirms the claim made by Kudrycki [24]. Figure 5 shows that when a ratio of 3:1 is maintained between the first and second hidden neurons, an increase in the size of the network is accompanied by an increase in the generalization ability of the networks. In Figure 3, the best generalization performance was recorded when 25 neurons were used in the first hidden layer and 75 in the second hidden layer. Results from Figure 4 and 5 indicate that with 2 hidden layers an ANN is most likely guaranteed to achieve a reasonably positive linear correlation and good fit between the predicted and observed values. The use of 3 hidden layers as shown in Figure 6 did not have any significant effect on the network's performance; in fact it produced slightly worse results than the smaller networks. It is interesting to note that the generalization error of the best performing architecture when 2 hidden layers are used i.e. (1:75:25) is equal to the best performance when 3 hidden layers are used i.e. (1:75:75:75:1), indicating that the use of a third hidden layer really has no significant impact on the generalization ability of ANNs, putting into effect the claim made by Cybenko [16] that 2 hidden layers are sufficient to approximate any function and that additional layers would be computationally redundant. Figure 7 summarizes the effects of different layers and neurons on the performance of ANNs. It indicates that, based on the generalization errors, compared to the 1 hidden-layer model, almost all 2 hidden-layer and 3 hidden-layer models did not perform significantly better. As Figure 5 depicts, the generalization errors achieved by 1 hidden layer architectures were in most instances better than 2 and 3 hidden layers considering the low amount of neurons. Nevertheless, this condition changes when the number of neurons in a single hidden layer is extremely low, which confirms the findings of Zhang [18] who proposed that simple network models are often adequate for forecasting linear time series. In his study, one hidden neuron was sufficient for optimal generalization and accurate forecasting. Nevertheless, the number of hidden neurons in his study was from 1 to 10, which is rather limited compared to our study which explored architectures as huge as 80 neurons in 3 hidden layers.

## VI. CONCLUSIONS

In this study the determination of the optimal network architecture for the prediction of TCP/IP network traffic has been investigated. The experimental results regarding the network architecture and generalization exhibited by ANNs trained on a fixed training set size indicate comparative levels of performance across a broad range of hidden units. Networks having fewer hidden units did not generalize as well as networks having more hidden units, whilst networks having 2 or 3 hidden layers did not perform any significantly better than networks having 1 hidden layer. The following conclusions can be drawn from the results:

- Small networks learn tasks more quickly, but not necessarily better.
- Increasing the number of neurons improves the performance of the system. Nevertheless, too many neurons decrease the generalization ability of the model due to over-memorizing. It means that too many or lesser amount of neurons in the model is detrimental to the generalization ability of ANNs.
- Although the size of the ANN affects the generalization ability of a network, it has little or no influence on the goodness of fit and linear correlation of the models. In our experiments the R and $R^2$ values for the train and test sets indicate a good fit and linear correlation between the network

targets and activations despite the size of the network examined.

- When the amount of neurons in the models is the same and less, the performance of l-hidden layer is better than that of 2 and 3 hidden layer models.
- Hecht and Nielsen [25] provided a proof that a single hidden layer of neurons, operating a sigmoidal activation function is sufficient to model any solution surface of practical interest. Our results seem to support their conclusion and further reaffirm the assertion made by Hush [26] regarding the sensitivity of ANNs to the number of hidden layers.
- Optimal network architecture is highly problem-dependent.

Perhaps, the most important conclusion derived from this study is that large networks do not always improve the generalization ability of ANNs. A network that is large enough to learn the characteristics of the data is usually sufficient; under this setup we recommend a single hidden layer with roughly 50 hidden neurons as being sufficient from a generalization view point. We recommend that for TCP/IP forecasting tasks one should start with a small number of hidden neurons, preferably 10 and slightly increase the size of the network until no further improvement in generalization ability of the network is achieved. Although this number is not ideal for all situations, it could be used as a starting point for the search towards the optimum number of hidden layer neurons. We are not really keen on the idea of a second hidden layer but should one for any reason feel they should use a second hidden layer, we would suggest they use a ratio of first to second hidden neurons of 3:1 and a network architecture of (1:60:20:1) should warrant good generalization performance. We refute the idea of using 3 hidden layers, it does not improve generalization performance by any standards and it is a total computational liability. We support the assertion made by Judd [27] that for a given number of hidden units it is better to contain the units in a single hidden layer than to distribute them over 2 or more layers. In our view keeping the number of neurons to a minimum has other several advantages as it (a) reduces the computational time needed for training (b) helps avoid overfitting (c) allows the trained network to be analysed easily. It is also vital to note that several factors, such as learning parameter, number of iterations, transfer function and the characteristics of the data, play very important role to get a network with high generalization capabilities. Investigating the effects of these factors would be very useful to understand the behaviour of ANNs.

## VII. REFERENCES


[1]. G. Box and G. Jenkins (1976), Time Series Analysis, Forecasting and Control. Holden Day, San Francisco, USA, 1976.

[2]. M.H. Hassoun (1995). Fundamentals of Artificial Neural Networks, MIT Press, Cambridge, 1995.

[3] A. Sang and S. Li (2002), A predictability analysis of network traffic, Computer Networks, vol. 39, no. 4, pp. 329.345, 2002.

[4]. P. Cortez, M. Rocha, J. Machado, and J. Neves(1995), A Neural Network Based Forecasting System, in Proceedings of IEEE ICNN'95, vol. 5, Perth, Australia, Nov. 1995, pp. 2689.2693.

[5]. Y.A.Hashash, S. Jung, and J.Ghaboussi (2004), Numerical implementation of a Neural Network based material model in finite element analysis, International Journal for Numerical Methods in Engineering, 59(7), 989-1005.

[6]. S.Haykin (1994), A Comprehensive Foundation, MacMillin College Publishing CO, New York.1994.

[7]. P. Cortez (2005), Time series forecasting by evolutionary neural networks, Chapter on Artificial Neural Networks, Idea Group Publishing, pp. 47–70, 2005.

[8]. E.Gately (1996), Neural Networks for Financial Forecasting. John Wiley, New York, 1996.

[9]. J. Moody (1992), The effective number of parameters, An analysis of generalization and regularization learning systems, in J. Moody, S.J. Hanson, and R.P. Lippmann, eds., Advances in Neural Information Processing, 4:847–854, 1992.

[10]. B. Ripley (1995), Statistical ideas for selecting network architectures. Invited Presentation, Neural Information Processing Systems 8, 1995

[11]. L.V Fausett (1994), Fundamentals neural networks: Architecture, algorithms, and applications, Prentice-Hall, Englewood Cliffs, New Jersey, 1994.

[12]. J. Faraway and C. Chatfield (1998), Time series forecasting with neural networks, A comparative study using the airline data. Applied Statistics, 47(2), 231-250, 1998.

[13] R. Gorman (1988), Analysis of hidden units in a layered network trained to classify sonar targets.1 (1): pp 75-88, 1988.

[14] P.M. Mather (1998), Assessing artificial neural network pruning algorithm, in Proceedings of the 24th Annual Conference and Exhibition of the Remote Sensing Society (RSS'98), pp. 603-609.

[15] K.Hornik, M. Stinchcombe and H. White (1989), Multilayer feed-forward networks are universal approximators, Neural Networks, 2, 359-366.

[16] G. Cybenko (1989), Approximation by superpositions of a sigmoidal function. Mathematical Control Signals Systems 2, 303–314, 1989.

[17] J.Setsma (1991), Creating neural networks that generalize, Neural networks 4(1): pp 67-79, 1991.

[18] X.Zhang (1994), Time series analysis and prediction by neural networks, Optimization Methods and Software 4, 151–170, 1994.

[19]. D. L .Chester(1990), Why two hidden layers are better than one, International Joint Conference on Neural Networks, 1, 265-268,1990.

[20]. E.B. Baum, D. Haussler (1989), What size net gives valid generalization? Neural Computation 1, 151–160, 1989.

[21]. L.R. Gay, *Educational research (4th Ed.)*, New York, Merrill, pg 298 1992.

[22] H. R Maier and G.C Dandy (2000), Neural networks for the prediction and forecasting of water resources variables, A review of modeling issues and applications, Environmental Modeling & Software, 15(2000), 101-12.

[23]. L. Gorr (1994), Research prospective on neural network forecasting. International Journal of Forecasting 10, 1–4, 1994.

[24]. T.P. Kudrycki (1988), Neural network implementation of a medical diagnosis expert system, MS thesis, College of Engineering, University of Cincinnati, 1988.

[25] R.Hecht-Nielsen (1990), Neurocomputing, Addison-Wesely Publishing Company, Reading, MA, 1990.

[26]. D. Hush, G.Zeng, N Ahmed (1989), An application of neural net in decoding error correcting nodes, in Proceedings of IEEE ICNN'89 Int Symp, pp782-785.

[27]. J .Judd, Neural Network Design and Complexity of Learning. MIT Press, Cambridge Massachusetts, 1990.


## VIII. ACKNOWLEDGEMENTS


This work is based on the research undertaken within the TELKOM Coe in ICTD supported in part by Telkom SA, Tellabs, SAAB Grintek Technologies, Eastell, Khula Holdings, THRIP, GMRDC and National Research Foundation of South Africa (UID: 86108). The opinions, findings and conclusions or recommendations expressed here are those of the authors and none of the above sponsors accepts liability whatsoever in this regard.



**Vusumuzi Moyo** received his B.Sc Honours (Computer science) from the University of Fort Hare and is presently studying towards his Master of Science degree at the same institution. His research interests include artificial intelligence and neural networks.

# STANDARDS, REGULATORY & ENVIROMENTALS

# Prototyping Machine-to-Machine Applications for Emerging Smart Cities in Developing Countries

Joyce Mwangama[1], Asma Elmangoush[2], Joseph Orimolade[1], Neco Ventura[1], Ronald Steinke[2], Alexander Willner[2], Andreea Corici[3], Thomas Magedanz[3]
[1] Communications Research Group, University of Cape Town, Cape Town, 7701, South Africa
Email: {joycebm, funsho, neco}@crg.ee.uct.ac.za
[2] Technische Universität Berlin, Marchstraße 23, 10587 Berlin, Germany
Email: asma.a.elmangoush@campus.tu-berlin.de,
ronald.steinke@mailbox.tu-berlin.de, willner@av.tu-berlin.de
[3] Fraunhofer FOKUS Research Institute, Kaiserin-Augusta-Allee 31, 10589 Berlin, Germany
Email: {andreea.ancuta.corici, thomas.magedanz}@fokus.fraunhofer.de,

**Abstract – Urbanisation trends in both developed and developing countries continue to grow significantly. This presents an opportunity for developing countries to increase future economic growth if they invest in the development of Smart Cities. One aspect of the realisation of Smart Cities revolves around the concept of using information and communication technologies to enhance the quality of living for citizens. This has led to the subject of Smart Cities becoming an important ongoing research topic in both academia and industry. Even as Smart Cities present lucrative and exciting opportunities for the developing world, many challenges stand in the way before they can actually be realised.**

**This work tackles the first steps of moving towards this objective by identifying potential use cases for the Smart City context in developing countries such as South Africa. The focus is on a common problem faced by many such developing countries which is the lack of reliable and efficient management of power distribution. This paper introduces a framework that will allow for the development of Smart City energy management applications. The proposed architecture is based on a Smart City platform and an ETSI M2M/ oneM2M compliant Machine-to-Machine (M2M) communication framework.**

**Index Terms— Smart City; M2M; Smart Home; Testbed, Mobile Applications**

## I. INTRODUCTION

The developing world is experiencing an unprecedented migration pattern that has resulted in mass populations moving from rural areas into urban environments or cities. By the year 2050 it is estimated that more than two thirds of the global population will be inhabiting cities [1]; this is going to place a large strain on these cities to be able to offer services and infrastructure that can cope with these populations. The strain will be felt more so in developing countries as water, sanitation, electricity, waste, transportation, communications, housing and food security will need to be managed efficiently or "in a smart manner".

This brings forward the notion of smart cities. Smart cities can be defined as follows according to Cohen [2]: "Smart Cities use information and communication technologies (ICT) to be more intelligent and efficient in the use of resources, resulting in cost and energy savings, improved service delivery and quality of life, and reduced environmental footprint all supporting innovation and the low-carbon economy." Within the developing world context for example, Accra's definition of a smart city [3] (Ghana): "A Smarter City is one that accelerates its journey towards sustainable prosperity by making use of new smart solutions and management practices. As one of Africa's fastest urbanising cities, in one of the world's fastest growing economies, Accra, Ghana has unprecedented opportunity to use transformative technologies as the foundation for future growth and development. From transportation, through water, sanitation, healthcare, energy to city management and public safety – there are multiple urban systems which technology can help to address holistically."

Future economic growth will be predominantly centred on cities. It therefore becomes crucial to begin to prepare current cities towards the future direction of planning and developing for smart city realisation. The realisation of smart city implementations will involve the integration of various domains within a Smart City concept, for example communication infrastructure; data storage and management; and analytics, to be able to run an effective smart city management platform.

This work tackles the first steps of moving towards this objective by the implementation of use cases by means of application development within a smart city prototyping testbed. We identify the common problem faced in developing countries which is energy shortage and subsequent management and propose possible solutions that can be developed to be utilised in future smart cities of these developing nations.

The objective of this paper is to present the challenges faced by developing cities as they move towards being Smart Cities. Some scenarios are highlighted that demonstrate the potential positive impact of research and development in this area. We also highlight the ongoing research project involving both industry and academia stakeholders in both developed and developing countries towards the realisation of smart city application prototyping. The rest of this paper is organized as follows: in section II, we present the identified requirements and scenarios that show the potential benefits of Smart City application. In section III, we overview the concept of Smart City and related work in this area. In section IV we present the implementation environment for proposed architectures of

Smart Cities. Section IV showcases the development of example Smart City application for future energy management. Finally, conclusion and further work is provided in section V

## II. Requirements and Scenarios

It is important to understand the requirements of the Smart Cities of developing countries, as these will differ somewhat from those of developed nations. We further describe some potential scenarios to highlight the benefits expected to obtain in future Smart Cities.

### A. Requirements

The following are key requirements that all smart cities must aim to achieve.

- Utilisation of existing underlying communication infrastructure;
- Cost effective data storage and management;
- Systematic computational analysis of data or statistics;
- And interoperability of connected systems and services.

#### a) Interconnection of multiple technology domains

In a Smart City, many different subsystems need to work together such that the Smart City system performs as intended. These integration points include the communication network, the internet, sensors, devices, gateways, and the resources or services of the Smart City. These different technologies, or Machine to Machine enabling technologies, will work together to enable the Smart City.

#### b) Integration of existing city information systems

Within a city already many sources of information exists. A Smart City will provide data aggregation and analysis tools to use this information in creating novel new applications for the benefit of city inhabitants. Machine-to-Machine (M2M) enabling technologies will be used in every business field and integrate application systems, data and Internet to be the core elements supporting urban operation and management. City support systems such as “energy”, “transport”, “safety”, “health”, “education” and “water”, stand poised to benefit from the Smart City enabling framework. It is clear that, taking into account the similarities and differences of each management domain there is a value in developing a global, standardised framework for Smart Cities.

#### c) Overcome challenges inherent in the developing world context

In the developing world, many other obstacles are faced that are different from those in the developed world. Most important is the availability of affordable and reliable connectivity, as connectivity can be expensive and limited. In a Smart City, connectivity is an important requirement. This means that when developing applications for smart cities in the developing world this limitation needs to be built into operational requirements. Other challenges include inefficient transportation systems, prevalence of informal settlements, high rates of service delivery protests which would halt delivery of city services, and energy related challenges such as power shortages resulting in frequent power cuts or load shedding situations. Figure 1 depicts such a real world example of load shedding due to the national demand exceeding the national capacity of South Africa.

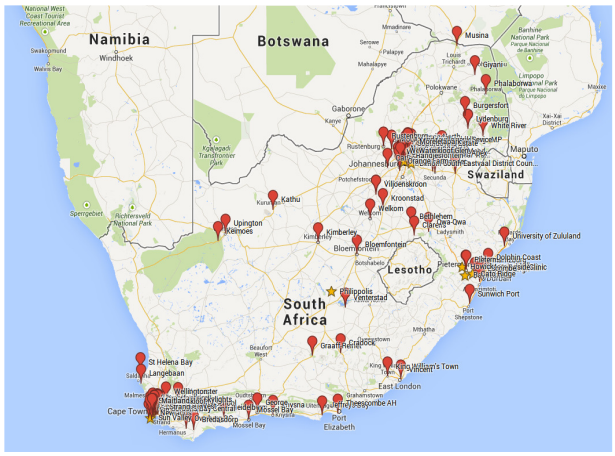



Figure 1: South African National Load shedding on 6 February 2014

### B. Potential Scenarios

Scenarios for the development of applications in Smart Cities are presented here below focusing on energy management. Many others exist for various Smart City domains, but these are not mentioned in this specific paper.

#### a) Simplified energy management by utility providers

Utility providers in the energy sector aim to provide uninterrupted power supply to all of the electrified population. It is therefore essential to always insure that demand is met in order to “keep the lights on”. When resources are scarce, smart management systems should enable the utility providers to employ some emergency procedures that mitigate the need to go into blackouts or load shedding modes. Ultimately, factors such as atypical weather or time of the year place heavy demands on municipalities, industries and home-owners.

For example, if a utility provider has an unscheduled maintenance on one of its power generation substations; this will result in an unstable power grid due to decreases in generation capacity. If the utility provider is able to access an aggregated view of both industrial and residential areas, they could have the ability to identify areas and entities of high utilisation. Furthermore, the utility provider could be able to notify targeted individuals who may be drawing higher than normal energy resources through a demand management signal to reduce consumption, rather than declaring a general call to the entire population to reduce consumption. Additionally, for critical situations, utility providers can send a command to devices that previously have been nominated by their owners as “controllable”. This emergency intervention has the potential to achieve the desired outcome of the grid returning to a stable state.

#### b) Simplified energy management by individuals

Giving energy consumers the ability to get information of their households or business energy consumption can also

be provided as a Smart City application. During the case when (as mentioned in the above scenario) a user is notified to reduce consumption because he/she has left, for example, the lights/hot water boiler/pool pump on, the user is then able to act accordingly. It may even be the case that the user is not on their respective premises, but still receives this notification, and is still able to act on it as the user has the ability to control devices remotely. The user is then able to assist the utility provider to mitigate a power outage situation from any location.

Additionally, from a safety point of view, users would be able to monitor and additionally be alerted if, for example, a heater has been left on whilst no one is at home. This Smart City energy scenario allows for many other applications at the disposal of the energy users in a Smart City.

#### *c) Energy management toward "Green Living"*

A further application of a robust energy monitoring and management scenario involves the drive towards achieving higher levels of energy usage efficiency. Smart Cities also aim to be Green; a main objective is considered to be the reduction of energy resource utilisations. Users of such a system may be environmentally conscious individuals who would like to use energy more efficiently. They would then have the ability to get a better understanding of their consumption patterns in order to make better and more informed decisions regarding their energy usage. The application is able to further process these usage statistics and formulate detailed analysis of consumption even to the benefit of being able to adapt the behaviour of users such that monthly utility bills would be reduced.

#### *d) Additional smart management scenarios*

**Energy management for service disruptions:** for situations where power outages are inevitable, utility providers can notify users that opt in to receive information about imminent unplanned power outages. This would enable users the ability to take action such as taking sensitive equipment offline to avoid damages.

**Energy management towards Smart Home automation:** for situations where users have the access to smart home control systems, these can be interfaced with the Smart City application such that the management and control of devices is automated without the necessity of human intervention. The smart home system can automatically make decisions and take actions based on notifications received from the utility provider, if users are for some reason not reachable or able to act.

**Energy management towards cost reductions:** the inclusion of additional features such as real time billing and energy reports incorporated into the Smart City application, allowing the ability to provide contextual information comparing individual usage to average usage of the surrounding area, users can be incentivised to adapt behaviour even further to reduce on bills.

## III. EXISTING SOLUTIONS

Even though Smart Cities are widely considered as a hot topic in academia and industry, there is no standardised definition of the Smart City concept among practitioners. Authors in [4] reviewed several working definitions and proposed a general Smart City framework based on eight factors: "management and organization, technology, governance, policy context, people and communities, economy, built infrastructure, and natural environment". The instrumentation of Smart Cities is considered as a key enabler that will leverage the understanding of the City operations by "making the invisible visible" [5].

The main goal of M2M platforms is to connect the growing number of devices, and associate them to a set of services addressing use cases from different industrial domains such as energy, automotive, health, transportation etc. The need to exchange information between actors at different domains in a Smart City motivates the need of an M2M middleware that mediates the communication between these systems. Developing a large-scale Smart environment, based on M2M communication, demands interoperability at all communication layers between devices, gateways, and services. However, most of existing M2M solutions are not interoperable and have been built in a highly vertical fashion, where data gathered by one platform can't be easily reused by others. A middleware M2M platform is required in Smart City infrastructure to provide:

- Reliable transportation and session control.
- Secure access to privacy-sensitive information.
- Standard open interfaces toward service layer.
- Efficient data/event processing methodology.
- Ease of participation and application development.

Building a Smart City requires the collaboration of various stockholders, to increase the efficiency of administrative services, and developing environment-friendly applications. Several works deal with Smart Cities frameworks and related issues. A good overview of Smart Cities initial examples and collaboration models is provided by [6]. The main technologies of interest in Smart City developments are: i) content fusion technologies to enable the collaboration between stockholders, ii) cloud service for federating all components, iii) scalable content management tools, and iv) intelligent high level solutions that use advanced sensors in an efficient manner.

Smart City principles in a developed world have been explored, and implementations are being realised (e.g. SmartSantander [7]). In a developing world, even though the potential of introducing the concepts are clearly evident, Smart City concepts have yet to make the required breakthrough.

## IV. IMPLEMENTATION ENVIRONMENT

The Smart City project, entitled "Testbeds for Reliable Smart City Machine-to-Machine Communication" [8][9], aims to address Smart and Green Cities challenges within underdeveloped countries. In this section, we describe the reference architecture. The overall architecture is presented in Fig. 2, which was defined to fulfil the following objectives:

- Deliver a specification of the overall architecture that involves an M2M communication platform [10] used as the basis for a Smart City platform.
- Interweave standard-based M2M platform with other sophisticated Smart City platform.

- Integrate resource-constrained devices over Delay Tolerant Networks (DTN).
- Perform the integration of the main building blocks (M2M, Smart City, Smart Energy) into a comprehensive platform using federation tools [11].
- Define specific enhancements for a Smart/Green City system, by implementing one pilot for Smart Energy consumption in the region Gauteng (South Africa) and one pilot for pollution monitoring in San Vicenç dels Horts (Spain).

The OpenMTC platform is used in the implementation and testbeds, which is a M2M platform compatible with ETSI M2M standards [12][13]. The platform is developed jointly by Fraunhofer FOKUS Institute and the Technical University Berlin (TUB) to act as a horizontal convergence layer supporting multiple vertical application domains such as logistics, automotive, energy, eHealth, etc. It provides open-standard interfaces that facilitate interweaving with other Smart City platforms. This will be validated by the implementation of the Smart Energy system in the region Gauteng (South Africa). Results obtained from this trial will be discussed in future publications.

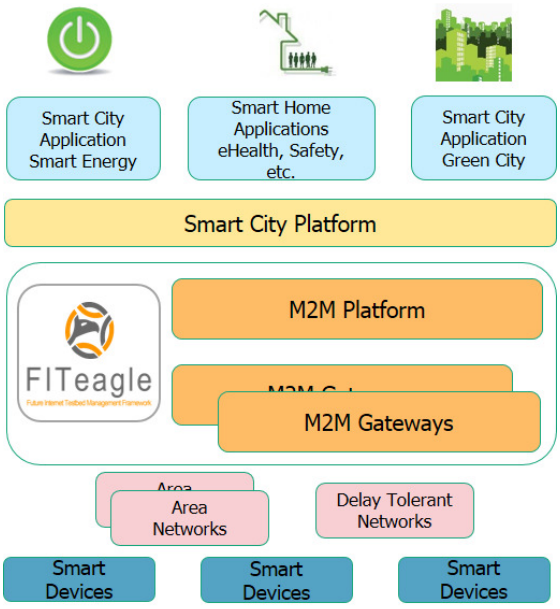



Figure 2: Smart City Reference Architecture

## V. Realisation of Example Scenario

The realisation of the following smart home application will demonstrate the OpenMTC functionalities, which monitors the energy production and consumption in an arbitrary number of buildings at a certain area in the city. Each device in a building is connected through an OpenMTC home gateway component. In the system, the user can zoom in on a particular building and discover the energy consumption of individual appliances and possibly to control them.

Technical components of the M2M Smart City system include the OpenMTC M2M software which interconnects and coordinates with an FS20 wireless smart home control system. Some of the actual devices in the smart home system include:

- A radio switching socket plug with the capability of control through the FS20 radio communication protocol
- A remote control that is able to be programmed to control the individual sockets and by proxy the devices connected through these sockets
- USB wireless transceiver that is able to act as a Gateway Interworking Proxy (GIP) enabler that allows for communication between the smart home devices and the OpenMTC M2M middleware software.

The OpenMTC platform runs both a Network Service Capability Layer (NCSL) and a Gateway Service Capability Layer (GSCL). The platform is designed for interworking with actual sensors and actuators (e.g. using ZigBee, FS20 enabled smart devices). These functionalities are exposed using the mIa reference point (between the NSCL and the network applications to be implemented) and dIa reference point (between the GSCL and the device to which communications occurs).

These reference points' communication are implemented in REST request and responses using CRUD (Create, Retrieve, Update and Delete) commands. The libcurl libraries which is client-side URL transfer library supporting various protocols e.g. FTP, FTPS, HTTP, SCP etc. could be used to interact with the platform over stateless transport protocols, like HTTP and Constrained Application Protocol (CoAP). CoAP is proposed to support essential features required for resource-constrained M2M devices, such as low overhead. The gateway interworking proxy (or GIP) is used for device interworking in the OpenMTC. The GIPs implemented are the ZigBee GIP, the FS20 GIP and HTMLv5 sensor GIP.

When developing a network or device application that allows for interworking with mobile devices, the HTMLv5 GIP exposes monitoring and control capabilities, this becomes handy if one wishes to develop mobile applications that can act on certain signals from the Smart City network applications alerts.

Figure 3 highlights the ETSI connected home example. This Smart Home system has the ability to offer a monitoring application for homes: i) monitoring the energy production and consumption of buildings in a certain area of the Smart City; ii) OpenMTC offers the integrated solution that allows home gateway devices to communicate with OpenMTC backend servers mediating the interactions between the network application and the home gateway; iii) allows for device access within the smart home using access technologies such as Zigbee or FS20. Figure 4 demonstrates how this is implemented within the OpenMTC M2M system.

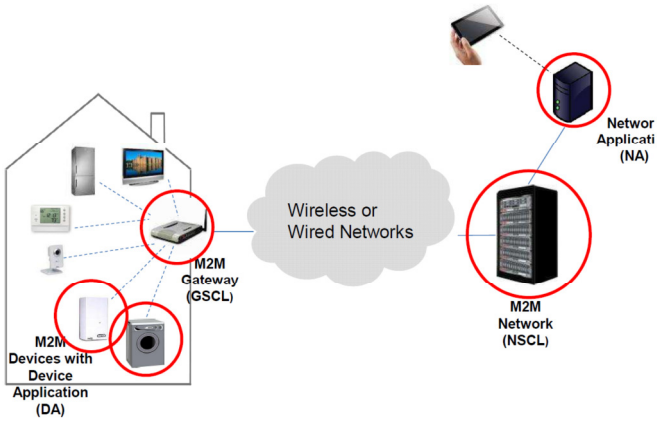


Figure 3: ETSI M2M connected home example

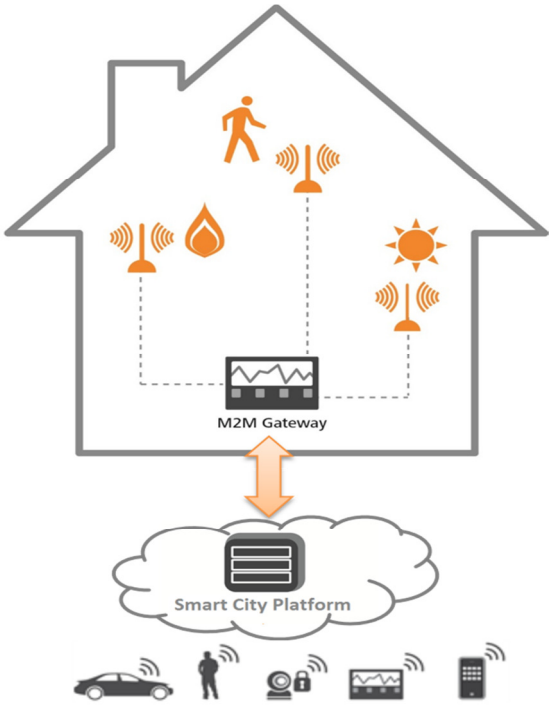


Figure 4: OpenMTC smart home model.

## VI. Conclusions and Future Work

Cities and communities worldwide are facing various challenges, due to increased populations and prospective economic growth. Furthermore, the connected world is extending exponentially including physical objects, computers and smartphones in a global Internet of Things (IoT). For Smart Cities to be successful, they will need to leverage the utilisation of past and future generations of Information, Communications and Technologies (ICT). Additionally, great benefit will be observed if the integration of different M2M enabling technologies with the different service and resource sectors of a Smart City is well planned. The collection and analytics of certain data and information will also enhance the quality of living for Smart City inhabitants. All these factors will work to achieve the goals of high efficiency of Smart City management, timely and convenient service delivery, reliable urban operations, green economy and comfortable living.

As interoperability is a very important issue, we believe that the architecture described here are based on the ETSI and oneM2M standards is a starting point when developing or deploying a smart energy management system. All types of access network might be used: from WiFi to GPRS, UMTS or LTE (for ensuring a low delay in communication). In the case of smart energy grids specific QoS policies managed by the Smart City application by interacting with the local telecommunication operators would be required to guarantee the network resources for time sensitive data.

The presented architecture is part of the research collaboration between partners from EU countries and South Africa during the European Union's Seventh Framework Programme funded project TRESCIMO [14]. In this project two pilots targeting both air quality monitoring in urban areas for the developed smart city and smart grid integration with smart city applications in the developing work Smart City will be deployed.

## Acknowledgements

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 611745, as well as the South African Department of Science and Technology under financial assistance agreement DST/CON 0247/2013.

## References

[1] H. Schaffers, N. Komninos, M. Pallot. "Cities – Smarter, Greener, Better." Scientific American Magazine, September 2011:

[2] B. Cohen, "The top 10 Smart Cities on the planet", Fast Company, www.fastcoexist.com/1679127/the-top-10-smart-cities-on-the-planet.

[3] IBM Ghana Ltd.,"A Vision for Smarter Growth, An IBM Smarter Cities report on Accra, Ghana", Apr 2013, IBM Media Relations, Growth Markets

[4] H. Chourabi, T. Nam, S. Walker, J. R. Gil-Garcia, S. Mellouli, K. Nahon, T. a. Pardo, and H. J. Scholl, "Understanding Smart Cities: An Integrative Framework," in 45th Hawaii International Conference on System Sciences, 2012, pp. 2289–2297.

[5] C. Harrison and I. A. Donnelly, "A Theory of Smart Cities," in Proceedings of the 55th Annual Meeting of the ISSS, 2011, pp. 1–15

[6] H. Schaffers, N. Komninos, M. Pallot, B. Trousse, M. Nilsson, and A. Oliveira, "Smart Cities and the Future Internet: Towards Cooperation Frameworks for Open Innovation," in The Future Internet, Springer Berlin Heidelberg, 2011, pp. 431–446

[7] L. Sanchez, L. Muñoz, J. A. Galache, P. Sotres, J. R. Santana, V. Gutierrez, R. Ramdhany, A. Gluhak, S. Krco, E. Theodoridis, and D. Pfister. "SmartSantander: IoT experimentation over a smart city testbed," in Journal on Computer Networks Special issue on Future Internet Testbeds –Volume 61, 14 March 2014, Pages 217–238

[8] J. Mwangama, A. Willner, N. Ventura, A. Elmangoush, T. Pfeifer, and T. Magedanz, "Testbeds for Reliable Smart City Machine-to-Machine Communication," in Southern African Telecommunication Networks and Applications Conference (SATNAC), 2013.

[9] A. Corici, A. Elmangoush, R. Steinke, T. Magedanz, J. Mwangama, and N. Ventura, "Utilizing M2M

technologies for Building Reliable Smart Cities" in SmartCity'2014 workshop at IEEE 6th International Conference on New Technologies, Mobility and Security (NTMS), 2014.
[10] OpenMTC platform." [Online]. Available: http://www.open-mtc.org/index.html.
[11] "FITeagle- Future Internet Testbed Experimentation and Management Framework." [Online]. Available: http://fiteagle.org/.
[12] ETSI TS 102 690 v1.1.1, "Machine-to-Machine communications (M2M); Functional architecture," 2011.
[13] ETSI TS 102 921 v1.1.1, "Machine-to-Machine communications (M2M); mIa, dIa and mId interfaces," 2012.
[14] "TRESCIMO Project." [Online]. Available: www.trescimo.eu.

**Joyce Mwangama** completed her BSc Eng. and MSc Eng. at the University of Cape Town in 2008 and 2011 respectively. She is currently working towards her PhD in Electrical Engineering in the Centre for Broadband Networks at the same institution. She is currently a Research and Teaching assistant within the electrical engineering department at UCT**.**

**Asma Elmangoush** completed her BSc Eng. And MSc Eng. at the Higher Institute of Industry (HII) –Libya. Afterwards, she joined the electrical engineering department at the Higher Institute of Industry as an Assistant Lecture. In October 2011 she joined the OpenMTC team with the chair of Next Generation Networks (AV) at the Technical University Berlin and works towards her PhD in electrical engineering. Her research focuses on machine-to-machine communication.

# Received power prediction of a terrestrial TV broadcasting transmitter using ordinary kriging interpolation

Willem H. Boshoff, Magdalena J. Grobler, Melvin Ferreira
TeleNet Research Group
School of Electrical, Electronic and Computer Engineering
North-West University, Potchefstroom Campus
Email: {21625158, leenta.grobler, melvin.ferreira}@nwu.ac.za

***Abstract* – Radio Environment Mapping (REM) provides information useful for many different applications in the telecommunications field. In this paper, the authors evaluate the ability of ordinary kriging (OK) to produce a REM without the large number of samples typically required. The kriging model is generated using different sizes of received power input sample sets over a 5 km radius surrounding a transmitter. The input sample selection is discussed as well as the implementation of the OK model in MATLAB. The authors compare contour maps of the received power predicted by the kriging model to the SPLAT generated map and the kriging model is validated through cross-validation and by inspecting the semi-variogram. The authors find that the OK model produces a Root-Mean-Squared Error (RMSE) of less than 6 dBm for sample set sizes larger than 500 samples as well as a very small ratio between the RMSE and the variation in the input data.**

***Index Terms* - Cognitive radio, Dynamic Spectrum Access, kriging interpolation, Longley-Rice ITM, ordinary kriging, Radio Environment Mapping, TV white space**

## I. INTRODUCTION

The concept of cognitive radio (CR) is defined as a radio system with the ability to assess its surrounding geographical and operational environment, to obtain knowledge and accordingly adapt to changes in its operating parameters and protocols in a dynamic and autonomous way [1]. This is done in an effort to provide reliable communication, independent of its location and which is spectrally efficient [2].

A difficult task in the journey toward CR functionality is providing the ability to dynamically access the frequency spectrum. This can, in turn, lead to effective utilisation of spectral opportunity inside the spectrum. Recently, spatial re-use techniques enjoyed increasing attention, where CRs are allowed to transmit and receive within specified interference constraints [3]. Therefore, research on Power Spectral Density (PSD) maps became of interest as a method of obtaining information on Radio Frequency (RF) traffic in terms of time, space and frequency.

A proposed solution to improve utilisation of the frequency spectrum is Dynamic Spectrum Access (DSA). This involves wireless devices sharing locally available spectrum based on real time demands rather than making use of statically allocated frequencies [4]. Proposed solutions to characterise the spectral use in the area of interest, are the use of interference cartography, channel gain maps or PSD maps. These solutions can collectively be referred to as Radio Environment Mapping (REM).

In order to generate REMs, many measurement samples covering the entire area of interest are required. If a propagation modelling approach is followed, the transmitter data as well as the topographical information of the entire area of interest are required.

Taking the required amount of empirical measurements is, although preferred, usually very time-consuming and the mobile measuring equipment needed is also expensive. Another problem arises when samples need to be measured using a grid fashioned approach. In this case, sampling locations can be difficult or impractical to reach with the equipment. A third problem is introduced by experimental variances due to the dynamic RF environment. Finally, in order to get a good resolution, a lot of measurements are required.

In this paper, the authors evaluate the ability of ordinary kriging (OK) to produce a REM without the large number of samples typically required. The output of the Longley-Rice Irregular Terrain Model (ITM) is considered as measurements at designated points from which a randomly selected sample set is collected. For the purpose of this evaluation, the ITM output is taken as the ground truth. Hence, the authors will evaluate the ability of OK to aid in constructing a REM using significantly fewer sample points in comparison to what would be required if one were to measure these points in the field. More specifically as a case study, the authors consider the Aggeneys Black Mountain transmitter located in Aggeneys in the Northern Cape province of South Africa.

The remainder of this document is structured as follows: a brief background on OK and the Longley-Rice ITM are given in section II, followed by section III which elaborates on the simulation tools used for this experiment. Section IV will give more insight into the research methodology that was followed and precedes the experimental results in section V. The results are evaluated in section VI and the authors come to a conclusion in section VII.

## II. BACKGROUND

### A. Spatial Interpolation

There are numerous different spatial interpolation techniques, geostatistical and non-geostatistical. Each of these has their own advantages and disadvantages for certain situations [5]. The main purpose of interpolation methods is to make inferences of certain properties at locations using only limited data of the surrounding spatial area.

Due to its optimal results when input assumptions are met and its robustness when they are not met, the authors have chosen to investigate kriging as a candidate for effectively producing REMs using the least possible irregularly spaced samples.

### B. Kriging Interpolation

Kriging is a very popular spatial interpolation technique used in geostatistics. Geostatistics is the area in statistics that focuses on geographical applications such as meteorology, mining exploration and other environmental sciences [6]. This technique was originally introduced by the South African mining engineer Danie Krige [7] by whom it was used to map mineral resources by using scattered measurements. The method is based on spatial autocorrelation which originates from the first law of geography (or Tobler's first law). This law states that everything is related to everything else, but near things are more related than distant things [6].

The kriging interpolation technique is generally used for spatial data and has a few variations of implementation. The three most common variations are: simple kriging (SK), ordinary kriging (OK) and universal kriging. Another variation is co-kriging. All of these variations are conceptually the same but differ in the parametrical assumptions that are made.

OK is a B.L.U.E. (Best Linear Unbiased Estimator) spatial interpolation method since the estimates are weighted linear combinations of the sample data used, it attempts to achieve a mean error of zero and minimises the error variance. Last mentioned feature distinguishes OK from other spatial interpolation techniques [8].

Another favourable characteristic of OK is the fact that the points are estimated using the covariances between the data samples and between the estimation point and the data samples. This means that the estimation does not depend on the locations of the sampled data points but rather the separation between them.

What defines OK as a spatial prediction method is the following two assumptions [9], [10]:

1. Model assumption:

$$Z(\boldsymbol{s}) = \mu + \delta(\boldsymbol{s}), \qquad \boldsymbol{s} \in \boldsymbol{D}, \quad \mu \in \mathbb{R}, \quad \text{and } \mu \text{ unknown.} \tag{1}$$

, where $\mu$ is a constant unknown regression function (i.e. $\mu = \alpha_0$ with $\alpha_0$ being a constant unknown value) and $\delta$ is a Gaussian process constructed from the residuals. The constant regression function means that the technique assumes an unknown constant trend in the data [11].

2. Predictor assumption:

$$p(\boldsymbol{Z}; B) = \sum_{i=1}^{n} \lambda_i Z(\boldsymbol{s}_i), \sum_{i=1}^{n} \lambda_i = 1. \tag{2}$$

,where $\lambda_i$ in this case represents the kriging weights [12], [13] which are assigned to each sample value, $Z(\boldsymbol{s}_i)$, and are used as a linear combination to predict the unknown value ($p$). The condition of the kriging weights summing to unity in equation (2) ensures the unbiasedness.

OK utilises a semi-variogram to determine the kriging weights. The model semi-variogram is a curve that is fitted to the experimental semi-variogram. The experimental semi-variogram is calculated by finding the average of the semi-variance cloud for each separation distance of the sample data points. It is also common practice to determine the semi-variogram from the correlogram. A correlogram is usually derived from the correlation matrix describing the correlation between the input samples. The relationship between the semi-variogram and the correlogram is shown in equation (3).

$$\gamma(h) = 1 - c(h) \tag{3}$$

where $\gamma$ is the semi-variance function, $h$ is the separation between sample points (known as the lag) and $c$ is the correlation function of the correlogram. The model semi-variogram is an indication of how the point being estimated is related to the sample points considering the distance between them.

### C. Longley-Rice Irregular Terrain Model

Currently, REMs are mainly generated by using radio wave propagation models. The Longley-Rice ITM is a very popular terrain specific model for this application.

The Longley-Rice model is a radio propagation model used to predict radio signal attenuation over irregular terrain relative to free-space transmission loss. This method is designed for the frequency range 20 MHz to 20 GHz and path lengths from 5 km to 2000 km [14]. It is also known as the Institute for Telecommunication Sciences (ITS) ITM. The model caters for two telecommunication links: point-to-point and point-to-area and it is path specific. The path parameters taken into account in this model are the effective antenna heights, horizon distances and elevation angles, terrain irregularity and reference attenuation. The reference attenuation includes line of sight, diffraction and forward scatter attenuation.

### D. ICASA data

The transmitter data used for this research is provided by ICASA (The Independent Communications Authority of South Africa). The data contains information of all analogue and digital broadcasting stations in South Africa as given by the ICASA final terrestrial broadcasting plan of 2013 [15].

## III. SIMULATION TOOLS

### A. MATLAB ooDACE toolbox

For this research application a kriging toolbox called "ooDACE" (object-oriented Design and Analysis of Computer Experiments) [16], [17] will be used. This

toolbox implements the "Gaussian Process based Kriging surrogate models" [18]. ooDACE [16], [17] is capable of implementing SK, OK, co-kriging, blind kriging and stochastic kriging and supplies several accuracy metrics such as the prediction variance, the process variance and cross-validation prediction error. These metrics can be used to determine the accuracy of the kriging model generated.

### B. SPLAT!

SPLAT! is an acronym for an RF Signal Propagation, Loss and Terrain analysis tool. This tool implements the Longley-Rice ITM with enhancements and one of its main applications is in analogue and digital television and radio broadcasting [19]. SPLAT! is a very popular tool among researchers in the radio propagation field [19]. Interoperability with Google Earth also provides scaled and good quality graphical presentation of results.

SPLAT! has different option for point-to-point predictions as well as area predictions. The output is given in the form of two text files, a portable pixmap (ppm) image and a data file.

The output file of importance for this experiment is the alphanumerical data file. This file contains the boundaries of the analysed area as well as the coordinates, azimuths, elevations to first obstruction and the received power in dBm (decibels of the power referenced to one milliwatt) at that specific point [20]. Another file of interest to us is the '.dcf' (SPLAT colour definition) file. This colour definition file assigns a RGB colour value to each 10 dBm increment of received power and it is used to plot the ppm image which marks the transmitter location and shows the received power in dBm for the specified radius.

## IV. Methodology

### A. Experimental flow

As described in section III, the OK model is implemented in the MATLAB software environment using the ooDACE kriging toolbox. The kriging model considers no topographical information whatsoever but does require measured samples of the received power over the entire area of interest.

The authors gather data for the experiment by implementing the Longley-Rice ITM in SPLAT using the information of the Aggeneys Black Mountain transmitter. The only information regarding the transmitting antenna that SPLAT requires for a received power analysis is the name of the transmitter, the latitude, longitude, transmitter height above ground level and the ERP in watts (W). Both the latitude and longitude can be specified in either decimal format or DMS (degrees, minute, second) format. The alphanumerical output data file used to obtain the samples to be used as input to the kriging model. The authors randomly select 1%, 5%, 10% and 20% of the SPLAT output points as the input sample sets for the OK model. After the samples are obtained, they need to be transformed into a format suitable for kriging.

One of the most important transformations is to convert the LatLon (Latitude/Longitude) geographical coordinates specifying the location at which the samples are taken, into a flat plane. The reason for this is the fact that the kriging model is based on the Euclidean distances between the samples and the geographical coordinates give the locations on the ellipsoidal surface of the earth (i.e. locations are separated by arc lengths). Although the effect of the ellipsoidal coordinates on the relatively small area of analysis is not great, it still affects the accuracy of the kriging predictions. For this experiment, the geographical coordinates were converted to UTM (Universal Transverse Mercator). The reference ellipsoid used is the World Geodetic System 1984 (WGS84). Thus, the input to the kriging model is UTM coordinates which can be used to calculate the distances between samples.

The authors in [12] and [21]–[23] propose *k*-fold and leave-one-out cross-validation (LOO-CV) to evaluate the kriging model. The authors implement a 10-fold cross-validation as well as the LOO-CV using ooDACE.

### B. Analysis parameters

The following command was used in SPLAT in order to implement the Longley-Rice ITM on the Aggeneys Black Mountain transmitter for a radius of 5 km.

*-t Aggeneys.qth -L 10.0 -R 5.0 -olditm -erp 10000 -dbm -metric -ano ReceivedPower_Aggeneys*

SPLAT automatically finds the required SRTM topographical data files stored in the execution directory, based on the location of the transmitter defined in the *.qth* file and the specified area of interest. The command also specifies the receiver antenna height to be 10 m and the *-metric* switch indicates that the specifications are in meters and kilometres instead of feet and miles. The *-erp* and *-dbm* switches invoke the received power analysis with the ERP antenna ERP specified in watts. For the implementation of the Longley-Rice ITM, SPLAT uses the 3-arc-second Satellite Radar Topography Mission (SRTM) data. 3-arc-seconds can be approximated to roughly 90 meters on the surface of the earth. This means that the points predicted by SPLAT are separated by approximately 90 m. Thus, 10533 points are predicted over the 5 km radius area.

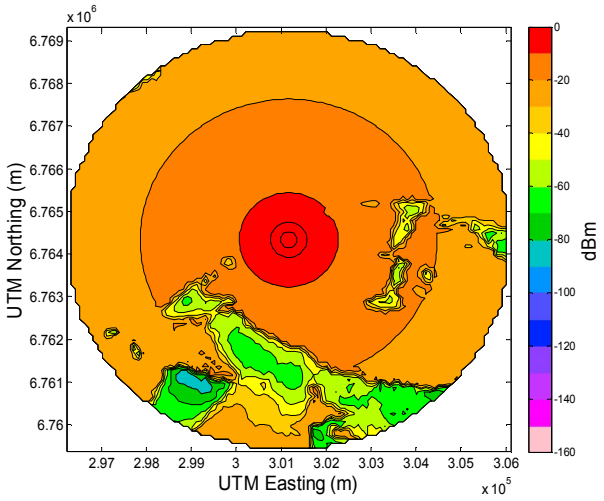
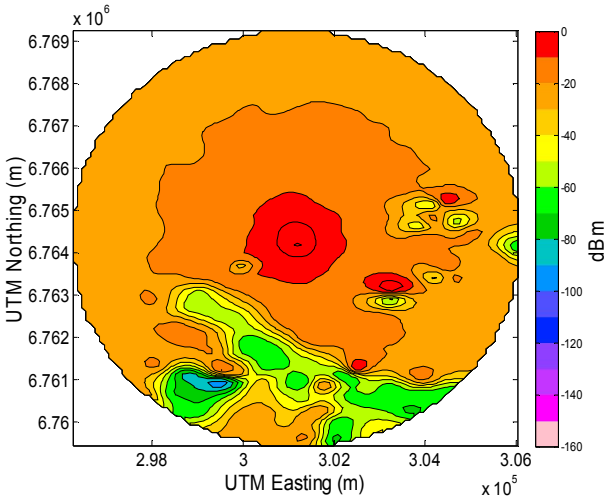


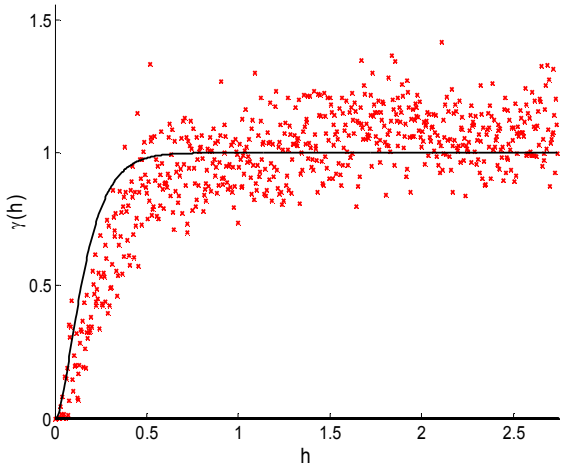

**Figure 1: Received power contour plot of the points predicted using the Longley-Rice ITM for a radius of 5 km**

## V. Kriging Results

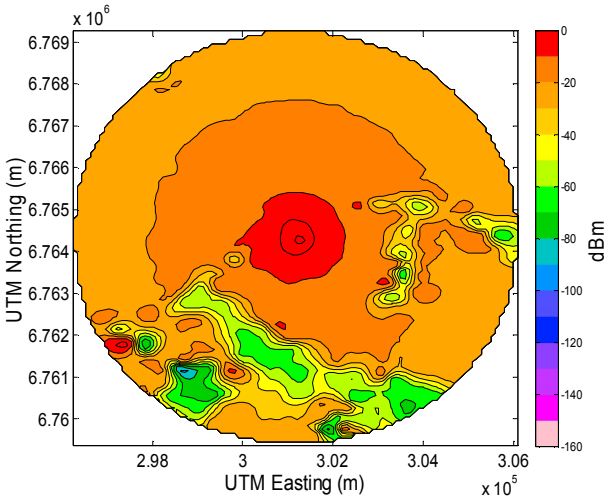
One of the conditions for kriging is that it requires a normally distributed data set. Thus, before the model is fitted to the input data, samples are normalised.



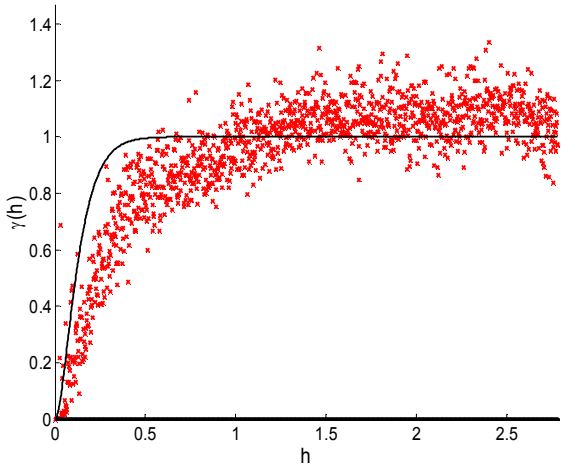

(a) Contour plot of kriging model for 5% samples



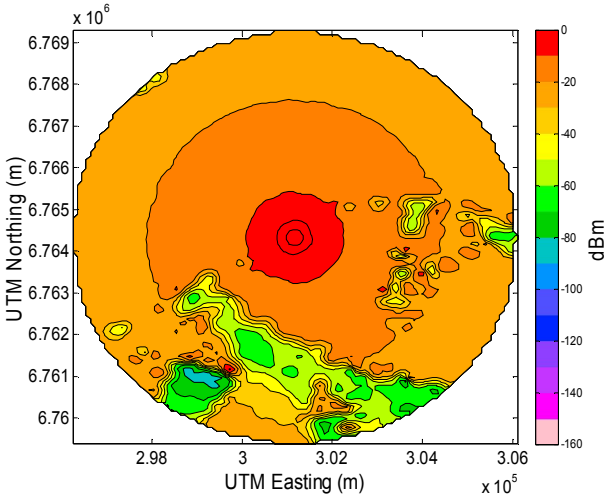

(b) Semi-variogram for 5% samples



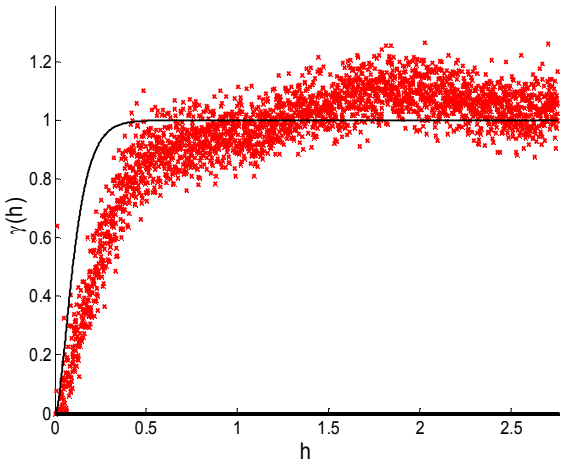

(c) Contour plot of kriging model for 10% samples




(d) Semi-variogram for 10% samples




(e) Contour plot of kriging model for 20% samples




(f) Semi-variogram for 20% samples

**Figure 2: Contour plots and semi-variograms of the models fitted to different sample set size**

The colour scale specified in the SPLAT colour definition file was used to create a colour map in MATLAB. This colour map was used for the contour plot in order to visually compare the results to the received power map generated by the Longley-Rice ITM in SPLAT (see Figure 1).

Since the authors are interested in mapping the received power of the broadcast signal, contour maps of the OK model output are plotted. Figure 2 shows the OK model contour maps as well as the experimental and the model semi-variograms fitted to the normalised data for three

different input sample set sizes. The ooDACE toolbox uses equation (3) to derive the model semi-variogram from the correlation matrix.

Figure 2 shows visually how the kriging model contour plot becomes more accurate in comparison to the contour plot produced by SPLAT for 5%, 10% and 20% samples in figures 2.a, 2.c and 2.e, respectively. This is also evident in the three semi-variograms for 5%, 10% and 20% samples in figures 2.b, 2.d and 2.f, respectively. The semi-variogram plots show both the experimental semi-variogram (designated by red x's) and the model semi-variogram (solid line) for each case. The semi-variances of the input sample data in the experimental semi-variogram becomes less scattered as the number of input sample points increase, which also leads to a better fit of the model semi-variogram.

## VI. Results Evaluation

As mentioned in section IV, the authors evaluate the generated OK model using two forms of cross-validation namely, 10-fold and LOO-CV. The authors cross-validate the model by dividing the sample set into disjoint subsets. One subset is then used for validation while the remaining subsets are used to fit the kriging model. The subset of sample points (or a single sample point in the case of LOO-CV) kept for validation is then predicted using the fitted model. These results are compared to actual sample values using the Root-Mean-Squared Error (RMSE) function. Since the larger errors are given more weight in this approach, the RMSE values also give an indication of the variation in the errors. The authors also considered the RMSE relative to the standard deviation (SD) of the input data using equation (4). This value is expressed as a ratio and is proposed to be a better representation of the actual performance of the model [24]. The RMSE relative to the SD is given by:

$$RMSE_{SD} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Z_{CV,i} - Z(s_i)\right)^2}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Z(s_i) - \bar{Z})^2}} \quad (4)$$

, where $n$ is the number of sample locations, $Z_{CV,i}$ is the cross-validation predicted value at sample location $s_i$ and $\bar{Z}$ is the mean of the input values of all sample locations.

For the *k*-fold cross-validation, the authors used a *k*-value of 10, which means that the sample data set is divided into 10 disjoint subsets. In order to divide the sample set into 10 disjoint subsets, it must be a factor of 10. Thus, each sample set size was rounded to the nearest factor of 10.

For the evaluation of the kriging model, the authors inspected the accuracy of the model using sample set sizes of 1%, 5%, 10% and 20%. Refer to Table 1. It was found that both the 10-fold cross-validation and the LOO-CV produced similar RMSE results. The calculated SDs (in dBm) of the input data was 13.83, 15.55, 14.39, and 14.17 for the respective sample sets in ascending order of sample set sizes. The SDs were calculated in order to obtain $RMSE_{SD}$ as suggested in equation (4).

**Table 1: RMSE value results from cross-validation**

| Number of samples | | 10-fold CV | | LOO-CV | |
|---|---|---|---|---|---|
| ***Count*** | ***%*** | ***RMSE (dBm)*** | ***$RMSE_{SD}$*** | ***RMSE (dBm)*** | ***$RMSE_{SD}$*** |
| 120 | 1 | 7.9224 | 0.5728 | 7.7679 | 0.5617 |
| 520 | 5 | 5.9992 | 0.3858 | 5.8019 | 0.3731 |
| 1040 | 10 | 5.6444 | 0.3922 | 5.4098 | 0.3759 |
| 2100 | 20 | 4.1140 | 0.2903 | 3.8956 | 0.275 |

It is evident throughout the different sizes of sample sets, that LOO-CV yields the smaller RMSE of the two. This is to be expected since the subset of sample data used to train the kriging model is always larger than the 10-fold cross-validation. Furthermore, the decrease in the RMSE value as the size of the sample data sets increase was also found as expected. The small values for $RMSE_{SD}$ indicate very good performance of the OK model relative to the extent of the variation in the input data.

The results shown in Table 1 are quite remarkable considering the fact that the largest number of samples used to predict the 78.5 $km^2$ area, was 2100 samples (i.e. 1 sample for every 37380 $m^2$). In research it was found that the accuracy of empirical propagation models is in the order of 8-10 dB for urban areas and between 10 and 15 dB in rural areas [25]–[27].

## VII. Conclusion

This paper introduced the problem of evaluating OK interpolation as a tool for REM. A brief background on the applicable study fields was given and the methodological approach taken to evaluate OK was discussed. It can be concluded that the OK results are comparable to results obtained using SPLAT with low RMSE as well as a very low $RMSE_{SD}$. This is noteworthy considering the small number of samples used to predict the received power for the entire area as well as the fact that the kriging model requires no topographical information at all. For future research, the authors will look into various methods to increase the accuracy of the OK model fit, which could lead to a reduction in the number of samples required and predictions with lower error. It can thus be concluded that OK shows great promise for effectively aiding in the generation of REMs. A case is therefore to be made for the use of a hybrid approach – measurements combined with kriging as an alternative to pure radio propagation modelling.

## VIII. Acknowledgement

The authors acknowledge the financial support of the Telkom Centre of Excellence (CoE) at the North-West University, Potchefstroom Campus.

## IX. REFERENCES


[1] IEEE Standards Association, "IEEE Standard Definitions and Concepts for Dynamic Spectrum Access: Terminology Relating to Emerging Wireless Networks, System Functionality, and Spectrum Management," *IEEE Std 1900.1a-2012*, no. January, 2013.

[2] S. Haykin, "Cognitive Radio : Brain-Empowered Wireless Communications," *IEEE Journal on Selected Areas in Communication*, vol. 23, no. 2, pp. 201–220, 2005.

[3] S. Kim, "Cooperative spectrum sensing for cognitive radios using kriged kalman filtering," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 1, pp. 24–36, 2011.

[4] L. Cao and H. Zheng, "Balancing Reliability and Utilization in Dynamic," vol. 20, no. 3, pp. 651–661, 2012.

[5] J. Li and A. D. Heap, "A review of spatial interpolation methods for environmental scientists." Geoscience Australia, Record 2008/23, p. 137, 2008.

[6] A. B. H. Alaya-Feki, S. Ben Jemaa, B. Sayrac, P. Houze, and E. Moulines, "Informed spectrum usage in cognitive radio networks: Interference cartography," *2008 IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 1–5, Sep. 2008.

[7] D. G. Krige, "A Statistical Approach To Some Basic Mine Valuation Problems on the Witwatersrand," *Journal of Chemical Metallurgical and Mining Society of South Africa*, pp. 201–215, 1952.

[8] E. H. Isaaks and R. M. Srivastava, *An Introduction to Applied Geostatistics*. New York, United States of America: Oxford University Press, 1989.

[9] G. Matheron, *The Theory of Regionalized Variables and Its Applications*, 5th ed. Fontainebleau, France, 1971.

[10] A. G. Journel and C. J. Huijbregts, *Mining Geostatistics*. Academic Press, London, 1978, p. 304.

[11] W. Kresse and D. M. Danko, Eds., "Geostatistics," in *Handbook of Geographic Information*, Springer, 2012, pp. 43–59.

[12] N. A. C. Cressie, "Spatial Prediction and Kriging," in *Statistics for Spatial Data*, Revised., Wiley Interscience, 1993, pp. 105–210.

[13] D. Arroyo, X. Emery, and M. Peláez, "Sequential Simulation with Iterative Methods," in *Quantitative Geology and Geostatistics*, 17th ed., P. Abrahamsen, R. Hauge, and O. Kolbjørnsen, Eds. Oslo, Norway: Springer Science & Business Media Dordrecht, 2012, pp. 3–15.

[14] O. Thoke, "The Longley-Rice Propagation Model and TV White Space for Ultra WiFi," 2011. [Online]. Available: http://www.brighthubengineering.com/consumer-appliances-electronics/102752-the-longley-rice-propagation-model-and-tv-white-space-for-ultra-wifi/. [Accessed: 16-Sep-2013].

[15] "Final terrestrial broadcasting frequency plan, 2013," *Government Gazette*, Republic of South Africa, vol. 574, no. 36321, 2013.

[16] I. Couckuyt, A. Forrester, and D. Gorissen, "Blind Kriging: Implementation and performance analysis," *Advances in Engineering Software*, vol. 49, no. February 2012, pp. 1–13, 2012.

[17] I. Couckuyt, F. Declercq, T. Dhaene, H. Rogier, and L. Knockaert, "Surrogate-Based Infill Optimization Applied to Electromagnetic Problems," *International Journal of RF and Microwave Computer-Aided Engineering (RFMiCAE)*, vol. 20, no. 5, pp. 492–501, 2010.

[18] T. Dhaene, "ooDACE toolbox," no. June, 2013.

[19] J. A. Magliacane, "SPLAT! Because the world isn't flat!," 2014. [Online]. Available: http://www.qsl.net/kd2bd/splat.html. [Accessed: 23-Jan-2014].

[20] J. A. Magliacane, D. McDonald, and R. Bentley, "SPLAT user manual," no. February, pp. 1–19, 2011.

[21] T. Hengl, *A Practical Guide to Geostatistical Mapping*, Second. 2009.

[22] V. Joseph, Y. Hung, and A. Sudjianto, "Blind kriging: A new method for developing metamodels," *ASME Journal of Mechanical Design*, vol. 130, no. 3, 2008.

[23] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, Massachusetts: MIT Press, 2006.

[24] P. Hiemstra and R. Sluiter, "Interpolation of Makkink evaporation in the Netherlands," De Bilt, 2011.

[25] C. Phillips, D. Sicker, and D. Grunwald, "Bounding the Practical Error of Path Loss Models," *International Journal of Antennas and Propagation*, vol. 2012, pp. 1–21, 2012.

[26] N. Faruk, A. A. Ayeni, and Y. A. Adediran, "On the study of empirical path loss models," *Progress In Electromagnetics Research B*, vol. 49, no. February, pp. 155–176, 2013.

[27] C. Phillips, S. Raynel, J. Curtis, S. Bartels, and D. Sicker, "The efficacy of path loss models for fixed rural wireless links," in *Passive and Active Measurement: 12th International Conference*, 2011, pp. 42–51.



**Willie Boshoff** received his B.Eng. degree in Computer and Electronic Engineering from the North West University Potchefstroom in 2012 and started pursuing a Master of Engineering degree in 2013 at the same institution. He is a Telkom CoE student and his research interests include efficient spectrum usage methods and radio environment mapping.

# Standard Compliant Channel Selection Scheme for TV White Space Networks

Moshe T. Masonta*†, Thomas Olwal†, Fisseha Mekuria† and Mjumo Mzyece*
*Department of Electrical Engineering and F'SATI
Tshwane University of Technology, Pretoria, South Africa
†Wireless Networking and Computing
Council for Scientific and Industrial Research (CSIR) Meraka Institute
Pretoria South Africa
Email: (mmasonta, tolwal, fmekuria)@csir.co.za, mzyecem@tut.ac.za

***Abstract*—In television white space networks, secondary users are required to query an authorised geo-location spectrum database (GSDB) in order to determine the vacant channels or white spaces. While recent development of the Protocol to Access White Spaces (PAWS) by the Internet Engineering Task Force (IETF) is intended to standardize communication between the GSDB and white space devices for sharing spectrum white spaces (WSDs) and their related parameters, the mechanism for channel selection remains an open issue. In this paper, an Analytic Hierarchy Process (AHP) based scheme is proposed for optimal channel decision. The best channel is selected from a pool of available channels provided by the GSDB. Each channel is ranked based on the current class of service offered (either best effort or real time) offered as well as multiple attributes sourced from GSDB. The numerical results show that the proposed scheme is capable of selecting the best channels to satisfy the users' preferences with lower decision latency than the compared existing solution.**

***Keywords*—Analytic hierarchy process, Channel selection, Geo-location spectrum database, TV white space, White space device.**

## I. Introduction

The television white spaces (TVWSs) are portions of radio frequency (RF) spectrum on the TV band that are not being used at a given time and location by the licensed TV incumbents as a result of frequency planning (guard bands) and as a by-product of the global digital switch over process. In our prior work on TV spectrum measurements in both urban and rural areas in Southern Africa, we found that TVWS availability ranges between 100 to 300 MHz [1], [2]. Due to their favourable propagation characteristics, the TVWSs are being considered for providing broadband access in rural areas [3]–[5]. Successful sharing of TV spectrum between broadcast and broadband services depends on adoption of dynamic spectrum access (DSA) regulatory approach. In DSA based broadband networks, cognitive radios (CRs) or white space devices (WSDs) operate as secondary users (SUs) who access the white spaces without creating interference to the licensed or primary users (PUs). Two techniques are commonly considered for discovering the white spaces: *spectrum sensing* and *geo-location spectrum database (GSDB)* [6]. Experimental TVWS broadband networks using GSDBs have been piloted in many parts of the world including South Africa [7].

Recently, the Internet Engineering Task Force (IETF) released an Internet draft on the Protocol to Access White Spaces (PAWS) which is a standardized protocol used for sharing TVWS and their related parameters between the WSDs and GSDBs [8]. While PAWS addresses the communication between the GSDB and WSDs, the mechanism for selecting or allocating the channels to the WSDs remains an open issue. In a large TVWS network that consists of multiple WSDs, optimal channel decision and allocation to satisfy all SUs' quality of service (QoS) requirements, while managing total interference, was found to be an NP-complete problem [9]. As a result, heuristic methods are being considered for finding optimal solutions.

One of the most critical challenges on channel decision in TVWS networks is the heterogeneous propagation characteristics among various TV spectrum bands [10]. This is due to the wider range between the lower TV channel (channel 21, at 470 MHz) and the upper channel (channel 48 at 694 MHz) which exhibits different interference relationships among the WSDs. There are several applications operating on the TV band such as Programme Making and Special Event (PMSE) services which are characterised by narrow-bands (in the range of 200 kHz) as compared to the 6 or 8 MHz wideband TV channel. PMSE devices may include hand-held devices which may use a combination of broadcasting (e.g. TV reception) and broadband such as an Internet Protocol Multimedia Subsystem (IMS) clients [11]. As a result, the TVWSs will not have equal channel bandwidth. Hence it is important to consider the user preferences, white space channel characteristics and other network attributes when designing channel selection solutions.

Over the past decade, multiple attribute decision making (MADM) techniques have been considered to address the decision making process in integrated and heterogeneous wireless networks [12]–[15]. As a subset of MADM technique, Analytic Hierarchy Process (AHP) provides a powerful and robust step-by-step decision-making process through pairwise comparisons that can be used to combine qualitative and quantitative factors for prioritizing, ranking and evaluating alternatives [16]. By using the hierarchical approach, a complex problem is broken-down into smaller and less-complex problems for simplified decision making. This makes AHP the most preferred MADM technique where several attributes, such as throughput, bandwidth, delay, QoS requirements, etc., are considered in decision-making process. In [12], AHP was applied to solve a network selection problem in an integrated

wireless local area network (WLAN) and 3rd generation cellular networks. Rodriguez-Colina *et al.* used AHP for spectrum decision making in cognitive radio networks (CRNs) [13]. Ramli *et al.* applied AHP to make a decision on how to fairly allocate the spectrum licenses [14]. For TVWS specific networks, a channel allocation algorithm based on Simulated Annealing (SA) is proposed in [9]. A distributed spectrum allocation algorithm based on spectrum fragmentation and non-uniform interference relationship among multiple access point (APs) is proposed in [10]. Both [9] and [10] offload the channel decision functionality to the local-white space database. While this might be a suitable solution for a small geographical area where such a GSDB is located within the vicinity of the TVWS network, it might not be feasible for wider TVWS network deployments where a national GSDB is located far away from the TVWS network. For instance, TVWS network might be deployed in a rural or remote area while a GSDB can be located hundreds of kilometres away or located in the cloud (or in a different province or country). In such scenarios, it becomes important for a TVWS base station (TVWS-BS) to manage and coordinate spectrum decision and allocation to its associated customer premises equipments (CPEs) or WSDs. Furthermore, the existing schemes, such as the Simulated Annealing (SA) scheme [9] are computationally complex and time-consuming, especially when determining and choosing channel parameters [9].

In this paper, we propose an AHP-based channel decision scheme which considers the user preferences and channel conditions based on the information collected from the national GSDB. Our proposed scheme uses AHP to determine the weights and ranking of suitable channels taking into account the white space channel attributes such as the available bandwidth, available channel occupancy time and the allowed transmit power limit. Based on the weights and ranking of white space channels (or alternatives), the national regulatory rules on PU protection against harmful interference are considered before the best channel (i.e. channel with higher relative weight) is selected and allocated to the WSDs. By considering these attributes, our scheme is compliant to the PAWS protocol and can be implemented to most WSDs because it is portable and not computationally complex.

The remainder of this paper is organized as follows. Section II describes the system model considered in this paper. An overview of AHP and the relevant PAWS messages with their parameters are discussed in Section III. The proposed AHP-based channel decision model is presented in Section IV. Section V presents and discusses the numerical results. The paper is concluded in Section VI.

## II. System Model

### A. Network Scenario

We consider a centralised TVWS network consisting of a single TVWS base station (TVWS-BS) with three sectors and multiple fixed WSDs or CPEs which provide broadband access to several schools. This is the same network scenario used during our recent TVWS trials in Cape Town, South Africa [7]. The TVWS-BS co-exists with a single frequency network TV broadcast system where multiple transmitters are co-located on a single mast.

### B. Access to GSDB

The TVWS-BS is connected to one of the approved GSDBs through either fixed line or alternative wireless technology (such as satellite links) which also provides the back-haul. Based on the Cape Town TVWS trial network, each of the three sectors operates on a different TV channel (thus, a minimum of three channels are required to provide minimum connectivity to the schools). Once switched "On" the TVWS-BS establishes an HTTP session with the national GSDB in order to register and initiate a query for available channels. The messages between the TVWS-BS and the GSDB are standardized by the PAWS [8], while messages between the WSD and the GSDB is use dependent, i.e. not standardized by the PAWS. Once selected a channel for itself, the TVWS-BS then uses each WSD's geo-locations (which are readily known to the TVWS-BS) to query the GSDB for more available channels using the AVAIL_SPRECTRUM_BATCH_REQ request. The WSDs are also called slaves because they do not have direct access to the GSDB [8]. Depending on the amount of traffic demand or class of service (i.e. real-time or best effort) by the WSDs and the distance from the BS, the best suitable number of channels will be selected (also depending on the availability of white spaces). Finally, by sending a SPECTRUM_USE_NOTIFY, the TVWS-BS notifies the GSDB of the selected channels used by the entire network. This will assist the GSDB not to provide the same list of available channel to other secondary networks, thereby preventing any co-channel interference [17].

### C. PU Protection Against Interference

By relying on the GSDB for white space availability, interference to the PUs is already minimized since the GSDB will only provide a list of channels which are not occupied by the PUs. However, there are some restrictions on the allowed maximum transmission power depending on whether the TVWS channel being used is adjacent to the PU channel or more than $n \pm 1$ away from the PUs channel (where $n$ is the primary channel number). The regulators can determine the allowed transmission power for adjacent and non-adjacent channels as specified by the FCC [3]. Operation on adjacent channels is likely to be kept lower than the operation on other channels which are more than $n \pm 1$ away from the primary channels. Therefore, it is important for the TVWS-BS to carefully select the best channel for its associated WSDs taking into account the transmit power limit set by the national regulator. The proposed system should have a list of all primary channels which must be avoided when deciding the best channels to select. This list is also useful to the TVWS-BS when checking for the adjacent primary channels.

We assume the primary TV channel size of 8 MHz wide. Thus the list of occupied frequency blocks are ordered and stored as $F_{L,i}$ and $F_{U,i}$, which represent the $i^{th}$ lower and upper channel edge frequencies. The PU Protection Rules is determined by checking whether the startHz and stopHz values of each white space channel is adjacent to the lower and upper edge frequencies of the primary channels, respectively. Thus, if startHz = $F_{U,i}$ (where $i$ is the index of the busy or occupied primary channels) then the lower edge of the available channel is adjacent to the upper edge of the primary channel. And if stopHz = $F_{L,i}$, then the upper edge of the available channel is

adjacent to the lower edge of the primary channel. To represent whether the white space channel is adjacent to the $n \pm 1$ primary channels, we define $\mathbf{PU_{n\pm1}}$ as in equation (1):

$$\mathbf{PU_{n\pm1}} = \begin{cases} 1, & \text{available channel is adjacent to the PUs,} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

## III. AHP Overview and PAWS Parameters

Before explaining how the model works, we start by providing a brief overview or introduction of the AHP and the PAWS message and parameters relevant to this paper.

### A. Analytic Hierarchy Process (AHP) Overview

First introduced by Saaty in the late 1970's, AHP is a theory of measurement through pairwise comparisons that can be used to combine qualitative and quantitative factors for prioritizing, ranking and evaluating alternatives [16]. Over the past thirty decades, AHP has been applied in different fields such as finance, politics, personnel, sports, social sciences, engineering and medical field [18]. AHP is generally performed in five steps (S1 - S5) [16]:

- S1: Structure a problem as a decision hierarchy of independent decision elements where the goal is at the top level and the set of alternatives on the lowest level;
- S2: Collect relevant data about the decision elements;
- S3: Compare the decision elements pairwise on each level based on their importance to the elements in the level above, thereby constructing a comparison matrix;
- S4: Calculate the relative priorities of decision elements in each level; and
- S5: Synthesize the above results to achieve the overall weight of each decision alternative.

### B. Weight Determination using AHP

From the developed pairwise comparison matrix $A$, the global priority vector (PV) or weights $w$ are found using the eigenvector method. So, $w$ is found by solving:

$$Aw = \lambda_{max} w, \tag{2}$$

where $\lambda_{max}$ is the largest or principal eigenvalue of matrix $A$. Global weights are determined by making pairwise comparisons of the alternatives with respect to the available criteria.

After completing step 5, the consistency test, which is one of the most important AHP features, is performed to measure the degree of inconsistency during the pairwise comparison. The consistency test is performed by first computing the matrix's consistency index (CI) as in (3):

$$CI = \frac{\lambda_{max} - n}{n - 1}, \tag{3}$$

where $n$ is the total number of activities in the pairwise matrix. Using the CI, we then calculate the consistency ratio (CR) as:

$$CR = \frac{CI}{RI}, \tag{4}$$

where RI is the random index related to the order ($n$) of judgement matrix as calculated in [16]. The CR of less than 10% ($CR \leq 0.1$) is acceptable (i.e. the pairwise comparison process was consistent). If the CR is higher than 10%, the process was inconsistent and the pairwise comparison must be re-evaluated. Finally, the results of the AHP are conveniently captured and visualized in the form of a decision profile, which offers a convincing view of the results of rating the alternatives.

### C. Available Spectrum Response Message

According to PAWS protocol, attributes are extracted from the AVAIL_SPECTRUM_RESP message which is sent to TVWS-BS by the GSDB [8]. This message contains several parameters which includes ***timestamp, spectrumSchedules, maxTotalBwHz, maxContiguousBwHz****, etc.*. The most important parameters used in our model are *maxTotalBwHz and spectrumSchedules* which are explained next.

*1)* ***The maxTotalBwHz Parameter****:* The maxTotalBwHz provides the maximum total bandwidth (in Hz) which may or may not be contiguous [8].

*2)* ***The spectrumSchedules Parameter****:* Provides a combination of EventTime and Spectrum elements. The EventTime element specifies the start (*startTime*) and stop (*stopTime*) times of an event. The TVWS-BS will then use these elements to calculate how long each white space channel is likely to be available.

The *spectrum* element in the spectrumSchedules parameter consists of two parameters which characterizes a list of frequency ranges and permissible power levels for each range. The *frequencyRanges* lists the maximum permissible power levels within a frequency range.

## IV. Proposed Channel Decision Scheme

The proposed channel selection model is performed based on the flowchart shown in Fig. 1. We assume that the TVWS-BS is authorised and registered with the national GSDB. The model starts when the TVWS-BS queries the GSDB after performing the necessary initialization and authentication processes [8]. Once the TVWS-BS receives the list of spectrum from the GSDB, it extracts important parameters or attributes which will be used for decision making. The maxTotalBwHz parameter is important for our model to decide whether to perform AHP analysis for spectrum selection or not. At least one full channel (equivalent to 8 MHz bandwidth) should be available for the the AHP analysis to be conducted. If no channel is available, the TVWS-BS will query the GSDB after a predefined period of time until at least more than one channel is available to allow the channel allocation process to start.

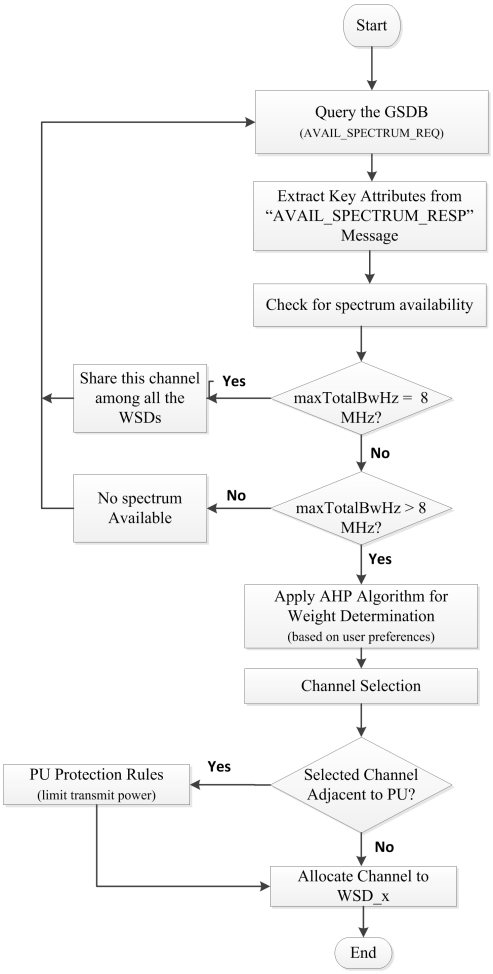



Fig. 1: Proposed channel selection scheme flowchart

### A. White Space Channel Attributes Collection

Based on the above parameters, the following three channel attributes are defined as criteria during the AHP analysis:

***Bandwidth*** = *stopHz - startHz*,

where startHz and stopHz are the start and stop of the frequency range in Hz, respectively and these values are sourced from the frequencyRanges element. This represents the usable channel width which may vary from a few kHz to a few MHz depending on whether there are contiguous channels or not. The reasons for such variable channel width are motivated by the types of other SUs (such as IEEE 802.22 or IEEE 802.11af standards) and PU technologies (such as narrowband PMSE).

***Transmit Power*** = *maxPowerDBm*.

***Event Time*** = *stopTime - startTime*,

where startTime and stopTime are the inclusive start and end of the channel availability.

### B. Proposed Spectrum Decision Hierarchy Structure

Fig. 2 shows the proposed AHP hierarchy structure where the top level presents the main goal on the structure, which

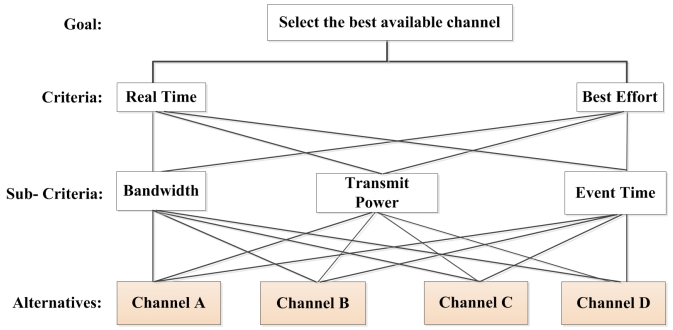



Fig. 2: Channel Selection hierarchy structure

is to select the best available channel. The selection of the channels is based on the user's preferences which requires a specific QoS for each class of service. In this paper, we consider two classes of service (CoS): *Real Time (RT)* and *Best Effort (BE)*, which are on the second level of the hierarchy. On the third level of the hierarchy are the three independent criteria to be compared when selecting the channels. At the bottom of the hierarchy are at least four alternative channels which are compared in order to select the best channel.

### C. Deciding the Best Channel

The final stage of the AHP analysis is reached when the overall weight vector $\mathbf{W}$ for each alternative (available channels) is found. From this weight vector, the best channel $C_{best}$ is selected by observing the highest weight from the alternatives as shown in equation (5).

$$C_{best} = \max(\mathbf{W}) \tag{5}$$

where $\mathbf{W} = \{w_1, w_2, \ldots, w_n\}$ and represents the overall weights for each alternative. However, if the consistency ratio (CR) was found to be bigger than 10%, the above analysis is repeated until the CR is less than 10%. In the next section we present the numerical results to demonstrate our channel decision scheme.

### D. PU Protection Rules

Once the best channel is selected, we use equation (1) to check whether the selected channel is adjacent to the primary channel or not. If it is adjacent, we apply the PU Protection Rules by making sure that transmit power is kept within the limit for adjacent channel operations. Table I shows the allowed transmission power for adjacent and non-adjacent channels as specified by the FCC [3]. The FCC limits are used here as a guideline and they may vary according to the individual country's regulatory power limits. If the selected channel is not adjacent to any PU, the channel is allocated to the WSD at the default transmit power. The default transmit power is the power that was provided by the GSDB. Once all the channels are selected and allocated to the WSDs, the TVWS-BS uses the SPECTRUM_USE_NOTIFY message to update the GSDB about the selected channels. Then the GSDB will mark those channels as unavailable.

TABLE I: FCC Compliant WSD Operation Characteristics

| FCC Approach (6 MHz TV channel bandwidth) | | | | | |
|---|---|---|---|---|---|
| **WSD Types** | **Transmit Power Limit** | **PSD Limit (100 kHz)** | **Adjacent Channel Limit (100 kHz)** | **Access to Database** | **Geo-location** |
| **Fixed WSD** | 30 dBm (1 W) | 12.6 dBm | -42.8 dBm | Mandatory | Mandatory |
| **Personal/Portable** | 20 dBm (100 mW) | 2.6 dBm | -52.8 dBm | Only Mode II devices (Moe I devices act as slaves to Fixed or Portable WSDs) | Mandatory for Mode II devices (master WSD) |
| **Personal/Portable (operating on Adjacent channel)** | 16 dBm (40 mW) | -1.4 dBm | -56.8 dBm | | |
| **Sensing only WSD** | 17 dBm (50 mW) | -0.4 dBm | -55.8 dBm | Not Required | Not Mandatory |

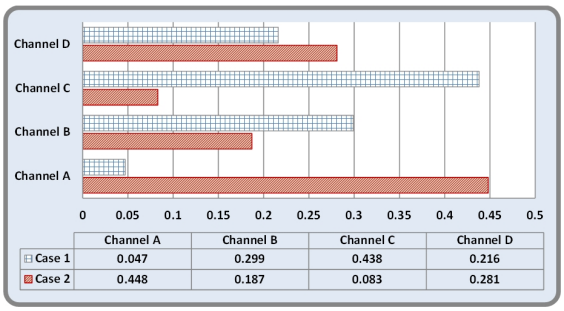



Fig. 3: Decision profile for channel selection in RT CoS.

TABLE II: Two case scenarios with available channels characteristics for Real-Time and Best-Effort CoS

| CASE 1: Available Channel Characteristics | | | | |
|---|---|---|---|---|
| | **CHANNEL A** | **CHANNEL B** | **CHANNEL C** | **CHANNEL D** |
| **TX limit (dBm):** | 25 | 30 | 20 | 17 |
| **Event Time(minutes):** | 60 | 240 | 360 | 100 |
| **bandwidth (MHz)** | 5 | 8 | 7 | 4 |
| CASE 2: Available Channel Characteristics | | | | |
| | **CHANNEL A** | **CHANNEL B** | **CHANNEL C** | **CHANNEL D** |
| **TX limit (dBm):** | 16 | 25 | 12 | 24 |
| **Event Time(minutes):** | 420 | 50 | 120 | 200 |
| **bandwidth (MHz)** | 8 | 5 | 5 | 8 |

## V. Results and Discussions

In this section we present simulation results related to the channel selection based on the white space attributes provided by the GSDB as well as the user preferences for the CoS (i.e. RT and BE).

### A. Simulation Scenario

In order to demonstrate our scheme, we consider a scenario where the TVWS-BS queries GSDB for available white spaces on behalf of its associated WSDs. Four channels with different characteristics are returned by the GSDB and the TVWS-BS must at least select one channel to allocate to each WSD. We evaluate two cases: one for RT CoS and the other one for BE services. Table II shows characteristics of the available channels received by the TVWS-BS from the GSDB for both RT and BE CoS.

### B. Discussions

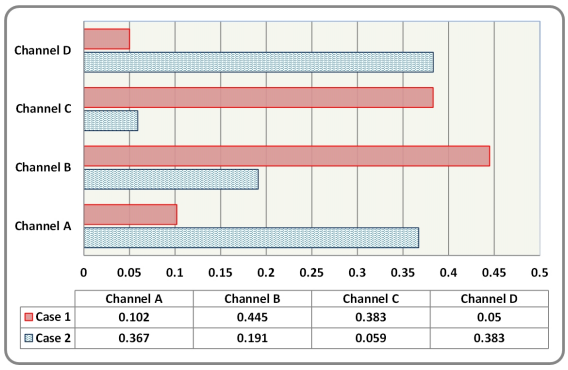
After receiving a set of available channels from the GSDB, the scheme extracts and computes the values of key parameters such as channel bandwidth, event time and the allowed transmission power. Using the AHP algorithm, the weights of these three parameters or criteria are determined for both RT and BE CoS, depending CoS for the active application to be transmitted. The weights and consistency ratio (CR) for the three criteria are shown in Table II. These weights shows the important relationship between the criteria and CoS. The criteria with the highest weight correspond to the highest degree of importance for the CoS considered. For instance, in RT services, the event time (duration of channel availability) and transmit power are more important than the channel bandwidth. For BE applications, of course the bandwidth is more important, followed by the transmit power. Furthermore, we have shown the CR for each CoS, which was found to be less than 10%. This means that our judgements in the pairwise comparison matrix were consistent.

TABLE III: Weights of criteria for RT and BE services

| **Criteria** | **Real-Time** | **Best Effort** |
|---|---|---|
| Bandwidth | 0.072 | 0.637 |
| Transmit Power | 0.279 | 0.258 |
| Event Time | 0.649 | 0.105 |
| Consistency Ratio (CR) | 0.056 (5.6%) | 0.033 (3.3%) |

Using the weights of the criteria, we then a developed pairwise comparison matrix comparing the four available alternative channels against each criterion (i.e. bandwidth, event time and transmit power). The same procedure that we used to find the criteria weights was followed for each case scenario as shown in Table II. Fig. 3 shows the decision profiles or relative weights for each of the four alternative channels RT CoS. It can be seen that our scheme was able to select *Channel C* for case 1 because it has the longest available time (event time) than other channels.Despite having the lower bandwidth and transmit power than*Channel B*, as well as lower transmit power than *Channel A*, *Channel C* is suitable for RT services where the SUs would not want to experience some call drops due to shorter event time. For case 2, *Channel A* was selected for RT transmission. Again, this decision was mainly based on the high degree of weight given to the event time criteria for RT services.

In Fig. 4, the BE CoS decision profiles for each of the four available channels are shown. For case 1, *Channel B* has the highest weight and was selected for BE transmission. This is mainly due to the size of the channel bandwidth which is higher than other alternative channels. As shown in Table III, BE CoS allocated more weight to the bandwidth criterion followed by the transmit power. Event time is ranked lower for BE services because such services are delay tolerant than RT services. Thus, the higher the channel bandwidth, the higher the transmission throughput. If the selected channel's event time elapses, the TVWS-BS will have enough time to allocated a new channel without compromising the SU's QoS.




| | Channel A | Channel B | Channel C | Channel D |
|---|---|---|---|---|
| Case 1 | 0.102 | 0.445 | 0.383 | 0.05 |
| Case 2 | 0.367 | 0.191 | 0.059 | 0.383 |

Fig. 4: Decision profile for channel selection in BE CoS.

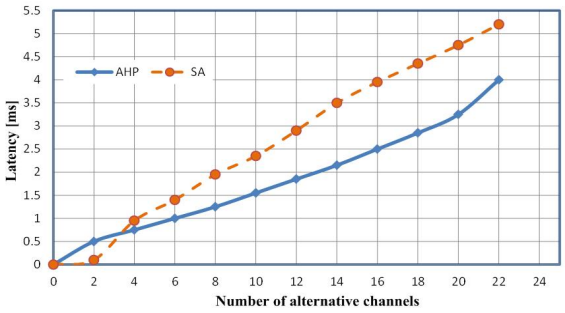



Fig. 5: Performance of AHP and SA schemes with respect to simulation latency

Fig. 5 compares our AHP-based scheme to the SA approach (used in [9]) based on the decision making average simulation latency as a function of alternative channels. We compared the simulation latency of each scheme for selecting the best channel between 2 to 22 alternative channels. The delay in SA scheme is due to the complexity of choosing the parameters before selecting the best channel for the available alternatives. Instead of a hierarchical structure for criteria and sub-criteria, which is used in AHP, SA scheme is based on simulating a random walk on the set of states [9]. Furthermore, the AHP scheme is more convenient to evaluate only two alternatives at a time, which makes it robust especially when dealing with a large number of alternatives. Whereas SA requires multiple combinatorial optimization steps to converge. Since the latency and low computational complexity are crucial parameters for a spectrum decision scheme, especially in DSA based networks [19], our proposed scheme performs best.

## VI. Conclusion and Future Work

In this paper, we have developed an AHP-based channel selection scheme which jointly considers multiple parameters provided by the GSDB as well as the users' preferences. It was shown that the proposed scheme performs adequately and yields acceptable and accurate results in selecting the best channels for different WSDs in a TVWS network. The scheme was also found to perform with low latency and easy to implement when compared to other existing schemes. Future work includes the integration of the AHP scheme with optimization techniques in order to allow optimal selection of the channels, and the use of a game theoretic approach to analyse the model for fairness and energy efficiency.

## References

[1] M. T. Masonta, D. Johnson, and M. Mzyece, *The White Space Opportunity in Southern Africa: Measurements with Meraka Cognitive Radio Platform*, R. Popescu-Zeletin, *et al.*, Ed. Springer, vol. 92, pp. 64-73, Feb. 2012.

[2] A. Lysko, M. Masonta, D. Johnson, and H. Venter, "Fsl based estimation of white space availability in UHF TV bands in Bergvliet, South Africa," in *SATNAC*, George, South Africa, Sep. 2-5 2012.

[3] Federal Communications Commission, "Unlicensed operation in the TV broadcast band," *Federal Register: Rules and Regulations*, vol. 77, no. 96, pp. 29 236 – 29 247, 2012.

[4] M. Fitch, M. Nekovee, S. Kawade, K. Briggs, and R. MacKenzie, "Wireless services provision in TV white space with cognitive radio technology: a telecom operator's perspective and experience," *IEEE Communications Magazine*, vol. 49, no. 3, pp. 64–73, Mar. 2011.

[5] W. Webb, "On using white space spectrums," *IEEE Communications Magazine*, vol. 50, no. 8, pp. 145–151, Aug. 2012.

[6] M. Masonta, Y. Haddad, L. De Nardis, A. Kliks, and O. Holland, "Energy efficiency in future wireless networks: Cognitive radio standardization requirements," in *IEEE CAMAD*, Barcelona, Spain, Sep. 17-19 2012.

[7] TENET, "The Cape Town TV white spaces trial," Available from: http://www.tenet.ac.za/tvws, 2013, [Accessed: 14/10/2013].

[8] V. Chen, Ed., S. Das, L. Zhu, J. Malyar, and P. McCann, "Protocol to access spectrum database," Jun. 2013, draft-IETF-PAWS-Protocol-06, Expires 21 December 2013.

[9] B. Ye, M. Nekovee, A. Pervez, and M. Ghavami, "TV white space channel allocation with simulated annealing as meta algorithm," in *CROWNCOM*, Stockholm, Sweden, Jun.18 – 20 2012.

[10] X. Feng, J. Zhang, and Q. Zhang, "Database-assisted multi-AP network on TV white spaces: Architecture, spectrum allocation and AP discovery," in *IEEE DySPAN*, Aachen, Germany, May 3 – 5 2011.

[11] M. Masonta, O. Oyedapo, and A. Kurien, "Mobile client for the next generation networks," in *BROADCOM*, Pretoria, Nov. 23-26 2008.

[12] Q. Song and A. Jamalipour, "Network selection in an integrated wireless LAN and UMTS environment using mathematical modeling and computing techniques," *IEEE W. Comm.*, vol. 12, no. 3, pp. 42–48, Jun. 2005.

[13] E. Rodriguez-Colina, P. C. Ramirez, and A. C. E. Carrillo, "MADM-based network selection in heterogeneous wireless networks: a simulation study," in *Int. WOCN*, Paris, France, Sep. 24–26 2011.

[14] R. Ramli, *et al.*, "Modeling of spectrum demands through hybrids of analytic hierarchy process and integer programming," *International Journal of Modeling and Optimization*, vol. 1, no. 5, pp. 368–374, 2011.

[15] F. Bari and V. Leung, "Automated network selection in a heterogeneous wireless network environment," *IEEE Network*, vol. 21, no. 1, pp. 34 – 40, 2007.

[16] T. L. Saaty, "Decision making with the analytic hierarchy process," *Int Journal of Services Sciences*, vol. 1, no. 1, pp. 83–98, 2008.

[17] C. Ghosh, S. Roy, and D. Cavalcanti, "Coexistence challenges for heterogeneous cognitive wireless networks in TV white spaces," *IEEE Wireless Communications*, vol. 18, no. 4, pp. 22–31, Aug. 2011.

[18] W. Ho, "Integrated analytic hierarchy process and its applications a literature review," *European J. of Op. Res.*, vol. 186, p. 211228, 2008.

[19] L. Mfupe, M. Masonta, T. Olwal, and M. Mzyece, "Dynamic spectrum access: regulations, standards and green radio policies consideration," in *SATNAC*, George, South Africa, Sep. 2-5 2012.

**Moshe Timothy Masonta** received M. Tech degree in Electrical Engineering from Tshwane University of Technology (TUT) in 2008 and an MSc in Electronic Engineering (2010) from ESIEE de Paris, France. He is a Senior Researcher at the CSIR Meraka Unit and a Doctorate candidate in Electrical Engineering at TUT. His research interests are in dynamic spectrum access and management, cognitive radio systems, television white space spectrum, spectrum regulations and energy efficiency in wireless networks.

# Improving Trustworthiness amongst Nodes in Cognitive Radio Networks

Efe Orumwense[1] Olutayo Oyerinde[2] Stanley Mneney[1]

School of Electrical, Electronic and Computer Engineering,[1]
University of KwaZulu-Natal, Durban, 4041, South Africa.
Tel: +27 74 7960548
School of Electrical and Information Engineering,[2]
University of the Witwatersrand, Johannesburg, 2050, South Africa.
efe.orumwense@gmail.com[1]; olutayo.oyerinde@wits.ac.za[2]; mneneys@ukzn.ac.za[1]

**Abstract** – *Cognitive Radio is a key technology whose major objective is to alleviate the problem of spectrum scarcity and inefficiency in spectrum usage. In a cooperative communication environment, a secondary (unlicensed) user is equipped with a cognitive radio whose nodes receive spectrum sensing occupancy information from a primary (licensed) user so as to promote dynamic spectrum access (DSS). Since spectrum sensing information is being received by the secondary users, it is important to verify if this information is actually coming from a genuine primary user. Without this verification, a malicious user may be able to falsify this spectrum sensing information thereby resulting interference to primary users and spectrum underutilization between secondary users. In this paper we examine a concept called trustworthy networks and propose techniques to verify if the source of the spectrum occupancy information is from a genuine primary user in order to evict malicious users from the network and maximize spectrum utilization efficiency.*

**Keywords** – *Cognitive Radio, Cognitive Radio Networks, Trustworthiness, Secondary User, Primary User, Spectrum Sensing, Primary User Emulation Attacks (PUEA).*

## I. INTRODUCTION

With the ever increasing scarcity of spectrum resources, the Federal Communications Commission (FCC), an independent agency regulating interstate and international communications, has decided to permit unlicensed (secondary) users to make use of the spectrum belonging to licensed (primary) users when it is not in use by the primary users. Cognitive radio [1] has been employed for this purpose and it has emerged to be an enterprising solution to reduce this problem of spectrum shortage by carrying out a spectrum sensing process for the purpose of identifying fallow or vacant bands in the spectrum. Spectrum sensing is one of the basic functionalities where cognitive radio networks monitor available spectrum band, capture their information and then identify spectrum holes or white spaces. Once these white spaces are identified, secondary users can opportunistically utilize them by operating in them without causing interference to the primary users. The primary user has the priority to access the spectrum band, which means that at any time the primary user decides to use this frequency band for transmission, the secondary user must vacate that specific band so as not to cause interference. But when there is no active primary user communication in the spectrum, all other users enjoy equal right to access the unoccupied spectrum bands.

The most efficient way for secondary users in identifying white spaces in a spectrum is by detecting primary users. Secondary users do not get a direct feedback from the primary users about their transmission so they strongly rely on the sensing ability of their nodes to do so and this can be done either cooperatively or non-cooperatively. In non-cooperative spectrum sensing, the secondary user needs to independently sense the spectrum for unused spectrum band and use the spectrum band without causing interference to the primary user [2]. While in cooperative spectrum sensing, multiple secondary users perform local spectrum sensing independently and then makes a binary decision and forwards this decision to the secondary base station or fusion center for fusion [3] which eventually leads to more accurate detection of primary signals. In this work, we will concentrate on cooperative spectrum sensing since it yields better sensing results [4] [5].

However, the secondary user ability to sense and exploit the spectrum in the cognitive radio network imposes some threats and provides an opportunity for some malicious users to intrude the network and disrupt the performance of cognitive radio spectrum sensing [6]. An example of this threat is called Primary User Emulation Attacks (PUEA). In this type of attack, malicious users send fake signals that resemble that of a primary signal over the licensed frequency band thereby causing genuine secondary users to erroneously identify the attacker as a primary user. This will cause secondary users to migrate from or not to make use of the vacant band thereby robbing the secondary users the opportunity to use the spectrum [7]. To mitigate such an attack, a

trustworthy network should be established whereby the trust level of every node in the system can be assessed individually or through a combination of other nodes and also a verification technique should be carried out to enable secondary nodes identify the information provided by genuine users. This way, the secondary user is sure that the information regarding the occupancy of the spectrum is provided by a genuine primary user and the nefarious activities of a malicious user are being thwarted.

Various techniques to evaluate the trust of a user in a cognitive radio environment have been proposed in literature. In [8], a technique called weighted trust evaluation method was proposed, it was developed for wireless sensor networks against selfish users where the sensing information is shared and the trust value degrades if the sensor node provides false information. However, the impact of malicious users is more severe than selfish users because interference to primary users can have serious legal consequences. Also in [9], Wang proposed a technique for assessing trustworthiness of a user by calculating the suspicious level of users. But the technique does not mention how to calculate the thresholds and does not utilize the location information of primary users. Even though several techniques have been proposed in literature for evaluating trust in wireless networks, most of them suffer from several shortcomings.

In this paper, we propose and analyze new techniques to verify spectrum sensing information provided by secondary users so as to identify malicious users in the system and create a trustworthy network in order to build a healthy relationship amongst nodes in cognitive radio networks.

## II. CREATING A TRUSTWORTHY COGNITIVE RADIO NETWORK

Trust is an important factor that cuts across many facets of disciplines and based upon it many relationships are formed. Whether it is used for security on recommended systems, the issue of trust will help in successful message transmission among network entities. In a cognitive radio network, trust amongst its nodes will improve the reliability of the spectrum occupancy information of the primary users and ease the decision making process of the fusion center in terms of cooperative spectrum sensing. A malicious user node might detect the absence of primary signal and sends false information that shows the presence of a primary signal to the fusion center. The fusion center erroneously decides that the primary signal is present, this way the malicious user selfishly uses the entire free spectral band [10].

To curb this menace, a trust value is assigned to secondary user nodes where it will be measured by other nodes in terms of the expected genuineness of its information amongst other decisions made from the collective information. This can be achieved when a secondary node sensing result is always different from all other nodes sensing results. A specific scenario is when all nodes report the absence of a primary user and a specific node reports the presence of a primary user. That node is then regarded as a malicious node and its sensing result is removed before a final decision is taken by the fusion center. In this way, a trustworthy network will be created where by the trust level of every component can be assessed individually or through a combination with other nodes.

In another sense, secondary nodes of a cognitive radio network form a social relationship between themselves to help build trust in the network. A set of nodes can form a sub group and give positive or negative rating on each other based on their previous encounters in order to determine and assess the trust rating of each other. In this way, malicious or untrustworthy nodes can be detected cooperatively because of their low trust rating. Therefore, the information can be diverted from these malicious nodes or the information originating from these nodes can either be ignored or disregarded before the fusion center makes a final decision. Also in a cooperative scenario, a node can change its association with a neighboring node when it finds out that the level of trust value of that node has drastically been reduced thus ensuring the network operates in a trustworthy manner.

In order to ensure a trustworthy cognitive radio network, a robust transmitter verification scheme [6] that can distinguish between trustworthy secondary users and malicious secondary users is also necessary. In hostile environments, such a mechanism can be integrated into the spectrum sensing process of a cognitive radio network to enhance its trustworthiness.

## III. SYSTEM MODEL

Considering a system as in Fig. 1, where all the secondary and malicious users are distributed across a circular grid, with a distance $d_P$ between a good secondary user and a primary user and a distance $d_M$ between a malicious secondary user and a good secondary user and the primary user is located at the center of the circular grid. We consider a cooperative cognitive radio environment where all secondary users can share their spectrum occupancy information and send this information to the fusion center for final decision. The secondary users broadcast their location information in order to detect unused spectrum bands. We assume that the secondary users can employ some positioning mechanisms to acquire positions, e.g., by using the global positioning system (GPS) [11]. It is assumed that the location of the primary user is also known by the secondary users. Since most modern wireless communication systems employ power control to adaptively adjust their transmit power [12], the transmit

powers of all the users are predefined and are known to all the users in the system. The cognitive radio user calculates the distance between the secondary user and other users based on location coordinates and also calculates the distance based on the received power from the primary user. If the distance calculated using the coordinates matches the distance calculated with received power level, then we can consider the user as trustworthy but if otherwise, the user is untrustworthy.

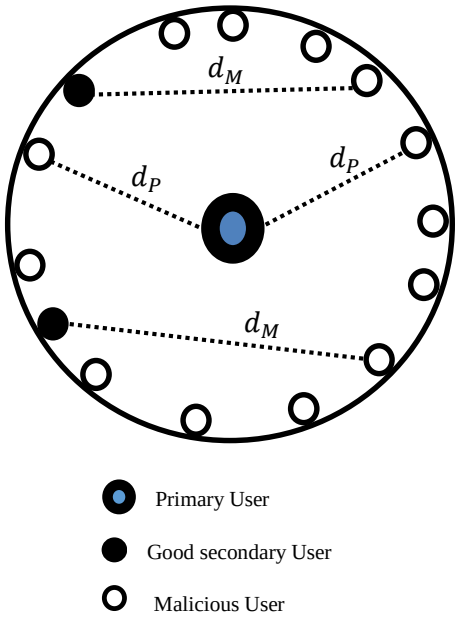



*Fig. 1. A typical cognitive radio network in a circular grid consisting of all the users in the system.*

### A. *Distance Estimated Based on Location Coordinates*

We analyze the proposed system based on location coordinates whereby all secondary users broadcast their location information. With this information, the distance between the users can be calculated.

For simplification in calculating the distance between users, we consider the location in a (2-D) plane, where $(x_{S\,i}, y_{S\,i})$ are the $x$ and $y$ coordinates of the $i^{th}$ secondary user, $(x_P, y_P)$ are the $x$ and $y$ coordinates of an existing primary user and $(x_M, y_M)$ are the $x$ and $y$ coordinates of the malicious user. The distance $d_P$ between the $i^{th}$ secondary user and primary user is given by

$$d_P = \sqrt{(x_{S\,i} - x_P)^2 + (y_{S\,i} - y_P)^2},\ i = 1,2,3 \ldots N \quad (1)$$

where $i$ is the particular secondary user and the distance $d_M$ between the $M^{th}$ malicious user and any good secondary user is also given by

$$d_M = \sqrt{(x_{m\,i} - x_S)^2 + (y_M - y_{S\,i})^2},\ i = 1,2,3 \ldots N \quad (2)$$

The decision making node can now use the estimated distance obtained using the coordinates to determine how trustworthy any of the secondary user in the system can be.

### B. *Distance measured based on received power level*

The whole idea of distance measurement by means of received signal strength (RSS) or received power level is based in ideal case on the assumption that the received power level is a function of the transmitting power and distance on the path between two radio devices. The distance between the secondary and other users can also be calculated by measuring the received power level with a known transmit power level. The received power level, $P_r$, with a given transmit power $P_t$ is given by the equation as in [13].

$$P_r(d) = P_t G_t G_r \frac{H_t^2 H_r^2}{d^4 L} \quad (3)$$

where $P_t$ is the transmit power level, $G_t$ and $G_r$ are antenna gain of both the transmitter and the receiver respectively, $H_t$ and $H_r$ are the heights of both the transmitter and receiver antennas respectively while $L$ is the system loss factor.

Considering $H_t$, $H_r$, $G_t$, $G_r$ and $L$ are constant and equal to $k$, therefore, the received power level will be solely dependent on the transmit power level and distance, expressed as,

$$P_r = \frac{P_t}{d^4} k \quad (4)$$

Based on the received power level the distance between the secondary user and the primary user is given by

$$d = \sqrt[4]{\frac{P_t}{P_r}(k)} \quad (5)$$

Hence, the distance between the users can be estimated based on the received power level given that the transmitter power is known. The distance calculated using the received power may not be 100% accurate due to the noise level and the impact of channel impediments and some other uncertainties caused by the signal propagation environment. However, many researchers still use the received power level based measurement method because of its simplicity and cost efficiency. The path loss model as proposed in [14] and [15] which is commonly used in received signal power based measurements is written as

$$\frac{P_r(d_0)}{P_r(d)} = \left(\frac{d}{d_0}\right)^n \quad (6)$$

where $P_r(d)$ is the received power at distance $d$, $P_r(d_0)$ is the received power at the reference distance $d_0$, $n$ is path loss exponent, $d$ is the distance between the transmitter and the receiver [Km], and $d_0$ is the reference distance [Km]. Due to the large dynamic range of received power levels, dBm or dBW units can be used to express received power levels.

$$[P_r\, d] = 10.log\frac{P_r(d)}{0.001(W)} \quad [dBm]$$

$$\Rightarrow P_r(d) = 10^{\left(\frac{P_r(d)-30}{10}\right)} \quad [\text{W}],$$

$$\log P_r(d_0) - \log P_r(d) = n.log\left(\frac{d}{d_0}\right)/10,$$

$$\frac{P_r(d_0)-P_r(d)}{10} = n.log\left(\frac{d}{d_0}\right),$$

$$P_r(d) = P_r(d_0) - 10.n.log\left(\frac{d}{d_0}\right) [\text{dB}], \qquad (7)$$

$$d = d_0.10^{\left(\frac{P_r(d_0)-P_r(d)}{10.n}\right)} [\text{Km}] \qquad (8)$$

$P_r(d)$ and $P_r(d_0)$ are in dBm units. Equation (7) is the so-called simplified log-normal shadowing model. Parameters $P_r(d_0)$, $d_0$, $n$, are the main parameters for log normal shadowing model formula and they define the properties of radio propagation environment. For this work, $n$ is taken as 2.8 as in [16].

In our proposed technique, the distance between a cognitive user and other users is calculated based on location coordinates and also received power level. If the distance calculated using either proposed methods matches or is extremely close to each other, then the user is regarded as a trustworthy user. If otherwise, then it will be regarded as a malicious user. The trust value is expected to be close to 1 for trustworthy users and low for untrustworthy or malicious users.

## IV. VERIFICATION OF SPECTRUM OCCUPANCY

In a typical cooperative cognitive radio network, secondary users sense if a particular spectral band is occupied or not before sending its spectrum sensing information to the fusion center for a final decision. During this process, it is imperative that these secondary users correctly sense that the spectrum is occupied by a primary user instead of a malicious user otherwise it will be sending a false spectrum sensing information to the fusion center. As a result, the trust of this particular secondary user node will be compromised.

So to verify the authenticity of the secondary user spectrum sensing information, i.e. to verify if the primary user is indeed using a specific spectral band, we propose a verification tag technique. In this technique, a verification tag is added to the primary user signal, the secondary user retrieves these verification tags from the primary user signal and uses the tag to verify whether a spectrum is currently being used by its legitimate owner or not.

The primary signal generates the following one way hash chain:

$$h_n \rightarrow h_{n-1} \rightarrow \ldots \rightarrow h_1 \rightarrow h_0, \qquad (9)$$

where $h_i = hash(h_{i+1})$ and $hash(.)$ is a hash function.

The last tag $h_0$ is broadcasted to all users, hence it is known to both the secondary users and the malicious users. The subscript $i$ of $h_i$ indicates the time index during which the primary user will transmit the tag $h_i$. At time $t = 1$, which is indicative of a short time window, the primary user transmits $h_1$. Because of the way the one-way hash chain is generated, the disclosure of $h_i$ does not lead to the disclosure of $h_j$ for $j > i$. So between time $t = 1$ and $t = 2$, the verification tag is simply $h_1$. That is the primary signal embeds $h_1$ into its signal as shown in figure 2 and during $t = 2$ and $t = 3$, $h_2$ is embedded in the signals and sent out repeatedly. The repetition is necessary because a secondary user might tend to sense the spectrum at any arbitrary moment. Once the secondary user senses the particular spectrum, it retrieves the verification tag from the signals; then using the current time and spectrum owner's $h_0$ value, the secondary user can verify the validity of $h_1$.

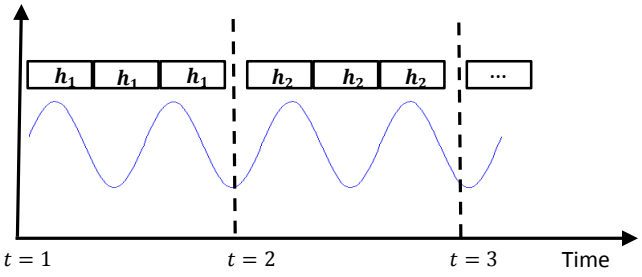



*Fig. 2. Verification tags*

If the malicious user decides to replay the verification tag into its signal so as to emulate the primary user, it will not be successful because the malicious user only replays what is sensed from the primary user. Since the goal of $h_1$ and $h_2$ is to prove to the receivers that the primary user is using the current spectrum at a specific time $t$, so when the specific time window expires, the malicious user will be needing the next verification tag to fool the secondary users which will eventually not be transmitted if the spectrum is no longer in use by the primary user. That means if the primary user is not using the spectrum, $h_{i+1}$ will not be sent out, so the malicious user will not be able to emulate the primary user hence the spectrum occupancy is verified.

## V. RELATIVE TRUSTWORTHINESS OF A USER

For spectrum occupancy sensing information to be trustworthy, it has to be received from a trustworthy user. According to the principle of object trust combination, if the final values of an object calculated by using significantly different methods are similar, then the evaluator places a higher level of trust in the results [17].

In an unfriendly environment, trust values are assigned to secondary users to ascertain and evaluate their behavior in the network. These trust values are assigned

to secondary users based on their evaluation of performance using our proposed techniques. Each time after cooperation, the behavior of the selected secondary users will be evaluated and the trust value will be updated accordingly. Then these trust values will be exchanged periodically between the users in the network. The fusion center often maintains and record identities and their corresponding trust values of all secondary users and keeps these trust values in its domain.

If a malicious user masquerades or poses as a primary user, the trust value assigned to that specific user with the aid of our proposed techniques can enable the fusion center to verify the genuineness of the spectrum occupancy sensing information being carried by that user thereby increasing the legitimacy of spectrum occupancy sensing results in the network and a more accurate detection of primary signals.

## VI. SIMULATION AND RESULTS

We use MATLAB simulation in verifying the proposed techniques and evaluate the results. To determine the location of both the primary and the malicious user, we considered a 10km by 10km area for our simulation with 4 secondary users and a malicious user present in the network. We also assume 100 instances random coordinates for 50,000 samples in the case of trustworthiness and the distance is calculated based on coordinates and received power levels.

***Table 1: Distances measured based on coordinates from each secondary user in the network.***

| Secondary User | Estimated Primary User Distance(Km) | Actual Primary User Distance (Km) | Estimated Malicious User Distance (Km) |
|---|---|---|---|
| 1st | 5.02 | 5.00 | 9.18 |
| 2nd | 5.04 | 5.00 | 6.30 |
| 3rd | 5.02 | 5.00 | 3.70 |
| 4th | 5.05 | 5.00 | 8.42 |

Fig. 3, and 4 shows the actual and estimated locations based on coordinates of the primary user and malicious user respectively. When their distances are calculated, the distance of the primary user from any of the secondary users in the network seems to be almost the same while that of the malicious user seems to vary considerably. The error arising from the estimated distance is very minimal and it is therefore not taken into consideration.

Fig. 5 shows the distance measured based on coordinates and the distance measured based on received power level of the primary user from the secondary user. We can see that both distance measurements matches considerably; that is an indication that the secondary user is actually communicating with a trustworthy user.

Fig. 6 shows the trustworthiness of a user in cognitive radio network. As the SNR value increases, correspondingly, the trustworthiness also increases. If the trustworthiness increases to 1, then we can conclude that we are communicating with the primary user and not the malicious or untrustworthy user. If the trustworthiness is nearly or approximately equal to 1, we can still conclude that it is a primary user because of some uncertainties which may tend to reduce the trustworthiness. Even as the value of SNR increases we can see from Fig. 6 that the malicious user trustworthiness remains constant at 0.6. Therefore, whatever spectrum occupancy information is given by that user, it is not taken into consideration in the final decision making process of the fusion center.

## VII. CONCLUSIONS

Trust and its management are exciting and important fields of research due to its employment on trust systems and other security and commercial applications. In our proposed method we are able to verify that the source of a spectrum sensing occupancy information is from a genuine primary user and not from a malicious secondary user masquerading to be a primary user (PUEA). It is seen from our results that high quality and trustworthiness of received spectrum sensing occupancy information is very crucial and important to the decision maker (fusion center) in a cognitive radio network so that any bias or untrue decision can be ignored or avoided.

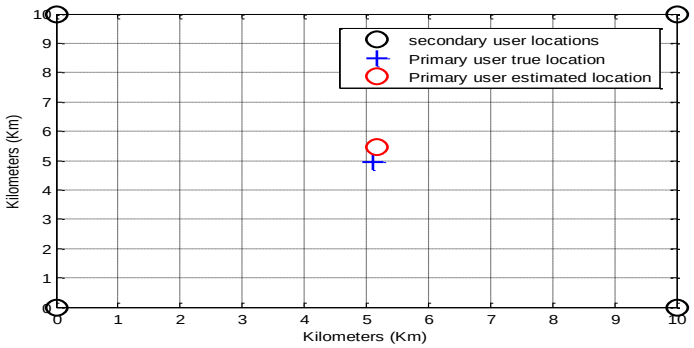



*Fig. 3 Location of the Primary user based on location coordinates*

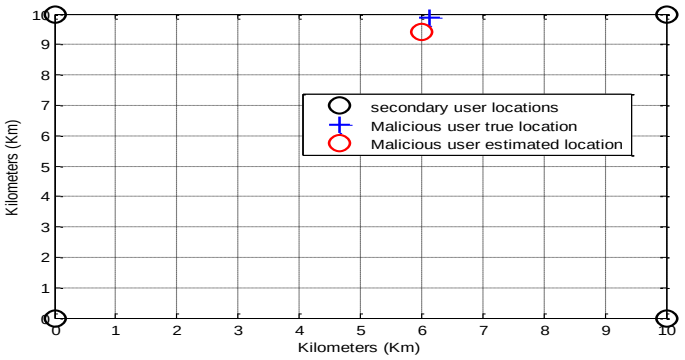



*Fig. 4 Location of the Malicious user based on location coordinates*

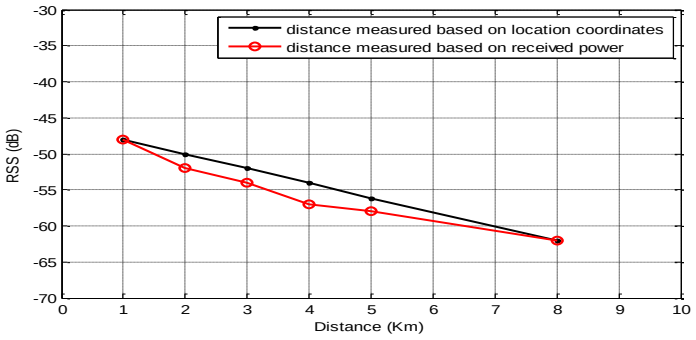



*Fig. 5 Distance measured based on location coordinates and received power*

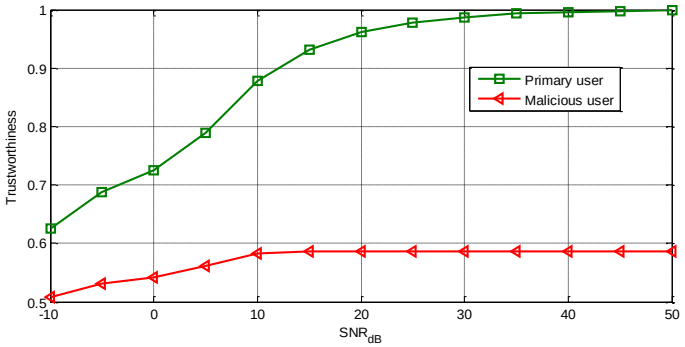



*Fig. 6 Trustworthiness of a user in a Cognitive radio network.*

## REFERENCES


**[1]** **Haykin S.** "Cognitive radio: Brain empowered wireless communication" *IEEE Journal on selected Areas in Communications* 2005; 23(2): 201-20.

**[2]** **Akyildiz, I. F, Lee. W, Vuran, M. C, and Mohanty, S.** "NeXt generation/dynamic spectrum access/cognitive radio wireless networks: A survey," (*Elsevier Journal), on computer Networks, Vol. 50, no. 13,* pp 2127-2159, Sept 2006.

**[3]** **Peh, E. and Liang, L.** "Optimization for Cooperative Sensing in Cognitive Radio Networks," in *IEEE Wireless Communications and Networking Conference*, pp. 27–32, March 2007.

**[4]** **Chen, C. Cheng, H. and Yao, D.** "Cooperative spectrum sensing in cognitive radio networks in the presence of primary user emulation attack." *IEEE Transactions on wireless communications 2011.*

**[5]** **Pehand, E. and Liang, C.** "Optimization for cooperative sensing in cognitive radio networks." *Wireless communications and networking Conference (WCNC), pp. 27-32, 2007.*

**[6]** **Chen, R. Park, J.** Ensuring trustworthy spectrum sensing in cognitive radio networks. *In Ist EEE workshop on networking technologies for software defined radio networks*. 2006. P110-9.

**[7]** **E. Orumwense, O. Oyerinde, and S. Meneny**, "Impact of primary user emulation attacks on cognitive radio networks," *International Journal on Communications Antenna and Propagation*, Vol. 4, pp. 19 – 26. April, 2014.

**[8]** **Atakli, I. M. Hu, H. Chen, Y. Ku, W. S. and Su, Z.** ''Malicious node detection in wireless sensor networks using weighted trust evaluation" *in SpringSim '08; Proceedings of the 2008 Spring Simulation Multiconference*, pp. 836-843, 2008.

**[9]** **Wang, W, Li, H. Sun, Y. and Han, Z.** "Attack-Proof Collaborative Spectrum Sensing in Cognitive Radio Networks" *Department of Electrical, Computer and Biomedical Engineering communications database (2009).*

**[10]** **Taghavi, E. and M. Abolhassani, B.** "Trustworthy Node detection in Cognitive radio in hostile environment" *International Journal of Information and Electronics Engineering Vol 3, No 2*, March 2013.

**[11]** **L. C. Wang and A. Chen,** "Effects of location awareness on concurrent transmissions for cognitive ad hoc networks overlaying infrastructurebased systems," *IEEE Trans. Mobile Comput.*, vol. 8, no. 5, pp. 577–589, May 2009.

**[12]** **A. J. Viterbi,** *CDMA: Principles of Spread Spectrum Communication*. Reading, MA: Addison-Wesley, 1995.

**[13]** **Xu, J. et al** "Distance Measurement model based on RSSI in WSN" *Communications in Wireless sensor Networks. Pp. 606 – 611, 2010.*

**[14]** **Kang, J. Kim, D. Kim, Y.** "RSS sef-calibration protocol for WSN *localization. In 2nd international symposium on wireless Pervasive Computing ISWPC'07. 2007*. San Juan, Feb 2007.

**[15]** **Botta, M. Simek, M.** Adaptive Distance Estimation Based on RSSI in 802.15.4 network. International Journal for Radio engineering, Vol.22, No. 4. December, 2013.

**[16]** National Telecommunications and Information Administration, NTIA Special Publication SP-04-409, *Proceedings of the International Symposium on Advanced Radio Technologies March 2-4, 2004*, at 101 (March 2004).

**[17]** **Zuo, Y. and Panda, B.** "Information trustworthiness evaluation based on trust" combination*" in SAC '06: Proceedings of the 2006 ACM symposium on applied computing, pp. 1880-1885, 2006.*



**Efe Francis Orumwense** received his B.Sc. (Hons) degree from the School of Engineering, University of Benin, Benin City, Nigeria in 2009. He is currently is working towards a Master's degree in the school of Electrical, Electronic and Computer Engineering, University of KwaZulu-Natal, Durban, South Africa. His research interest includes cognitive radio technology, wireless network security, and orthogonal frequency division multiplexing systems.

# Analysis of Spectral Opportunity in the UHF Terrestrial TV frequency band

Melvin Ferreira, Albert Helberg
TeleNet Research Group
School of Electrical, Electronic and Computer Engineering
North-West University, Potchefstroom Campus
Email: {melvin.ferreira, albert.helberg}@nwu.ac.za

***Abstract*—In this paper an analysis to quantify the Spectral Opportunity (SO) or so-called TV white space that can be made available to White Space Devices (WSDs) as Secondary Users (SUs) of the Ultra High Frequency (UHF) terrestrial TV frequency spectrum is presented. The system model and analysis approach is discussed and the input data and assumptions used, are detailed. The system model is applied to the entire geographical region of South Africa. Results from the analysis performed indicate that a significant number of channels in the UHF terrestrial TV frequency band (470 - 862 MHz) are available for secondary re-use by WSDs. The number of channels that are available are quantified by area and by population count and indicate that the location of the available channels are dispersed between rural and urban areas, with good availability in both.**

***Index Terms*—Dynamic Spectrum Access (DSA), secondary access, Spectral Opportunity (SO), TV white space**

## I. Introduction

WORLDWIDE administrations are under increasing pressure to keep up with demand for licensing more and more spectrum as the requirements for wireless data and communications continues to grow. The concept of Dynamic Spectrum Access (DSA) allows the utility of the spectrum to be raised, as the management paradigm of DSA dictates that this spectrum may be shared on a dynamic basis with Secondary Users (SUs) for communication purposes. Aforementioned statement holds true whether the spectrum in question is already licensed to an incumbent or service or is provisioned to be licensed in the near future.

In the context of Spectral Opportunity (SO) in the Ultra High Frequency (UHF) terrestrial TV frequency bands, also referred to as TV white space, the Independent Communications Authority of South Africa (ICASA) has expressed early interest in the possibilities of this technology and its possible utility in exploiting the SOs in the terrestrial TV frequency spectrum [1], [2].

An SO can be formally be defined as the existence of a frequency band segment, satisfying an availability criterion, that DSA capable devices can exploit for their communications purposes [3]. In terms of secondary access to the terrestrial TV frequency bands, the availability criteria are typically as follows:

- The White Space Device (WSD) must not cause harmful interference to the Primary User (PU) (incumbent) TV transmitters or receivers.
- The WSD must accept interference from the PU (incumbent) TV transmitters and receivers.

In our previous work [4], the authors quantified the probability of finding SO in the form of contiguous channels in the UHF terrestrial TV frequency band. The fraction of geographical locations as well as the fraction of the population that has access to contiguous channels of a specified length, was given. Keeping in mind that the channel bandwidth in the UHF terrestrial TV frequency band is 8 MHz, it was noted that in excess of 5 contiguous channels (40 MHz), was available to a large fraction of the population and a large number of geographical areas.

In this paper the authors extend and build forth on their previous findings by analysing the amount of SO or so-called TV white spaces that are available in the UHF terrestrial TV frequency band. In this paper the availability of single 8 MHz channels are considered, arguing that these channels present significant utility in the context of secondary re-use, i.e. for WSDs as defined in IEEE 802.22 [5]. To this end the number of channels available on national level are quantified, weighting the number of channels available by area as well as population that are available as SOs.

The rest of the paper is organised as follows: Section II details relevant related work on the national level. Section III briefly touches on the most important design aspects of the system model. Section IV discusses the overall system methodology, detailing the analysis parameters, and metrics used. Finally section V presents the analysis results.

## II. Related work

Nationally, TV white space is gaining increased attention and awareness in South Africa. ICASA has been working in collaboration with key stakeholders on the launch of a TV white space trial network in the greater Cape Town area [6]. This white space trial has now been completed and there is plans for a second trial in the Limpopo province [7]. The final reports on trial indicate positive results with regards to co-existence and negation of interference to the incumbent. Suggested technical rules and parameters for the use of WSDs in the UHF terrestrial TV spectrum has also been put forth [7].

Furthermore, with the exception of our previous work in [8] and [4], to the best of our knowledge, no further work on the provincial and/or national quantification of SO available in the

UHF terrestrial TV frequency bands in South Africa, has been conducted.

## III. System model

### A. Propagation Model

The propagation model used in the analysis is the ITU-R P.1546 model. The model is a statistical prediction model that is derived from experimental data. Our model choice is motivated by the widespread acceptance and use of ITU-R P.1546 for network planning of terrestrial broadcast services [9].

Our implementation of the model complies with revision 4 of the model, with the exceptions of sea based and land/sea based propagation paths. These path modes are not implemented because they are not required for the analysis region. ITU-R P.1546 predicts the mean electrical field strength at a specified distance from the transmitter under analysis. The model predicts the field strength values for a specified location probability and time probability. Our implementation has been verified and yields the same prediction curves as the reference curves found in the recommendation [10]. For a more detailed treatment of the propagation model implementation, refer to our previous work in [8].

### B. TV transmitter database

The TV transmitter database is populated with information obtained from the Final Terrestrial Broadcasting Plan of 2008 [11], the two erratums to the Final Terrestrial Broadcasting Plan [12], [13] and the ICASA TV transmitter database. The Final Terrestrial Broadcast Plan and erratums are publicly available in the government gazettes, whilst the TV transmitter database was made available for research purposes by ICASA.

The PU database includes the following entries as obtained from the sources mentioned above:

- All analogue channel entries with service status operational (OPE / OP) or licensed (LIC / LI), as found in the ICASA TV transmitter database.
- All digital channel entries as found in the Final Terrestrial Broadcasting Plan of 2008.

It is important to note that the TV transmitter database is regarded as a living document. New entries are added continually and technical parameters of the present channel entries may change. Technical parameters of analogue and digital entries may change as new Digital Terrestrial Television (DTT) sites are deployed. Furthermore, analogue switch-off still needs to take place.

These imminent changes will affect the temporal and geographical relevance of predictions made, using the given input data. The results obtained with the system model should therefore always be viewed with due consideration of the input data used.

### C. Analysis region exclusions

In addition to the terrestrial TV transmitters and receivers, the radio astronomy service also operates in the UHF terrestrial TV frequency bands and require protection from harmful interference. The Square Kilometre Array (SKA) is afforded protection from harmful RFI by other services through the Astronomy Geographic Advantage Act [14].

The use of any Radio Frequency (RF) equipment without the necessary permissions in the core advantage area is prohibited by law [15]. The proposed Karoo central radio astronomy advantage area is divided into areas 1, 2 and 3 for frequency bands 70-1710 MHz, 1.71-6 GHz and 6-25.5 GHz respectively. Of interest to our work is the proposed central advantage area 1 and coordinated advantage area 1 [16]. For completeness of the analysis results, the effects of excluding the core and central astronomy advantage areas from the analysis region are considered.

### D. Temporal analysis scenarios

Given the input data considered for the analogue and digital service, the SO can be analysed for three discrete points in time.

- The SO before the start of the dual illumination period, i.e. the analogue service only, denoted by $t_A$.
- The SO right after analogue switch-off, i.e. the digital service only, denoted by $t_D$.
- The SO at a predefined point in time in-between, i.e full dual illumination right before analogue switch-off, denoted by $t_{DU}$.

These three temporal points in time can be defined by the set $\mathcal{T} = \{t_A, t_{DU}, t_D\}$. The analysis will be performed for these discrete points in time in order to compare the change in SOs with time.

### E. TV transmitter protection requirements

For the protection requirements of the DTT service the recommendations as described ITU-R BT.1368 [17] are followed. Typically, the protection requirements for the digital service is specified at the point of signal failure.

For the protection requirements of the the analogue TV service the recommendations as provided in ITU-R BT.417 [18] are followed. These protection requirements are specified to ensure an acceptable perceived picture quality for the TV signal received. Due to the analogue nature of the signal increasing interference will cause gradual reduction in the perceived picture quality, however.

The UHF terrestrial TV channel assignments with their corresponding frequency ranges and ITU band allocation are channels 21-34 (470-582 MHz) for band IV and channels 35-68 (582-854 MHz) for band V. The range of minimum electrical field strength ($E_{min}$) values as determined from aforementioned ITU-R recommendations are displayed in table I. The variation of field strength values within each band are due to the frequency dependant term that changes with the channel number.

### F. Availability criterion formulation

In line with the definition given for an SO in the introduction of this paper, the availability criterion should be chosen as such so as to prevent WSDs from causing harmful interference to

TABLE I
MINIMUM ELECTRICAL FIELD STRENGTH [17], [18]

| | Band | |
|---|---|---|
| $E_{min}$ (dB$\mu$V/m) | IV | V |
| Analogue | 62 - 63,7 | 63.8 - 67,1 |
| Fixed DTT | 44,4 - 46,1 | 46,3 - 49,5 |
| Mobile DTT | 48,4 - 50,1 | 50,3 - 53,5 |

the incumbent or PU transmissions. The availability criterion as formulated by [19] and [20] is adopted for use in this work.

The availability criterion can be better described by means of fig. 1. The two circles represent two PU TV transmitters, transmitting on the same channel. The noise-limited protected contour radius of a PU, i.e. before the introduction of WSDs to the same spectrum, is denoted by $r_{nl}$. It is argued in [20] and [19] that all locations where TV reception was possible prior to the appearance of WSDs (i.e. $r_{nl}$) cannot be declared protected. To account for this, [20] introduces the concept of an erosion margin. The protected contour of the PU, taking the erosion margin into account, is denoted by $r_p$. Effectively the erosion margin accounts for the degradation in the noise-limited protected contour radius $r_{nl}$.

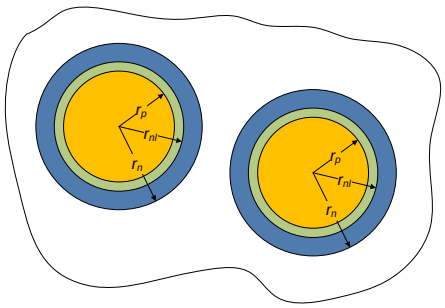



Fig. 1. Representation of the availability criterion [20]

Radius $r_n$ includes an additional keep-out region beyond radius $r_p$, which is a function of the WSD transmitter power, and WSD antenna height Above Ground Level (AGL). The additional keep-out region provides an additional buffer zone in which PU transmitters and receivers may not seek protection, but WSDs may also not transmit. The keep-out region accounts for a worst case placement of WSD relative to PU receivers at the edge of the protected contour ($r_p$), to ensure that the PU protection requirements are not violated.

However, to determine an upper bound on the available SO, the case of an WSD transmitting at a hypothetical 0 W considered. Fig. 1 shows that this has the same effect as letting $r_n = r_p$, since an WSD transmitting at 0 W would not require an additional keep-out region.

Harmful interference is therefore defined as any interference that erodes the noise-limited protection requirements ($r_{nl}$) of the PU service by more than $\psi$ dB. The protected contour ($r_p$) is then set as follows:

$$E_{protect} = E_{min} + \psi \tag{1}$$

where $E_{protect}$ is the field strength at the protected contour, $E_{min}$ is the minimum electrical field strength for the relevant service to be protected (table I) and $\psi$ is the allowable PU erosion margin in dB. For the analysis presented $\psi = 0.1$ dB.

The grid system used in the system model is defined in the geographic coordinate system. Note that the ground distance of such a system varies with latitude. With this in mind, let the analysis region in question be described in a geographical coordinate system, tied to a square grid system, with grid spacing $\Delta x_{degree}$. Let the grid system contain grid cells $i = 1, .., n$.

SO is a function of time, frequency, and space. Let the predicted electrical field strength of the PU transmitter be denoted by $FS_i(c,t)$, where the field strength at the $i$th grid cell is a function of the time ($t$) and the channel ($c$) selected. $FS_i(c,t)$ is determined by considering the effect of all entries in the PU data structure at time $t$ and on channel $c$ for the geographical location of grid cell $i$.

The maximum of the field strengths predicted for channel $c$, time $t$ and location $i$ is in each instance considered the wanted field strength and assigned to $FS_i(c,t)$. The SO of a given grid cell is regarded as a spatio-temporal bin, denoted by $x_i(c,t)$. The availability criterion is then modelled as a binary value, which can be expressed as a unit step function:

$$x_i(c,t) = \begin{cases} 1, & \text{if } FS_i(c,t) < E_{protect} \\ 0, & \text{otherwise} \end{cases}, \quad c \in \mathcal{C}, \quad t \in \mathcal{T} \tag{2}$$

where $x_i(c,t)$ is the co-channel availability decision for grid cell $i$, channel $c$ and time $t$, evaluated for the co-channel protected region $E_{protect}$ as defined in equation 1. The set $\mathcal{T}$ contains the temporal points for which the SO is analysed, as discussed in section III-D. The set $\mathcal{C}$ contains the channel numbers of channels considered in the analysis, so that:

$$\mathcal{C} = \{21, 22, 23, .., 67, 68\}$$

where $\mathcal{C}$ denotes all the channels in band IV and V (470-854 MHz), with a channel bandwidth of 8 MHz each.

## IV. METHODOLOGY

### A. Analysis parameters

The system model parameters as used in this analysis are detailed in table II. The grid spacing of 60 arc-seconds corresponds to a quadrangle of approximately 1.85 km$\times$1.85 km at the equator. The maximum effective height, $h_{max}$, is used as the input for the effective antenna height parameter. The procedure to determine $h_{max}$ from topographical data is provided in [10]. The void-filled 3 arc-second SRTM v4 Digital Elevation Model (DEM) [21] is utilised to extract elevation values of the topography surrounding each transmitter site to compute aforementioned parameter. The other parameters denoted in table II relate to propagation model settings and assumptions.

The protection requirements for the analogue, fixed DTT and mobile DTT service are calculated using the process as described in section III-E. The protected contour in equation 1 is determined by setting the erosion margin, $\psi$ as specified. From using these parameters the availability decision in equation 2 can be applied and the necessary SO metrics can be determined.

TABLE II
SYSTEM MODEL PARAMETERS FOR ANALYSIS

| Parameter | Value |
|---|---|
| Reference ellipsoid | WGS84 |
| Grid spacing ($\Delta x_{degree}$) | 60 arc-seconds |
| Propagation path | Land only |
| Location probability | 50% |
| Time probability | 50% |
| Diffraction loss correction | Terrain approximation method |
| Tropospheric scatter correction | No |
| TCA correction | No |
| Receiver antenna height | 10 m |
| Transmitter antenna pattern | Omnidirectional |
| Effective antenna height | Maximum effective height, $h_{max}$ |

### B. Metrics

The general formulation for the weighted mean of a number of samples is given by [22]:

$$\overline{x}_w = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i} \tag{3}$$

where $\overline{x}_w$ is the weighted mean, $x_i$ is the $i$th sample, $w_i$ is the corresponding weight of the $i$th sample and $n$ is the number of samples.

*1) SO weighted by area:* To quantify the SOs weighted by area, equation 3 can be rewritten as follows:

$$\overline{s}_a = \frac{\sum_{i=1}^{n} a_i s_i}{\sum_{i=1}^{n} a_i} \tag{4}$$

where $\overline{s}_a$ is the mean available channels weighted by area, $s_i$ is the number of available channels in the $i$th grid cell, $a_i$ is the corresponding area of the $i$th grid cell and $n$ is the number of grid cells in the analysis region.

The area of the each of the grid cells are estimated by projecting the analysis region onto an equal area projection, and then estimating the area on a flat surface by applying Green's theorem [23].

*2) SO weighted by population:* To quantify the SOs weighted by population, equation 3 can be rewritten as follows:

$$\overline{s}_p = \frac{\sum_{i=1}^{n} p_i s_i}{\sum_{i=1}^{n} p_i} \tag{5}$$

where $\overline{s}_p$ is the mean available channels weighted by population, $s_i$ is the number of available channels in the $i$th grid cell, $p_i$ is the population count residing in the $i$th grid cell and $n$ is the number of grid cells in the analysis region.

To obtain the population counts for each grid cell, the population data from Census 2001 on enumeration area level is used [24]. The population count is determined by implementing an areal weighting technique to approximate the population count for each grid cell from the original Census enumeration areas [25], [26].

TABLE III
WEIGHTED MEAN VALUES FOR UHF SO IN SOUTH AFRICA

| | Core area excluded | | Central area excluded | |
|---|---|---|---|---|
| Time | $\overline{s}_a$ | $\overline{s}_p$ | $\overline{s}_a$ | $\overline{s}_p$ |
| $t_A$ | 45.6 (95%) | 40.4 (84%) | 45.4 (95%) | 40.4 (84%) |
| $t_{DU}$ | 40.2 (84%) | 33.7 (70%) | 39.7 (83%) | 33.7 (70%) |
| $t_D$ | 42.5 (89%) | 40.9 (85%) | 42.1 (88%) | 40.9 (85%) |

## V. RESULTS

The analysis results for the SO in the UHF terrestrial TV frequency bands are now discussed and elaborated upon for South Africa. The results presented give an upper bound on the available SO in a given frequency band, based on the availability criterion that was developed in section III-F. The results is compared for three distinct temporal points in time, as was discussed in section III-D.

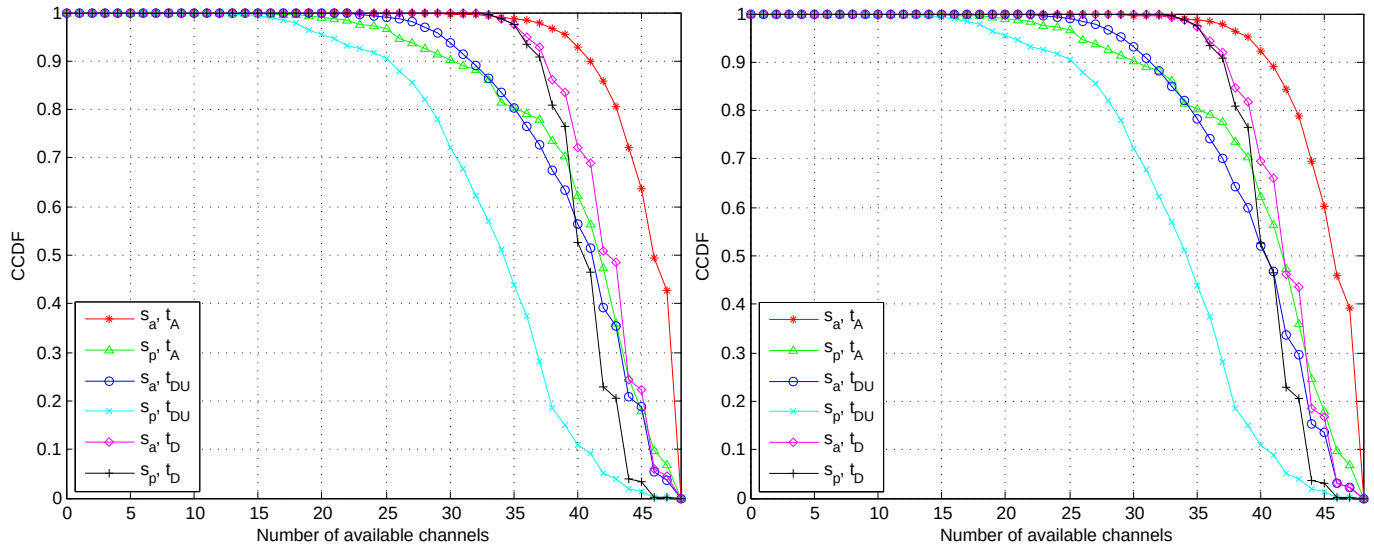
Table III details the mean SO weighted by area and population for the the three temporal scenarios, excluding the core and central advantage area from the analysis region. The mean values reported in each instance for the whole of South Africa appear promising. On national level, the additional area excluded by the central advantage area is seen to not influence the population mean ($\overline{s}_p$) values as much as the area mean ($\overline{s}_a$) values, indicating that the area excluded has a low population density, relative to the rest of the country.

Table III indicates that the mean SO recovered after analogue switch-off (i.e. time $t_D$) is in each instance less than before the commencement of dual illumination (i.e. time $t_A$). Interestingly, the population weighted mean values suggest that the SOs recovered after analogue switch-off by the co-channel criterion is more than before the commencement of dual illumination.

Aforementioned observation can be explained by the revisiting the TV transmitter protection requirements as discussed in section III-E. From table I it can be seen that the minimum electrical field strength ($E_{min}$) required for successful reception of a DTT signal is lower than for an analogue signal. Considering a site with the same transmitter Effective Radiated Power (ERP) for analogue and digital transmissions, the service radius that needs to be protected with a digital transmission will be larger due to the minimum electrical field strength requirement.

The empirical Complementary Cumulative Distribution Functions (CCDFs) for the number of available channels, weighted by area and population, are shown for the co-channel availability criterion for the core area (fig. 2a) and the central area exclusions (fig. 2b) respectively. The CCDFs indicate the distributions for the three temporal scenarios, which will now be discussed accordingly.

Comparison of fig. 2a and fig. 2b reveals that the CCDF curves follow the same trend, albeit with different probabilities for a given number of channels available.




(a) CCDF: Core area excluded

(b) CCDF: Central area excluded

Fig. 2. UHF band: SO for South Africa

Fig. 2a indicates for time $t_A$, that at least 13 channels are available to 100% of the population and at least 15 channels are available at 100% of the locations. At least 31 channels are available to 90% of the population, whilst at least 42 channels are available at 90% of the locations.

For time $t_{DU}$, the SO decreases to at least ten channels for 100% of the population and at least 12 channels for 100% of the locations. At least 32 channels are available at 92% of the locations and to 68% of the population. At least 26 channels are available to 90% of the population.

For time $t_D$ the SO increases, with at least 31 channels available to 100% of the population and locations. At least 38 channels are available at 93% of the locations and to 91% of the population.

The results for fig. 2b follow the same trend and absolute results for the availability for 100% of the population and locations. Availability of channels to the population ($s_p$) remains the same for time $t_A$, $t_{DU}$ and $t_D$ as discussed in the preceding paragraph. However, the availability of channels at locations ($s_a$) is lower.

Refer to fig. 2b. For time $t_A$, at least 42 channels are available at 89% of the locations. For time $t_{DU}$, at least 32 channels are available at 91% of the locations. For time $t_D$, at least 38 channels are available at 92% of the locations.

It follows from comparison of the results discussed above for the SKA core and central advantage area exclusions that the SO on a national level is still significant, even if legislation is put in place to prevent additional Radio Frequency Interference (RFI) from originating within the central advantage area. It should be emphasised that the draft regulations at the time of writing suggest that additional coordination of RFI in coordinated zone 1, i.e. outside the central advantage area, will only be required from transmitters with an ERP exceeding 60 dBm [16]. This ERP is unlikely to be exceeded by any WSD equipment intended for the market.

## VI. Conclusion

In this paper the authors presented a system model that is capable of determining the SO available in a given frequency band. The system model is applied to the entire geographical region of South Africa to quantify the SO or so-called TV white space that can be made available to WSDs as SUs of the UHF terrestrial TV frequency spectrum. The analysis results indicates that significant SO exists in the UHF terrestrial TV band and that a significant number of channels are available to a high percentage of the population and geographical locations. The authors are of the opinion that there are definite utility for these SOs in the context last-mile broadband access, given the low penetration rate of broadband in general.

Future work includes incorporating the amendments of version 5 into our ITU-R P.1546 implementation, evaluation of the secondary use case where a WSD emits radiation at a given ERP above 0 W, and assessing what the effect of changes proposed in the Draft Terrestrial Broadcasting Frequency Plan of 2013 [27] will have on the available SO after analogue switch-off.

## Acknowledgement

The authors would like to acknowledge the financial support of THRIP and the Telkom CoE at the North-West University. The authors would also like to thank the anonymous reviewers for their constructive feedback.

## References

[1] "Draft policy directions on exploiting the digital dividend," in *Government Gazette*, Republic of South Africa, 14 Dec 2011, no. 34848, notice 898.

[2] "Notice inviting comments regarding the 2nd draft frequency migration regulation and radio frequency migration plan," in *Government Gazette*, Republic of South Africa, 24 Dec. 2012, vol. 36031, notice 1064.

[3] "IEEE standard definitions and concepts for dynamic spectrum access: Terminology relating to emerging wireless networks, system functionality, and spectrum management," *IEEE Std 1900.1-2008*, pp. c1 –48, 26 2008.

[4] M. Ferreira and A. Helberg, "Contiguous spectral opportunity available in the UHF terrestrial TV frequency bands," in *Southern African Telcommunications and Network Access Conference (SATNAC)*, Stellenbosch, South Africa, Sep 2013.
[5] *IEEE Standard for Information Technology–Telecommunications and information exchange between systems Wireless Regional Area Networks (WRAN)–Specific requirements Part 22: Cognitive Wireless RAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Policies and Procedures for Operation in the TV Bands*, IEEE Std. 802.22-2011, Jul. 2011.
[6] M. Masonta, R. Makgothlo, and F. Mekuria, "Setting the scene for TV white space and dynamic spectrum access in South Africa," in *IST-Africa Conference Proceedings*, 2012.
[7] (2014, Aug) The Cape Town TV white spaces trial. TENET (Tertiary Education and Research Network of South Africa). [Online]. Available: http://www.tenet.ac.za/about-us/the-cape-town-tv-white-spaces-trial
[8] M. Ferreira and A. Helberg, "Spectral opportunity modelling in the terrestrial broadcast frequency spectrum," in *Southern African Telcommunications and Network Access Conference (SATNAC)*, George,South Africa, Sep 2012.
[9] "Final acts of the regional radiocommunication conference for planning of the digital terrestrial broadcasting service in parts of regions 1 and 3, in the frequency bands 174-230 MHz and 470-862 MHz (RRC-06)," in *RRC-06*. ITU, 2006.
[10] *Method for point-to-area predictions for terrestrial services in the frequency range 30 MHz to 3000 MHz*, ITU-R Recommendation P.1546-4.
[11] "Final terrestrial broadcasting frequency plan, 2008," in *Government Gazette*, Rebuplic of South Africa, 18 Nov 2009, no. 32728, notice 1538.
[12] "Erratum to the final terrestrial broadcasting plan," in *Government Gazette*, Rebuplic of South Africa, 24 Jun 2011, no. 34386, notice 414.
[13] "Second erratum to the final terrestrial broadcasting plan of 2008," in *Government Gazette*, Republic of South Africa, 21 Dec. 2012, vol. 36026, notice 1061.
[14] "Astronomy geographic advantage act, 2007," in *Government Gazette*, Rebuplic of South Africa, 17 Jun 2008, no. 31157, notice 666.
[15] "Regulations to prohibit or restrict certain activities in core astronomy advantage areas in terms of the astronomy geographic advantage act, 2007," in *Government Gazette*, Rebuplic of South Africa, 22 Jun 2012, no. 35450, notice R465.
[16] "Notice of intention to make regulations regarding the karoo radio astronomy advantage area for square kilometre array radio telescope," in *Government Gazette*, Republic of South Africa, 5 Feb 2009, vol. 31855, notice 118.
[17] *Planning criteria, including protection ratios, for digital terrestrial television services in the VHF/UHF bands*, ITU-R Recommendation BT.1368-9.
[18] *Minimum field strenghts for which protection maty be saught in planning an analogue terrestrial television service*, ITU-R Recommendation BT.417-5.
[19] J. van de Beek, J. Riihijarvi, A. Achtzehn, and P. Mahonen, "TV white space in Europe," *Mobile Computing, IEEE Transactions on*, vol. 11, no. 2, pp. 178–188, 2012.
[20] M. Mishra and A. Sahai, "How much white space is there?" EECS Technical Report, Tech. Rep. EECS-2009-3, January 2009. [Online]. Available: http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-3.html
[21] A. Jarvis, H. Reuter, A. Nelson, and E. Guevara. (2008) Hole-filled seamless SRTM data V4, international centre for tropical agriculture (CIAT). [Online]. Available: http://srtm.csi.cgiar.org
[22] D. Wong and J. Lee, *Statistical Analysis of Geographic Information with ArcView GIS and ArcGIS*. John Wiley & Sons, Inc., 2005.
[23] J. Stewart, *Early Transcendentals Multivariable Calculus*, 5th ed., K. Townes, Ed. Thomson Wadsworth, 2003.
[24] (2012, Nov) Census 2001: Spatial metadata for digital spatial boundaries. Statistics South Africa. [Online]. Available: http://www.statssa.gov.za
[25] M. Goodchild and N. Siu-Ngan Lam, "Areal interpolation: a variant of the traditional spatial problem." *Geo-Processing*, vol. 1, no. 3, pp. 297–312, 1980.
[26] M. Goodchild, L. Anselin, and U. Deichmann, "A framework for the areal interpolation of socioeconomic data," *Environment & Planning A*, vol. 25, no. 3, pp. 383–397, 1993.
[27] "Notice on the update of the terrestrial broadcasting frequency plan 2013," in *Government Gazette*, Republic of South Africa, 2 Apr. 2013, vol. 36321, notice 298.

**Melvin Ferreira** is a lecturer in Computer Engineering at the North-West University. He received his Ph.D. in 2013, M.Eng (cum laude) in 2010 and B.Eng (cum laude) in 2008 from the the North-West University. His research interests include dynamic spectrum access, spectrum sharing, spectrum management, propagation modelling, testbed experimentation, packet-optical transport and network planning.

# Impact of Regulation and Policy on Secondary User Pricing Strategies in a Cognitive Radio Environment in South Africa

E. Naidu, R Van Olst
School of Electrical & Information Engineering, University of the Witwatersrand
Johannesburg, South Africa
Tel: +27 82 6159310, Fax: +27 86 719 7012
email: {elicia.naidu@gmail.com, rex.vanolst@wits.ac.za}

***Abstract*** **–There has been a growing demand for spectrum availability due to inefficient management of the radio frequency spectrum and underutilization of all spectrum bands. Spectrum has been managed with the same approach for over the last decade and only recently due to the phenomenal growth in mobile and broadband communications, has it has been given attention. In this paper game theory along with economic theory is used to analyse the relationships/cooperation between the users and service providers to determine the best pricing strategy for secondary users. The pricing strategies modelled and simulated in MATLAB include the market-equilibrium pricing strategy and the competitive pricing strategy. These two strategies are chosen as they were the most relevant in South Africa at the time. The two pricing strategies are compared in terms of advantages and disadvantages as well the revenue earned by each of the primary services and it is shown that the competitive pricing strategy produces greater revenue than the market-equilibrium pricing strategy for certain bandwidths. The influence of telecommunication regulations and policy in South Africa on the strategies are analysed and discussed. A survey was used to determine the role-players reaction to spectrum trading, it is viewed as desirable in South Africa and the strategy that would best suit the South African market at present is the competitive pricing strategy. It was also worthwhile investigating which band should be used for spectrum trading.**

**Key words: Cognitive radio, spectrum trading/sharing, opportunistic spectrum access, South African Regulations**

## I. INTRODUCTION

The radio frequency spectrum is a natural resource critical for delivering electronic communication services and for building a knowledge-driven economy and society. With the growing demands for spectrum in the country and the poor assignment of spectrum bands, it can be regarded as a scarce resource. In order to improve the utilization of the radio frequency spectrum, intelligent wireless communication systems such as cognitive radio technologies can be utilized. Cognitive radios are able to improve the capability of a wireless receiver by allowing it to operate in multiple frequency bands using multiple transmission bands.

Studies have shown that while some frequency bands are heavily used there are many bands that are unoccupied most of the time, dependent on geographical areas and population. These potential unoccupied bands can be optimised by use by unlicensed users.

Previous works have focused on the use of Cognitive Radio and currently there has been a drive to focus on the various pricing strategies for spectrum sharing between primary and secondary users. Nel and Zhu [1] as well as Hossain and Niyato [2] have performed extensive work on pricing strategies pertaining to opportunistic spectrum access.

Through spectrum trading, the licensed users (or primary service provider) are able to sell a portion of the unused spectrum to the unlicensed users (or secondary service provider) for a price. Pricing for both the licensed users selling spectrum and the unlicensed users buying the spectrum is important. Therefore, an optimal and stable solution for spectrum trading in terms of price and allocated spectrum is required to maximise the revenue of the seller and utility of the buyer while still satisfying both the seller and buyer and their solutions. Game and economic theory are used to analyse the relationships and interactions between the players (primary and secondary users). Concepts such as utility theory, market-equilibrium, oligopoly market and auction theory define the incentive for licensed users to yield the right of spectrum access to the unlicensed users.

Important concerns in spectrum trading are policy, regulation and spectrum management. Spectrum regulations and policies define rules of cooperation between primary and secondary users and spectrum management has been practised around the world since the 1920's. The use of spectrum is regulated by the Independent Communications Authority of South Africa (ICASA) and the management and development of the spectrum plays an important role in developing a knowledge-driven economy and society. The regulator directly impacts the pricing strategy used in the country by the regulations and policy it enforces.

In South Africa, the spectrum management arrangements are a shared responsibility between the policy maker and the regulatory authority, i.e. the Department of Communication (DOC) coordinates spectrum for government services, whilst ICASA regulates all other spectrum requirements. Currently spectrum trading is seen as illegal and is being investigated by ICASA and the DOC.

The rest of this paper is organised as follows, section II provides a motivation for this work, while section III reviews the related work. The system model and pricing strategies are discussed in section IV. Section V provides a background to spectrum trading regulations and policy. The numerical performance and results are discussed in section VI. The social and economic impacts of spectrum sharing are discussed in section VII and the conclusions are stated in section VIII.

## II. MOTIVATION

With the exponential growth in wireless services and technologies, the demand for radio spectrum is steadily increasing. With the current spectrum management policy in which spectrum bands are assigned statically, there is the issue of crowded spectrum as well as underutilization of spectrum at various times and bands. Spectrum trading is seen as one approach to improving the efficiency of spectrum use, and one of the key issues related to spectrum trading is finding the optimal pricing strategy for primary and secondary spectrum users.

This work is motivated by consideration of dynamic spectrum access from an economic perspective and can benefit the regulator in providing them with a perspective that considers both the regulatory issues as well as engineering concerns.

## III. RELATED WORKS

With the use of different game models, a number of proposals have been made regarding pricing strategies and/or perspectives. Nel and Zhu [1] discuss the expansion of the work performed in [3] to model the use of the opportunistic spectrum access allowing secondary users to share the spectrum resources with primary users on an opportunistic basis using a three player Stackelberg game model. The simulation results show that under certain assumptions, dynamic spectrum access with secondary sharing can greatly improve the revenue for the service provider when the channel is under-utilised or over utilised; however compensating the primary users due to interference caused by secondary users could result in a loss of revenue to the service provider. The paper further shows that by exploiting the secondary user's willingness to pay (in demand by the secondary users), the service provider could earn more revenue from fewer secondary users when the channel utilization of the primary users is lower. This is due to the fact that secondary users who can be allocated more bandwidth with better channel conditions are willing to pay more. [1]

Niyato and Hossain present three different pricing and market models: market-equilibrium, cooperative and competitive as discussed in [4]. These models are used to compare the prices offered by the service providers at equilibrium as well as the revenue attained at equilibrium. The model assumes a primary service provider services a number of primary users as well as secondary users. The service providers can set the offered price accordingly with one of the three pricing models. The market-equilibrium and cooperative pricing models are based on optimization problems whereas the competitive pricing model is based on a non-cooperative Bertrand game assuming the players are selfish and compete against one another for price. The authors simulate static and dynamic models of the three pricing strategies and determine that the cooperative pricing strategy returns the highest revenue with the lowest stability whereas the market equilibrium pricing strategy returns the lowest revenue with the highest stability. [4]

The Bertrand model used for competitive pricing is expanded in [5] and [6]. In [5] D Niyato and E Hossain discussed the problem of spectrum sharing among primary users and multiple secondary users and used a non-cooperative Bertrand game to model a spectrum overlay-based cognitive radio wireless system with one primary user and a number of secondary users. A static and dynamic game is simulated and compared and the inefficiency of Nash Equilibrium is explored. The major observations in this paper are that the spectrum sharing solution in case of the dynamic strategy adaptation depends on the given system parameters as well as the algorithmic parameters (e.g. learning rate) and the Nash Equilibrium does not necessarily maximise the total revenue of the secondary users, however it does provide a fair solution. To expand on the concepts examined in [5], the same authors further investigate the inefficiency of the Nash Equilibrium as well as the concept of collusion and shows that collusion returns higher revenue than Nash Equilibrium by ensuring the primary services are aware of punishment due to deviation by properly weighting the revenue in the future. [6]

## IV. THE MODEL

The system considered is from [4] where secondary users can opportunistically exploit the wireless spectrum licensed to primary users as shown in Figure IV-1. Figure IV-1 illustrates the basic system design of spectrum sharing between the primary and secondary users.

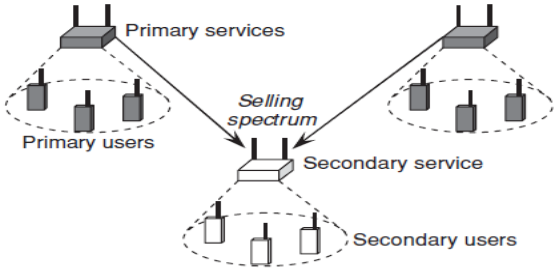



Figure IV-1: Illustration of spectrum trading [2]

The revenue of a primary service provider depends not only on the cost of sharing the spectrum with the secondary service providers (e.g. due to performance degradation of primary users), but also on the strategy chosen by other primary service providers. The price, $p_i$, can be set by three different pricing strategies, namely market-equilibrium, competitive and cooperative pricing strategies. These different pricing strategies result in different behaviors of the service providers in achieving the best and stable decisions.

For the purpose of this paper, a duopoly is considered, therefore there are only two primary services each servicing a number of primary users and a group of secondary users. The application of these pricing strategies extend beyond two-user games, however the assumptions, derivations and submissions here can be generalized to multi-user scenarios.

If spectrum trading is allowed on the network, a primary service provider has two sources of revenue, the primary service users and the secondary users. Although spectrum trading can generate higher revenue to the primary service provider by selling spectrum to secondary users, it comes at the cost of degraded QoS performance on the primary users. This is due to interference caused by secondary users sharing the radio spectrum with primary users.

### A. *Market-Equilibrium Pricing Strategy*

It is assumed that the primary service provider is not aware of any others and hence there is no competition or cooperation and the spectrum price is set naively based on spectrum demand from the secondary users (demand function). The price is set based on the willingness of the primary service provider to sell spectrum and this is determined by the supply function. The supply function indicates the size of radio spectrum shared by a primary user with the secondary user, whereas the demand function indicates the size of radio spectrum required by secondary users. [2] The equilibrium point is where the supply is equal to the demand.

As a primary service provider is unaware of the existence of other primary service providers, it can be noted that the supply function is independent of the prices offered by other primary service providers. [2]

### B. *Competitive Pricing Strategy*

Each of the primary service providers is aware of the competition amongst each other and each of the primary service providers aims to maximise their own revenue. The primary service providers compete through price adjustment, i.e. given the spectrum prices offered by other primary service providers, one primary service provider will choose the price for its own spectrum such that its individual profit is maximised. [2]

To model the competition for price among the primary service providers, a non-cooperative game model is used where the players (sellers in an oligopoly market) are the primary service providers, the strategy of each player is the price offered by unit of spectrum and the payoff for each player is the individual profit due to spectrum trading under competition to the secondary users. [2]

The solution of this game is Nash Equilibrium and this can be obtained by using the best response function of the players, which is the best strategy of one player given the other's strategies. [4]

## V. SPECTRUM TRADING REGULATIONS AND POLICY

Spectrum management involves technical and regulatory mechanisms that are designed to achieve the optimal use of the radio frequency spectrum with the key purpose of maximising the value society gains from the radio spectrum. Earlier techniques of spectrum management may have been effective when utilising radio communication systems however due to the technological progress and innovative applications to utilise radio spectrum, the spectrum management process has become out-dated as it has not kept up with the major changes in technology. This has led to growing technical and economic inefficiencies as well as obstacles to growing innovation. These inefficiencies have provided a basis for spectrum management reform. [7] [8] Zimri details the key events towards spectrum reform in [7].

At present, the spectrum management policy in South Africa is for ICASA to assign and allocate spectrum bands statically (known as "command-and-control"), which results in spectral under-utilisation. In September 2010, ICASA embarked on a public process to review the existing radio regulations established under the Post Office Act and Radio Act of 1952, these regulations aimed to introduce a market-based spectrum management approach as opposed to a command-and-control mechanism, however these regulations were withdrawn in the final radio spectrum regulations. Spectrum trading in South Africa is seen to be an illegal process that the regulatory framework and policy does not cater for.

Due to the limited available documentation on spectrum trading in South Africa, the data collected is performed through document analysis of material that could be obtained through one-on-one interviews with decision makers of the regulator and policy maker which have been purposely selected for this research and may not be in the public domain. The interviewees were presented in advance with the questionnaire. The interviewees were selected, due to their expertise in the field of spectrum management and the organisation they represented as major spectrum holders.

The aim of the surveys were to determine the role-players reaction to spectrum trading, if it is viewed as desirable in South Africa as well as which strategy would best suit the South African market. It is worthwhile investigating which band should be used for spectrum trading as well as the market that this can be posed too.

## VI. RESULTS

The analysis and simulation results presented in this paper consist of two parts, the engineering approach and the regulatory approach. The engineering approach gives insights into two pricing strategies, market-equilibrium and competitive pricing, while the regulatory approach provides insight into the impact of the policy and regulations of South Africa. Two pricing strategies were considered as these were the most relevant in South Africa at the current time due to the maturity of the market.

The two pricing strategies were simulated in MATLAB using the same initial conditions (number of primary services is set to two, total frequency spectrum available to each primary service is 20 MHz, each primary service serves 10 primary users, target Bit-Error-Rate (BER) for the secondary users is $BER^{tar} = 10^{-4}$, channel quality for secondary users can vary between 9 and 22 dB and is initialised to 9 dB, spectrum substitutability is set to 0.7. The price offered by each primary service, $p_i$ and $p_j$, are both varied as per the scenarios below). The total revenue of both primary services achieved with market-equilibrium pricing and competitive pricing is shown in Figure VI-1Figure VI-2Figure VI-3 with varying rates of the bandwidth requirement for each of the primary users, also known as the efficiency of the pricing solutions.

The required bandwidth for the primary users was varied to show the effects of the bandwidth on the revenue. The bandwidth requirements of the primary service provider fluctuates as their requirement is not the same at all times, as the requirements of the primary service provider changes, either more or less spectrum is available to the secondary service provider.

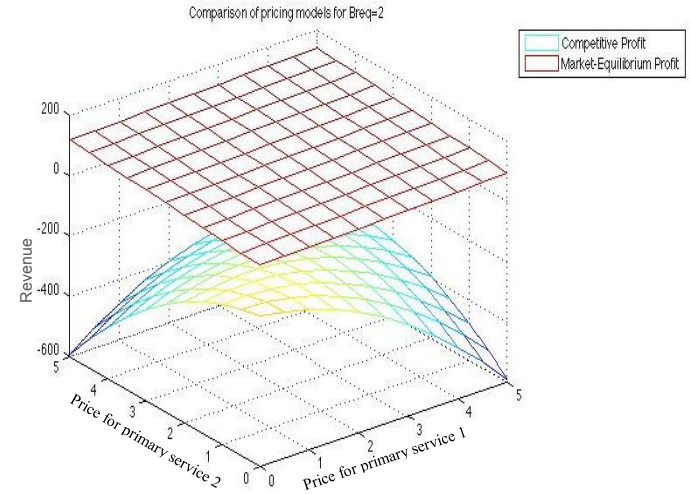



Figure VI-1: The total revenue for both pricing strategies with $B_i^{req}$=2

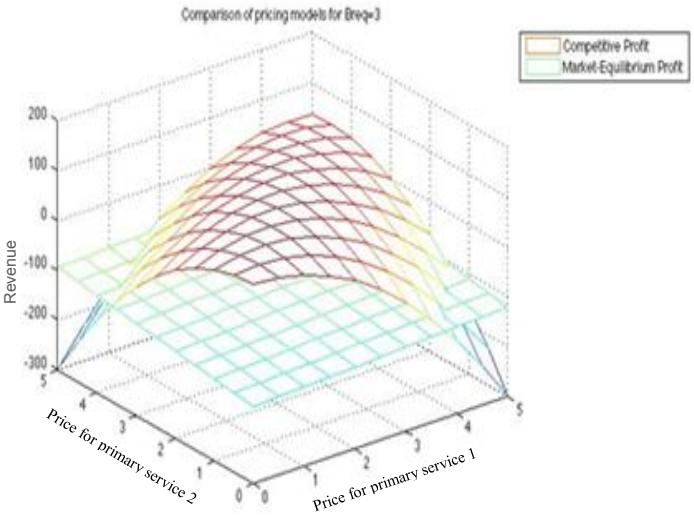



Figure VI-2: The total revenue for both pricing strategies with $B_i^{req}$=3

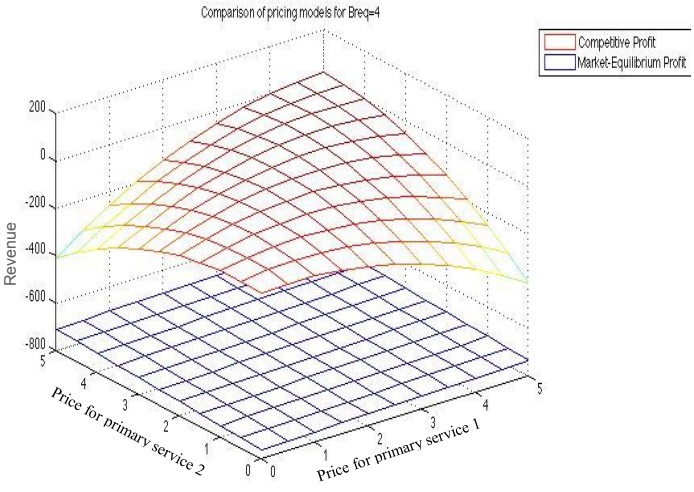



Figure VI-3: The total revenue for both pricing strategies with $B_i^{req}$=4

Table 1: Total revenue from each of the pricing strategies

| $B_i^{req}$ | Competitive pricing strategy (approximate max revenue) | Market-equilibrium pricing strategy (approximate max revenue) |
|---|---|---|
| 2 | -200 | 100 |
| 3 | 50 | -100 |
| 4 | 50 | -700 |

It can be seen from Table 1 that the higher the bandwidth requirement for each of the primary users, the higher the revenue earned from the competitive pricing strategy, alternatively, the revenue earned from market-equilibrium pricing strategy decreases as the bandwidth requirement for each of the primary users increases. The revenue earned from the market-equilibrium pricing strategy decreases as a result of the decrease in supply as the bandwidth requirement increases. The market equilibrium point is where the supply is equal to the demand; therefore the equilibrium point will be lower resulting in less revenue. For the competitive pricing strategy, as the bandwidth requirement of each of the primary user's increases, the cost discount decreases due to the improved QoS allowing the primary service provider to gain more revenue (this does not take the secondary users willingness to pay into account).

It is important to verify the existence of the pricing solutions as this shows the effect on the equilibrium point (price of the spectrum band) for varying values of the bandwidth required for each of the primary users. This is achieved by verifying if there is an equilibrium reached. FiguresFigure VI-4 and Figure VI-5 show the equilibrium points for a market-equilibrium pricing solution whereas Figure VI-6 and Figure VI-7 show the Nash equilibrium points for a competitive pricing solution.

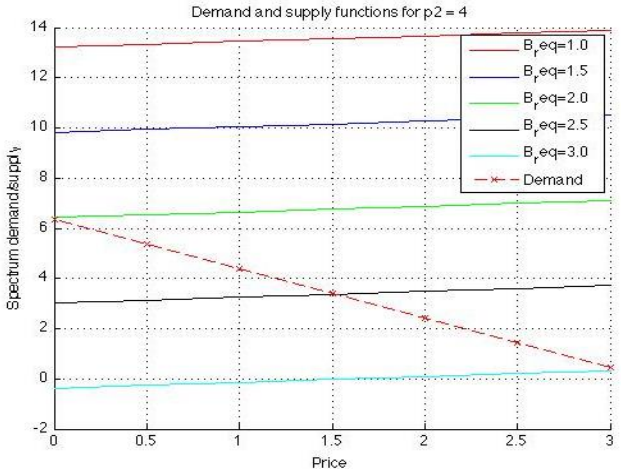



Figure VI-4: The spectrum supply and demand functions for a varying p1 and p2 set to 4

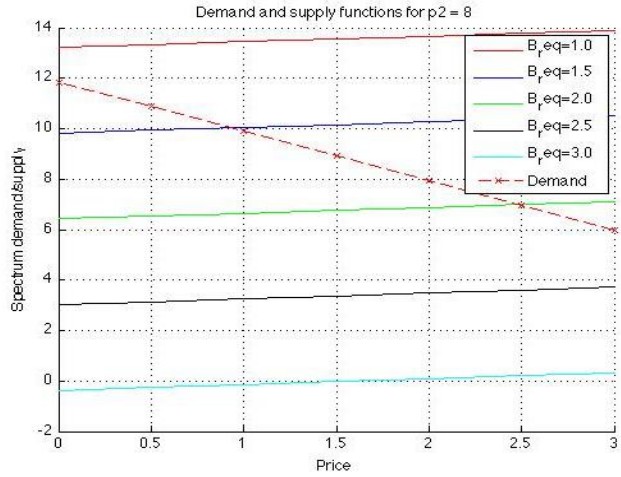



Figure VI-5: The spectrum supply and demand functions for a varying p1 and p2 set to 8

The figures above (Figure VI-4 and Figure VI-5) show that spectrum supply is dependent on the number of primary users and their bandwidth requirements. The figures below (Figure VI-6 and Figure VI-7) show that the existence of Nash equilibrium is dependent on the number of primary users and their bandwidth requirements. It is observed that market-equilibrium and Nash equilibrium exists only for certain values of offered prices and certain ranges of bandwidth requirement.

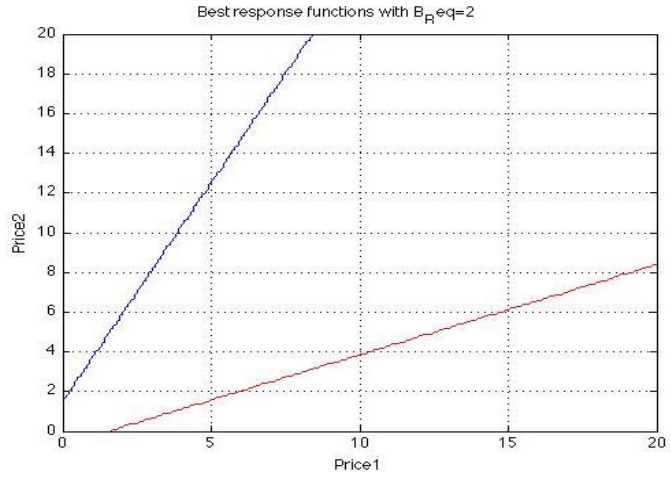



Figure VI-6: The best response functions for a fixed p1 and p2 respectively with $B_i^{req}$=2

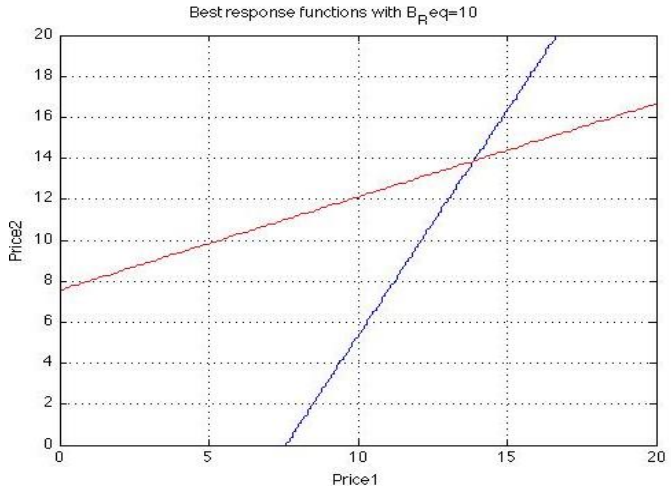



Figure VI-7: The best response functions for a fixed p1 and p2 respectively with $B_i^{req}$=10

For the competitive pricing solution, as the revenue of primary service two increases, the revenue of primary service one decreases, this is due to the effect of the offered price on the demand. If the price is too low, spectrum demand from the secondary service becomes high and the performance of the primary service degrades (causing a loss of revenue due to the cost discount), however as revenue from selling spectrum to secondary users is higher than the cost discount due to the performance degradation, the revenue of the service increases. If the price is too high, the demand for bandwidth becomes low and the revenue decreases. It can be seen that optimal values for the prices offered by primary service provider 1 and 2 are when they are the same.

By varying the offered price for primary service one and two, the effect on the revenue of primary service one can be determined. The maximum revenue attained is at the highest point of the curve in Figure VI-8. As the offered price for primary service one increases, the cost due to the QoS degradation to the primary user's increases, resulting in a negative profit for one primary service.

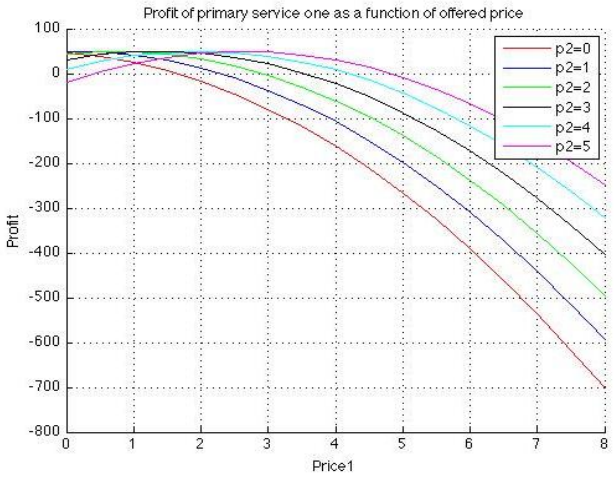



Figure VI-8: The revenue of primary service one for a varying p1 and p2

The bandwidth required is an important factor and affects the existence of a solution for both strategies as well as the revenue earned. Spectrum substitutability can influence the revenue as if it is too high, users can move around the frequency spectrum bands freely depending on the offered price and if it is too low, users would not want to choose that service. This highlights the importance of the variables in the pricing strategies and shows the impact they have on the revenue. From [4], it is shown that the cooperative pricing strategy has the highest revenue; however it is also the least stable for a distributed implementation (which is a more realistic model) and market-equilibrium pricing strategy has the lowest revenue of the three strategies, however the regulator feels a cooperative pricing strategy is the most desirable for a developing country. One of the downfalls of a competitive pricing strategy is that there is decrease in revenue when there are more primary services competing.

From the survey it is seen that the market would welcome sharing of underutilised whitespaces in spectrum with secondary users provided it does not cause interference; however operators do not agree for it to be given to secondary users for free as this would distort their economic model as they pay very high rates for the spectrum licenses. The suggested bands for this service are the freed up spectrum bands after the digital migration, 700 MHz and 800 MHz (TVWS) as well as the 2.6 GHz band. Therefore it can be said that the best band for spectrum sharing is the TVWS band due to the band being available as well as its desirable properties of radio signal propagation (long range, excellent in-building penetration and abundant bandwidth). However the regulator feels all the bands can be used for spectrum trading as all bands are underutilised.

The behaviour of the market if spectrum trading were allowed would be reflective on the market, as South Africa is not a mature market, the operators would act

competitively to outdo the other operators and earn the most revenue in line with the Competition Act. If operators cooperate with one another, jointly they can earn more revenue, however this is dependent on the market and the operators. It is believed that it is possible for the primary service providers to cooperate with the smaller role players; however the larger operators could attempt to eliminate the smaller role players from the market. The best pricing strategy currently suited to primary services in South Africa is the competitive pricing strategy where the primary services compete with each other to produce revenue.

The market considered for spectrum trading could include provision of backhaul links, rural sectors and point to multipoint sites for access networks. In rural areas, this comes at high costs and operators feel these services to such areas can be offered as a premium service with reduced rates. If spectrum trading is used for backhaul, there may be an issue of who the primary user is and who the secondary user is.

## VII. SOCIAL AND ECONOMIC IMPACTS OF SPECTRUM SHARING

"*Spectrum trading can provide significant economic and social benefits only if they become widely available and they are utilized"* [9] The benefits currently derived from Television Whitespaces (TVWS) that was made possible through Dynamic Spectrum Access (DSA) has enhanced socio-economic development of the end-user through the provision of high-speed Internet access to its citizens [10]. The TVWS has also enhanced radio spectrum availability without any national or regional re-structuring of the current international radio spectrum allocation policy. Likewise, the flexibility involved in DSA permits a dynamic spectrum market where licensed owners can lease out their unused radio spectrum to generate revenue, not only provides more income for the licensed owners, but also enhances radio spectrum availability and its utilization. Furthermore, the lower entry costs provided by DSA has contributed to both product and business model lifecycles by enhancing production of more communication equipment and services as well as promoting more job opportunities. This increase in worldwide production as well as the provision of more job opportunities has positively impacted a number of nations' GDP and worldwide economic growth in general.

## VIII. CONCLUSION

In this paper, solutions to the problem of secondary user pricing strategies in a cognitive radio environment have been approached from the perspective of game and economic theory. In terms of game theory, an attempt has been made to characterise the resolution of conflict among multiple cognitive radio users involved in selfish interaction. In terms of economic theory, an attempt has been made to characterise the conflict among primary services and the competition between them to maximise their revenue by selling spectrum to secondary users. Two strategies, the competitive pricing strategy and the market-equilibrium strategy, which have been derived from economic theory and game theory jointly, have been introduced and represented in sufficient theoretical details. The results from the simulations are discussed with relation to the regulatory and policy views in South Africa.

The simulation results presented indicate the competitive pricing strategy produces greater revenue than the market-equilibrium pricing strategy for certain bandwidths and the competitive pricing strategy describes the behaviour that primary service providers in South Africa will follow provided regulations are imposed to allow spectrum trading. Until ICASA promulgates regulation for implementing spectrum trading and revises the Radio Frequency Spectrum Regulations, spectrum trading is just an idea for the future.

## REFERENCES


[1] H Zhu and A Nel; "Investigating Opportunistic Spectrum Access from a Pricing Perspective"; Proceedings of the IEEE Africon Conference, Zambia, Sept 2011

[2] E Hossain, D Niyato, Z Han; "Dynamic Spectrum Access and Management in Cognitive Radio Networks"; Cambridge University Press, New York, July 2009

[3] A Ercan, J Lee, S Pollin and J Rabeay; "A Revenue Enhancing Stackelberg Game for Owners of Opportunistic Spectrum Access"; Proceedings of Dynamic Spectrum Access Networks (DySPAN 2008), Chicago, USA, October 14 – 17 2008

[4] D Niyato, E Hossain; "Market-Equilibrium, Competitive, and Cooperative Pricing for Spectrum Sharing in Cognitive Radio Networks: Analysis and Comparison"; Transaction on Wireless Communications, Vol. 7, No. 11, November 2008

[5] D Niyato; "Competitive Spectrum Sharing in Cognitive Radio Networks: A Dynamic Game Approach"; IEEE tranactions on wireless communications, Vol. 7, No. 7, July 2008

[6] D Niyato, E Hossain; "Competitive Pricing for Spectrum Sharing in Cognitive Radio Networks: Dynamic Game Inefficiency of Nash Equilibrium and Collusion"; Selected Areas in Communications, IEEE Journal, Vol. 26, No. 1, January 2008

[7] P Zimri; "Radio Frequency Spectrum, The Out of Sight, Out of Mind National Strategic Resource"; Still to be published, Johannesburg, 2013

[8] B Wellenius, I Neto; "Managing the Radio Spectrum: Framework for Reform in Developing Countries", Policy Research Working Paper, The World Bank, Global Information and Communication Technologies Department, Policy Division, March 2008

[9] GI Alptekin and AS Bener; "Spectrum trading in cognitive radio networks with strict transmission power control"; European Transactions on Telecommunications, Vol 22, 2011, published by Wiley Online Library

[10] J Popoola and R van Olst; "A Survey on Dynamic Spectrum Access via Cognitive Radio: taxonomy, requirements and benefits"; Journal paper prepared for submission to appropriate Telecommunications Policy Journal, February 2013.



**Elicia Naidu,** received her BSc Electrical Engineering (Information) degree in 2010 from the University of Witwatersrand. She is currently a part time candidate for the MSc 50/50 program at the University of Witwatersrand.

**Rex Van Olst**, is an Associate Professor with the University of the Witwatersrand, Johannesburg. He heads the telecommunication postgraduate research group at the School of Electrical and Information Engineering.

# Practical glycerol water solution measurements to determine the effects which the fluid properties has on the drop formulation process for 3D printers

PJM van Tonder, HCvZ Pienaar and DJ de Beer
Telkom Center of Excellence
Department: Electronic Engineering
Vaal University of Technology, Andries Potgieter Blvd, Vanderbijlpark, 1900
Tel: +27 83 3964440, Fax: +27 08 66526275
email: {malanvt}@vut.co.za

**Abstract - This paper describes the effect of different glycerol water mixtures on the solution's viscosity, surface tension and density. It also addresses the major disadvantage of a standard commercial 3 D inkjet printing process in terms of print material. A possible solution to this problem is also given.**

**Index Terms—3D printing, Glycerol water mixture, Viscosity, Surface tension, Density.**

## I. INTRODUCTION

The 3D printer technology is part of the additive manufacturing family originally developed by Massachusetts Institute of Technology (MIT) in 1993. The technology was licensed to several companies for the use of different applications. The Z-Corporation commercialised the technology to produce machines for printing of plaster and starch parts, while ProMetal Inc., Soligen Inc. and Therics commercialises the process respectively for metal, investment casting and pharmacology applications [1].

The 3D printing process works in the following manner: Firstly a model, normally in a .stl format, is sliced into layers by the printer software. Each of the layers is then sent to the printer in XYZ coordinates. After the printer received all of the printing data the printer spreads a layer of powder from the powder container onto the printing platform. A print is then made on the powder with a fluid binder using inkjet technology, where the printing head is enabled to move in the X-Y direction. The powder is "glued" together where the binder is printed. The remaining powder serves as support for other layers. When the printer has completed a layer, the printing platform moves down in the Z direction. Another layer of powder is spread onto the printing platform in order to print the next layer. Once the process is complete, the excess powder is vacuumed away and parts are lifted from the bed [2]. The 3D printing process is demonstrated in Figure 1.

However, one major disadvantage of the 3D printer technology, mentioned above, is that the printing material which can be used is limited to that of the supplier. For instance, when a prototype or product must be manufactured from materials like graphite or bone, the supplier would not be able to supply the user with that specific material. Graphite can be used to manufacture bipolar plates for Fuel Cells (FCs) or the bone can be used for medical implants.

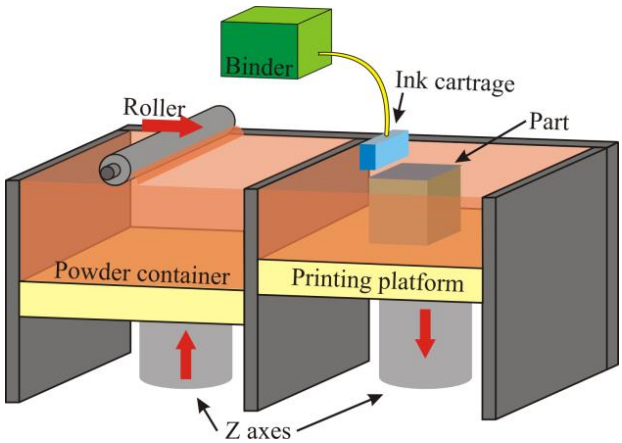


Figure 1. 3D printing process

## II. INKJET TECHNOLOGY

The fluid binder used in the Z 310 printer must have a low viscosity because of the HP C4800A inkjet technology used. The viscosity of the ink which HP uses in the inkjet technology is 2 mPa-s [3]. To match this ink viscosity, a water mixture is used [4]. Thus the real binder is in the powder and not the fluid. The water mixture is used to activate the binder in the powder to cause hardening [5]. This can be problematic when working with expensive powders and powders not supplied by the manufacturers (uncommon powders). When the binder is mixed with the powder it becomes almost impossible to extract that binder from the original powder, if it must be used for other applications. A possible solution to this problem is if the binder is the fluid and not the powder. This is not possible with current inkjet technology. When a glue-like fluid is used for the binder, it will increase the viscosity and will not extrude from the printing head. In order to determine the effects which the fluid properties have on the drop formulation process, different fluid solutions were prepared. These fluid solutions had different viscosities, densities and surface tensions as described later in the document.

## III. Fluid Solutions

When a solute substance dissolves in a solvent substance it is known as a solution. Solvents can be divided into the following classes:

- Polar solvents – Solvents which consist of strong dipolar modules having hydrogen bonding.
- Semi polar solvents – Solvents which consist of strong dipolar modules without hydrogen bonding
- Non polar – Molecules which consists of small or no dipolar characters

In order to obtain a solution the solute and solvent must be of the same class, in other words "like dissolves like". Solvents can fit into more than one of the classes, mentioned above. For example glycerol is considered to be polar and semi polar solvent, thus glycerol is an excellent solvent [6]. Glycerol solutions are widely used in experiments to determine the effect which the change in viscosity has in the fluid flow behavior. If glycerol is used as the solvent the solute must also be polar in order to obtain a solution. There are a number of polar solutes for example water, ethanol, methanol, benzene, acetone etc. [6]. Polar solvents and solutes use hydrogen bonds to form a solution. Hydrogen bonds consist of OH (Hydroxyl) groups to form bonds between a solute and a solvent. A hydroxyl group consists of an oxygen and hydrogen atom. As can be seen in Figure 2, an oxygen and hydrogen atom is considered to be neutral of charge.

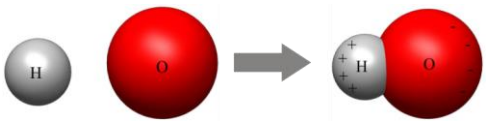



**Figure 2. Hydroxyl group**

An atom that is neutral of charge consists of an equal number of protons and electrons, thus balancing each other out. However when the oxygen and hydrogen atoms form a bond, a hydroxyl group is formed. Even though the group is still considered to be neutral of charge, the uneven distribution of the charges results in a partially positive and negative. The hydrogen in the hydroxyl group is partially positive where the oxygen is partially negative. Due to the partially charges formed, the hydroxyl group is considered to be polar. Hydroxyl groups will form weak bonds with each other because of the partially charges. The partially positive charged hydrogen will form a weak bond with a partially negative charged oxygen from a different hydroxyl group. However the bonds are constantly ripped apart, due to the fact that the molecules always move around [7].

## IV. Glycerol Solutions

Distilled water was chosen as solute due to its structure. However distilled water will be referred to as water in this document. Water mixed with glycerol will form a solution due to the fact that they are all polar elements with hydroxyl groups. Figures 3 and 4 show the structures of glycerol, and water, the hydroxyl groups are indicated by the red dashed lines. The viscosity, density and surface tension of the glycerol solutions were not calculated theoretically, due to the high error margin [8]. The solutions properties were rather determined practically using a viscometer and tension meter. The initial glycerol water solution was 15% glycerol and 85% water, where the % glycerol increased by 5% increments to a maximum of 80%. As the % glycerol increased the % water decreased by 5% decrements to a minimum of 20%.

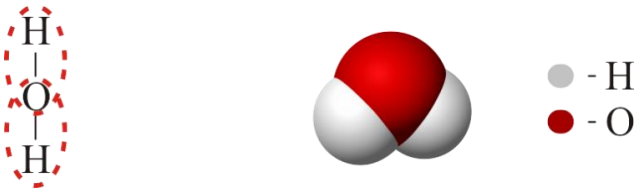



**Figure 3. Water molecule**

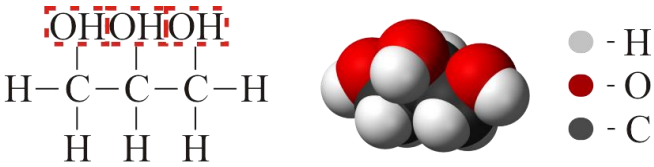



**Figure 4. Glycerol molecule**

## V. Viscosity and Density Measurements

The device used to determine the viscosity of a fluid is known as a viscometer. Many studies have been done to find methods to accurately determine the viscosity of fluids, the main measurement methods are the Rotational, Capillary, Vibratory and Ultrasonic methods [9]. The viscosity and density of the different glycerol mixture was determined using an Anton Paar SVM 3000 rotational viscometer due to its high accuracy. The SVM 3000, shown in Figure 5, uses the Stabinger and the oscillating U tube method to determine a fluid's density, kinematic and dynamic viscosity. The Stabinger measuring principle is demonstrated in Figure 6. The Stabinger measuring method consists of a copper housing, outer tube and a light weight rotor. Within the copper housing the outer tube, filled with the sample fluid, rotates at a constant speed.

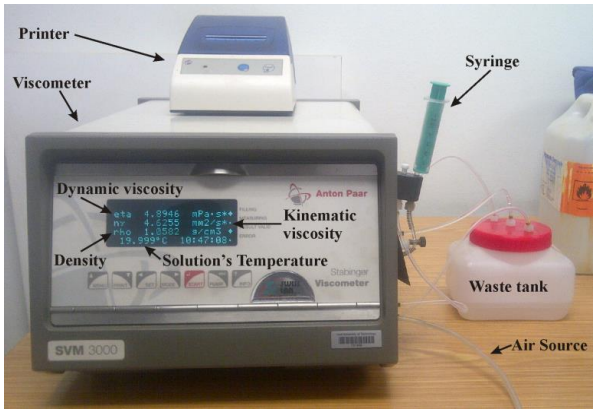



**Figure 5. SVM 3000**

A short description on the working of the SVM 3000 is given so that the principle is well understood. A light

weight rotor containing magnets floats in the sample fluid. The rotor will start to turn due to the shear force of the sample fluid, it will also be centered in the outer tube due to its low density and centrifugal forces. By using the speed difference between the rotating rotor and outer tube the dynamic viscosity can be calculated. The speed of the rotor is determined by means of a hall-effect sensor, thus there is no physical connection to the rotor [10].

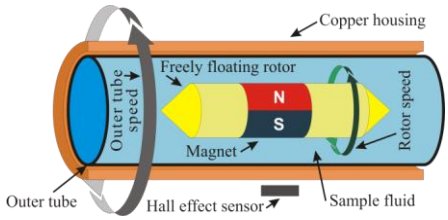



**Figure 6. Stabinger measuring principle**

The density of a fluid, or gas, can be determined by its resonance frequency, which can be obtained using the oscillating U tube method. The oscillating U tube method, as demonstrated in Figure 7, consists of a cantilever mounted U tube, mounting base, permanent magnets and inductor coils. The U tube is usually made out of glass or stainless steel, which is mounted onto a base. A pair of permanent magnets is bracketed onto the U tube, which extends through two inductor coils. Alternating Current (AC) is fed to one of the coils, which will cause a magnetic flied around the coil. The permanent magnet, which extends through the inductive coil, will start to swing due to the generated magnetic field. This will cause the U tube to vibrate at a certain frequency. The frequency can however be adjusted by changing the supply AC frequency.

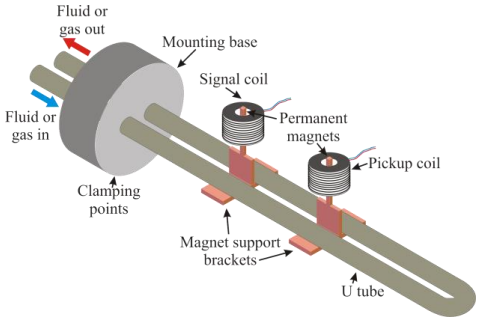



**Figure 7. U tube method**

An AC signal will be introduced in the second coil due to the swing of the permanent magnet through the second inductive coil. The AC signal introduced in the second inductor coil can be used to determine if the U tube is at its resonance oscillating frequency. The resonance frequency is inversely proportional to the square root of the summed up mass of the tube and its contents from the clamping points onwards. Thus the density of a fluid or gas can be determined from a one-time fill sample or a circulating sample [11].

However it must be kept in mind that the temperature will have an effect on the viscosity and density of a fluid, thus the temperature of the solutions must be kept constant through all of the experiments. The solution's temperature can be controlled by the viscometer. The temperature of the solutions was set to lab temperature which in this case was 20°C. Using a syringe a 5 ml fluid sample was taken from a solution. It must be ensured that there are no air bubbles present in the sample, due to the fact that this will cause an incorrect viscosity and density reading. The syringe was connected to the filling support connector, which is connected to the viscometer's density and viscosity measuring cells. By gently pushing the syringe plunger rod down the fluid sample enters the measuring cells. The measuring cells fluid outputs are connected to a waste tank, thus when the fluid sample enters the waste tank the measuring cells will be filled with the fluid sample. When the measuring cells were filled with the fluid sample the measuring cycle was started. It took 5 minutes to complete a measuring cycle, which determined the kinetic viscosity, dynamic viscosity and density of a fluid sample and would be displayed on the LCD screen. In order to prepare for the next fluid sample the measuring cells must be rinsed and dried to ensure the previous fluid sample would not affect the readings of the current sample. The necessary steps used to rinse and dry the measuring cells were:

1. The first step was to connect an air source to the filling support connector to push the solution out of the measuring cells and into the waste tank.
2. Acetone was then supplied to the measuring cells to dissolve all of the remaining solution.
3. The air source was then used to push the acetone-solution mixture into the waste tank and was also used to speed up the evaporation process of the remaining acetone- solution mixture. It took 1 min for the acetone-solution to evaporate. Once the density returned to 0 all of the acetone-solution evaporated.

### *A. Results*

The experimental results of the SVM 3000 can be seen in Table 1. It is clear that an increase in the % volume to volume mixture of glycerol and water (% v/v glycerol) caused an increase in the solution's viscosity and density.

**Table 1. Viscosity and Density**

| % v/v Glycerol | Solution volume (ml) | Viscosity (mPa.s) | Density (g/cm³) |
|---|---|---|---|
| 15 | 500 | 1.63 | 1.03 |
| 20 | 500 | 1.88 | 1.05 |
| 25 | 500 | 2.32 | 1.06 |
| 30 | 500 | 2.47 | 1.07 |
| 35 | 500 | 3.76 | 1.09 |
| 40 | 500 | 4.17 | 1.10 |
| 45 | 500 | 6.20 | 1.12 |
| 50 | 500 | 7.90 | 1.13 |
| 55 | 500 | 11.46 | 1.15 |
| 60 | 500 | 15.26 | 1.16 |
| 65 | 500 | 20.35 | 1.17 |
| 70 | 500 | 26.27 | 1.18 |
| 75 | 500 | 57.78 | 1.20 |
| 80 | 500 | 76.94 | 1.21 |

Using the data in Table 1 a graph % v/v glycerol vs. Viscosity and Density was plotted and can be seen in Figure 8. As seen in the graph the plot of the % v/v glycerol vs. density can be classified as a linear plot. A statistical analysis, using CurveExpert, was done on the plot to determine the mathematical relationship between the % glycerol water mixture and the solution's density. The mathematical equation is given below:

$$y(x) = 0.0028 \cdot x + 0.99$$
$$\therefore Density = 0.0028 \cdot (\%\ \text{v/v glycerol}) + 0.99 \quad (1)$$

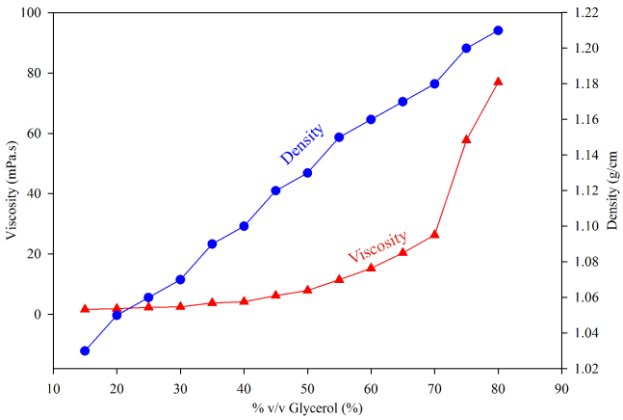



**Figure 8. % v/v Glycerol vs. Viscosity and Density**

However, as seen in Figure 8, the plot of the % v/v glycerol vs. viscosity can be classified as a nonlinear plot. A statistical analysis was also done on the % v/v glycerol vs. viscosity plot to determine the mathematical relationship between the % glycerol water mixture and the solution viscosity. The mathematical equation can be seen below:

$$y(x) = 0.53\left(1 + \frac{0.15 \cdot x}{-22.98}\right)^{-6.67}$$
$$\therefore Viscosity = 0.53\left(1 + \frac{0.15 \cdot (\%\ \text{v/v glycerol})}{-22.98}\right)^{-6.67} \quad (2)$$

## VI. Surface Tension Measurment

The surface tensions of the different fluid samples were determined by using an Optical Contact Angle (OCA) measuring device, manufactured by Dataphysics. The OCA 20 can be used to determine the following: fluid static contact angle, fluid dynamic contact angle, fluid surface tension, material surface energy and dispersion of polar contributions of surface free energy [12]. The main components of the OCA, demonstrated in Figure 9, are a video camera, light source, syringe (filled with the sample fluid), dosage needle and an electronical dispensing unit. In order to obtain the surface tension, of a fluid, the pendant drop-method is used.

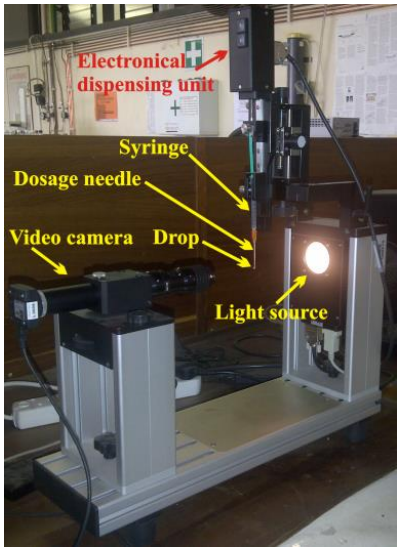



**Figure 9. OCA 20**

The pendant drop method will be explained so that the principle of the method can be understood. Using the pendant drop-method the syringe is filled with the sample fluid, however it must be ensured that there are no air bubbles present in the fluid due to the fact that this can cause and incorrect surface tension reading. A dosage needle with a specific size is then connected to the syringe's tip. The electonical dispensing unit will press down on the syringe's plunger rod, which will cause a drop to form out of the lower end of the dosing needle. The drop is monitored with the video camera, which is connected to the OCA Personal Computer (PC) software. The image of a drop, on the OCA software, can be seen in Figure 10.

The shape of the drop will depend on two forces namely: the gravitation force and the surface tension of the fluid. Thus the surface tension of a fluid can be mathematically determined by using the Young-Laplace equation on the drop shape. Two parameters must be defined in the OCA PC software to calculate the surface tension of a fluid namely: the fluid density and the dosage outer diameter size. The dosage needle's outer diameter is used as a reference to determine the actual size of the drop [12].

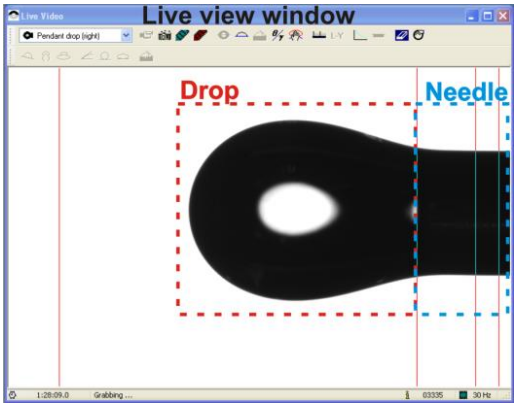



**Figure 10. OCA software**

The following steps were followed analyzing the fluid samples:

1. Using the syringe, 5 ml was taken from the fluid sample. A 0.165 mm dosage needle was then

attached to the syringe's tip. The syringe and dosage needle was then placed into the electronical dispensing unit.

2. The electronical dispensing unit was triggered via the OCA software to dispense a drop. The surface tension was calculated at the maximum drop size. If the drop size exceeds the maximum drop size it would detach from the dosage needle.
3. After the sample was analysed the syringe and dosage needle was disposed to prevent contamination of the fluid samples, which would affect the surface tension readings.

### *A. Results*

The experimental results of the OCA 20 tensio meter can be seen in Table 2. It can be clearly seen that an increase in the %v/v glycerol caused a decrease in the surface tension of the solution.

**Table 2. Surface tension**

| % v/v Glycerol | Solution volume (ml) | Surface tension (mN/m) |
|---|---|---|
| 15 | 500 | 70.84 |
| 20 | 500 | 70.52 |
| 25 | 500 | 70.15 |
| 30 | 500 | 69.34 |
| 35 | 500 | 68.67 |
| 40 | 500 | 68.4 |
| 45 | 500 | 68.3 |
| 50 | 500 | 67.85 |
| 55 | 500 | 67.4 |
| 60 | 500 | 66.8 |
| 65 | 500 | 66.5 |
| 70 | 500 | 66.08 |
| 75 | 500 | 65.44 |
| 80 | 500 | 65.12 |

Using the data in Table 2 a graph % v/v glycerol vs. Surface tension was plotted as can be seen in Figure 11. As seen in the graph the plot of the % v/v glycerol vs. surface tension can be classified as a linear plot.

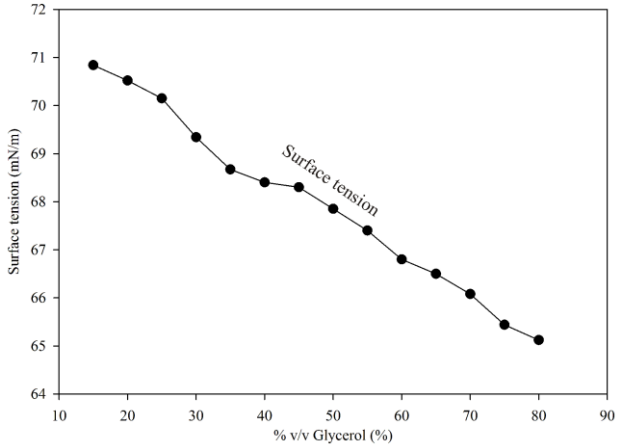



**Figure 11 % v/v glycerol vs. Surface tension**

A statistical analysis was also done on the % v/v glycerol vs. surface tension plot to determine the mathematical relationship between the % glycerol water mixture and the surface tension. The mathematical equation is given below:

$$y(x) = -0.088 \cdot x + 72.12$$

$$\therefore \textit{Surface tension} = -0.088 \cdot (\%\ \text{v/v glycerol}) + 72.12 \quad (3)$$

## VII. CONCLUSIONS

It was determined by changing the glycerol water mixture the viscosity, density and surface tension of the different solutions could be changed. The initial glycerol water solution was 15% glycerol and 85% water, where by the % glycerol increased by 5% increments to a maximum of 80%. As the % glycerol increased the % water decreased by 5% decrements to a minimum of 20%. The solution properties were not determined theoretically, due to the high error margin, but were determined practically. A SVM3000 viscometer was used to determine the viscosity and density of the solution and the OCA 20 tensio meter was used to determine the surface tension of the solution.

It was found that the viscosity of the glycerol water solutions increased as the % glycerol mixture increased. The increase in solution's viscosity was not directly proportional to the increase of the % glycerol mixture. A statistical analysis was done on the % v/v glycerol vs. viscosity plot, to determine the mathematical relationship between the % glycerol water mixture and the solution viscosity and can be seen below:

$$\therefore \textit{Viscosity} = 0.53\left(1 + \frac{0.15 \cdot (\%\ \text{v/v glycerol})}{-22.98}\right)^{-6.67} \quad (4)$$

It was further found that the density of the glycerol water solutions increased linearly as the % glycerol mixture increased. A statistical analysis was done on the % v/v glycerol vs. density plot to determine the mathematical relationship between the % glycerol water mixture and the solution density and can be seen below:

$$\therefore \textit{Density} = 0.0028 \cdot (\%\ \text{v/v glycerol}) + 0.99 \quad (5)$$

It was also observed that the surface tension of the glycerol water solutions decreased linearly as the % glycerol mixture increased. A statistical analysis was done on the % v/v glycerol vs. surface tension plot to determine the mathematical relationship between the % glycerol water mixture and the surface tension and can be seen below:

$$\therefore \textit{Surface tension} = -0.088 \cdot (\%\ \text{v/v glycerol}) + 72.12 \quad (6)$$

By using equation 4, 5 and 6 the viscosity, density and surface tension of a glycerol water solution can be calculated. Different glycerol water solution will have different fluid properties. Thus it will be used to determine the effect of which the viscosity, density and surface tension has on the drop formulation. The effects which the fluid properties has on the drop formulation process will be used to optimize the more "rugged" printing head, used in a 3D printer, to print with different fluid properties.

## VIII. REFERENCES


[1] E. RaminN, "*Automated design of trabecular structures*", England: Loughborough University**,** 2010.

[2] C.K. Chua, K.F. Leong, C.S. LIM, "*Rapid prototyping: Principles and Applications second edition",* USA: World Scientific Publishing Co. Pte. Ltd. , 2003, pp197.

[3] J. Wei, "*Product engineering Molecular structure and properties*". New York: Oxford University Press Inc., 2007.

[4] H. Butt, B. Kappl, " *Surface and Interfacial forces*", Germany: Wiley-VCH verlag GmbH & Co. KGaA Weinheim, 2010, pp164.

[5] O. Diegel, Verbal communication with Prof Olaf Diegel, Additive Manufacturing (3D Printing) expert, New Zealand: Massey University, 2012.

[6] D.B. Troy, "*Remington: The science and Practice of Pharmacy*", USA: Lippincott William & Wilkins, 2006.

[7] C. Masterjohn*," Cholesterol's Hydroxyl Group",* Availible at: < http://www.cholesterol-and-health.com/Cholesterol_Hydroxyl.html>, *Accessed: 15/05/2013, 200.*

[8] N.S. Cheng, "*Formulation for viscosity of glycerol-water mixture*". Industrial and Engineering Chemistry Research. 47:3285-3288, 2008.

[9] R. Kazys, & R. Rekuviene. "*Viscosity and density measurement methods for polymer melts*". Ultragarsas. 66(4):20-25. 2011.

[10] A. Paar, "*Measuring the Physical and Chemical Properties of Oil sample*", USA: Anton Paar, 2013.

[11] A.J. Hebra, "The Physics of Metrology", Germany: Springer, 2010.

[12] G. Maier, "*Operating manual OCA*", Germany: Dataphysics Instruments GmbH, 2002.


## IX. BIBLIOGRAPHY


[13] I. Gibson, D.W. Rosen, B. Stucker, "*Additive Manufacturing Technologies: Rapid Prototyping to Direct Digital Manufacturing*",USA: Springer New York Heidelberg Dordrecht London, 2010

[14] P. Glynne-Jones, M. Coletti, N.M. White, S.B. Gabriel, C. Bramanti, "*A feasibility study on using inkjet technology, micropumps, and MEMs as fuel injectors for bipropellant rocket engines*", Acta Astronautic. 67:194-203, 2010.

[15] H. Kipphan, "*Handbook of Print Media*", New York: Springer, 2004.

[16] C. Caglar, "*Studies of inkjet printing technology with the focus on electric materials*", Finland: Tampere University of Technology, 2009.

[17] K.C. Chaudhary, I.G Redekopp, T. Maxworthy, "*The Non-Linear Capillary Instability of a Liquid Jet*", Journal Fluid Mech. 96(2): 257-312, 1979.



**Malan van Tonder** received his M Tech degree in 2012 from the Vaal University of Technology and is presently studying towards his D Tech degree at the same institution. As a Telkom Center of Excellence student the research manly focuses on Additive Manufacturing.

# Experimental assessment of PV module cooling strategies

A. Ozemoya, A.J Swart and HCvZ Pienaar
Department of Electronic Engineering
Vaal University of Technology, Private Bag X021, Vanderbijlpark, 1900
Corresponding author email: austin4king@gmail.com

**Abstract – The main limiting factors to the extensive use of Photovoltaic (PV) modules include the high initial investment cost and the relatively low conversion efficiency. The issue of increasing the PV efficiency has been of great interest since the 1950's, both from a research and economic point of view. Temperature, however, exerts considerable influence on PV modules, with cell efficiencies decreasing as the cell's operating temperature increases. Higher surface temperatures mean lower output voltages and subsequent lower output power. This paper focuses on cooling techniques for controlling a PV module's surface temperature and the effect of different cooling techniques on the output voltage of the PV module and subsequently on the output power. Two cooling systems were investigated; a water cooling system and a forced air cooling system. A comparison was made between three PV modules, with water cooling, forced air cooling and without cooling. The results show a direct correlation between temperature rise and voltage decrease. It further reveals that water cooling is more effective than air cooling, with a water cooling system producing 4% more than a system with no cooling.**

**Keywords - Photovoltaic module, temperature degradation, tilt angle, ambient temperature, PV cooling.**

## I. INTRODUCTION

Solar power production using PV modules has increased and is currently one of the fastest growing energy technologies worldwide, leading to speculation that it will be the main source of electrical power in future [1]. Solar energy has been considered a promising solution to the global energy and environmental challenges facing mankind, including global warming [2]. Renewable energy sources contributed only about 13.1% of total primary energy supply in 2009; the share of solar PV was only 0.04% and is estimated to reach a maximum of 1% by 2030 [3]. Over the past five years, solar PV has averaged an annual growth rate of over 50%. Growth has generally been concentrated in a few countries, where solar PV currently generates only a low percentage of the total yearly electricity production [4].

This growth is driven not only by the progress in materials and processing technology, but also by market introduction programs in many countries around the world due to the increased volatility and mounting costs associated with fossil fuels [5]. However, owing to the recent economic slowdown and its associated developments, one might doubt the predicted magnitude or contribution of solar power to the global primary energy domain despite the proven potential of the sun in providing enough energy in one hour to meet the annual global consumption of energy. Achieving such ambitious targets demand reduction in the cost of solar energy per kilowatt hour (kWh) when compared to those of conventional fossil fuels. The current cost of production of PV modules has prevented it from being widely used, especially in the public sector where home residents could benefit greatly from its use. This high cost can be attributed to the fact that presently PV modules are made of expensive semiconductor materials, which are most commonly crystalline silicon (c-Si) [5].

Moreover, low efficiencies of PV modules still exist. A typical PV module converts 6-20% of the incident solar radiation into electrical energy, depending upon the type of PV module and climatic conditions [6]. This means that a larger surface area PV cell is required to produce the same amount of electrical energy, which could otherwise be produced by a smaller and more efficient PV cell. Therefore, to address the high manufacturing and installation costs associated with PV modules, two different approaches have been adopted, namely:

- Increase the efficiency of PV modules (PV efficiency).
- Use cheaper materials for the construction of PV modules.

The adverse effect of temperature increase on the performance of a PV module is a significant factor to take into consideration. A PV cell suffers from high temperatures reached under high irradiation conditions and can reach as high as 60-80°C [7]. An effective way of improving efficiency and reducing the rate of thermal degradation of a PV module is by reducing the operating temperature of its surface [8]. This can be achieved by cooling the module and reducing the heat stored inside the PV module during operation. Decreasing the temperature of PV module can boost the electrical efficiency [9].

The purpose of this paper is to optimize the available output power from a PV module by comparing the electrical performances of PV modules with and without cooling techniques. Literature relating to factors affecting PV module temperature is also discussed. The research methodology is then introduced and the practical set-up is explained. Initial results are presented in a number of graphs. Future work and conclusions are finally presented.

## II. FACTORS THAT CONTRIBUTE TO EFFICIENCY LOSSES IN PV MODULES

Factors that influence the operating condition of a PV cell are the total irradiance, the spectral distribution of the irradiance and the temperature [10]. Conditions in real life situations are not standard; instead, they vary strongly and influence the electrical performance of a PV cell, causing an efficiency loss with respect to the standard testing condition

(STC) nominal value [10]. This loss can be divided into different categories [11], namely:

**Angular distribution of light:** Due to the movement of the sun and the diffuse components of the radiation, light does not fall perpendicular onto the PV module's surface, as is the case when measurements are done and the nominal efficiency is determined.

**Spectral content of light:** For the same power content, different spectra produce different cell currents according to the spectral response. The solar spectrum also varies with the sun's position, weather, pollution, etc. and never exactly matches the AM1.5 standard.

**Irradiance level:** For a constant cell temperature, the efficiency of the module decreases with diminished irradiance levels. This is primarily due to the logarithmic dependence of open-circuit voltage on photocurrent; at very low illumination the efficiency loss is faster and less predictable.

**Ambient temperature:** The surface temperature of PV modules rises with longer exposure periods to sunlight and high ambient temperature. The elevated temperatures directly impact the PV modules efficiency. The ambient temperature changes because of the thermal insulation provided by the encapsulation. This is usually the most important performance loss. However, prediction of the module response under different conditions is required to correctly assess the yearly production of a PV system in the field [11].

**Surface orientation of PV modules:** The radiation falling on a tilted surface will be the sum of direct radiation, diffuse radiation and reflected radiation or albedo [12]. The sum of these three components is called global radiation. The correct installation of a collector can enhance its application advantage, as the amount of radiation flux incident upon the collector is mainly affected by the surface orientation i.e. azimuth and tilt angles of installation. In the northern hemisphere, the best azimuth is due south (facing equator), but the tilt angle varies with factors such as the geographic latitude, climate condition, utilization period of time, etc. [13].

**Thermal effect:** Thermal response of the PV module affects the electrical power output. The PV module receives the incident irradiation where a portion of it is converted to electrical energy in proportion to the module's efficiency. The rest of the incident irradiation heats up the PV module and increases its operating temperature in relation to the PV material's specific heat capacity [14]. PV module voltage is reduced when compared to the increase of current at higher operating temperatures, so the generated power is reduced. A portion of the absorbed heat is dissipated into the surroundings, occurring through conduction, convection and radiation [14].

## III. PROPOSED SYSTEM

The test was designed to investigate the electrical performance of a PV module with different cooling techniques. The practical set-up consists of a SW220 poly-crystalline PV module by SOLARWORLD with its peak efficiency of 13.12% under STC (25°C, 1000 W/m$^2$). The system was installed on the roof of the S-Block at the Vaal University of Technology (VUT), with latitude 26°S and longitude 27°E . Figure 1 and 2 presents the block diagram of the practical set-up and experimental set-up inside the laboratory.

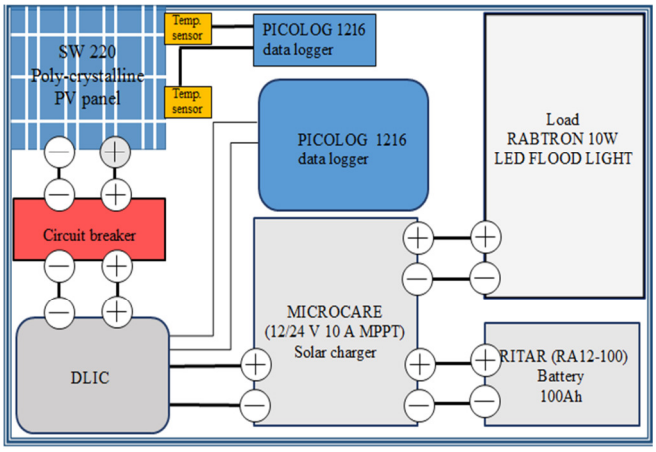



Figure 1: Block diagram of the practical set-up

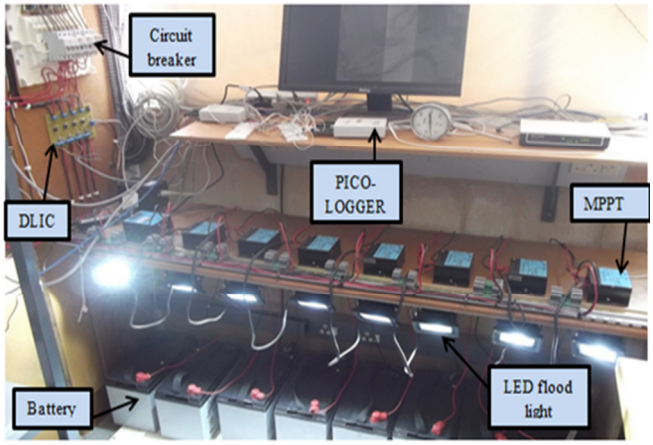



Figure 2: Experimental set-up of the monitoring station

The block diagram of the practical set-up comprises of a PV module connected to a data logging interface circuit (DLIC) via a circuit breaker. The data acquisition equipment consists of a DLIC and a PICOLOG 1216 data logger. The 12/24 V solar charger with maximum power point tracker (MPPT) was connected via several channels of the DLIC. A PICOLOG 1216 data logger was connected on each corresponding channel of the DLIC where it was used to record PV module output voltages and currents. A PICOLOG 1216 data logger was used as it has 16 analogue input channels that can accommodate all PV modules used in the set-up. The loads i.e., LED FLOOD LIGHT was coupled to the MPPT along with (RA12-100) lead acid discharge batteries (LADDB). Three identical PV systems are considered for this research. The first test was to investigate the factors that impact the PV module surface temperature as the transformation of solar energy into electrical energy depends on the operating temperature of the module [15]. Therefore, the surface temperature at different points was measured jointly with the air temperature. A pilot study was undertaken to investigate which tilt angles produce maximum PV module surface temperature and how it affects the output power. Results

show that a PV module's surface temperature is highest at a tilt angle of 16° during the day and lowest at night time [15]. This tilt angle was therefore chosen in the main study with the intention of maximizing the effect of the cooling system. It further reveals that the front and back-surface module temperature can be distinctly different [15]. The back-surface temperature is a good approximation of the actual cell temperature, and studies have shown that cells typically run 3°C warmer than back-surface temperatures for glass–glass laminate construction [16]. The electrical performance and reliability of PV modules can be severely affected by elevating cell operating temperatures due to elevated ambient temperature.

The study features three identical SW 220 poly-crystalline PV modules set to the same tilt angle of 16° with an orientation angle of 0° [15], connected to a data logging interface circuit (DLIC) via a circuit breaker (CB). The set-up has been developed to study the effect of different cooling techniques on the output voltage of the PV module and subsequently on the output power. Table 1 indicates selected parameters for a PV module which was used in this research due to its lower cost and better performance in areas of direct solar radiation [17].

Table 1: Electrical characteristic of SOLAR WORLD SW220 poly-crystalline PV module

| Specification | Abbreviation | Value |
|---|---|---|
| Maximum output | $P_{max}$ | 220 W |
| Open circuit voltage | $V_{oc}$ | 36.6 V |
| Rated voltage | $V_m$ | 29.2 V |
| Short circuit current | $I_{sc}$ | 8.08 A |
| Rated current | $I_m$ | 7.54 A |
| Nominal operating cell temperature | $NOCT$ | 46°C |
| Efficiency | ŋ | 13.12% |

A water cooling system is used with the first PV system, where water is sprayed onto the PV module's front surface at specific time intervals. A forced air cooling system is used with the second PV system, where fans are used to blow air onto the back of the PV module. The third system has no cooling system employed as it serves as a reference to make a reasonable comparison to the other two cooling systems. The aim of using water or forced air is to keep the PV modules operating temperatures within limits so as to achieve higher cell efficiencies, since generation of heat within the PV cell accounts for the increase in cell temperature and the decrease in its conversion efficiency [9]. The surface temperature of the PV module, the output voltage and current are monitored using a data logger and all measurement is saved to an MS Excel file for further analysis to draw reasonable conclusions

**Water cooling setup**

Water is used as the cooling medium in the first PV system. The cooling circulation system consists of a brushless DC pump (12 VDC 10.5 W), a 12 V DC solenoid valve, sprayers, a garden hose and a 50 litre water tank. The pump is powered from a 12 V/ 1.4 A power supply. Its maximum head being 4 m with a maximum flow rate of 450 l/h. The water pump circulates the water through a spray system connected by a hose located between the aluminium frames that separate two PV modules (see Figure 3).

Figure 3: Water cooling system. (Outdoor setup 1 - PV module, 2 - pump, 3 - solenoid valve, 4 - sprayers, 5 - hose, 6 - water tank covered with cardboard box.

The pump and solenoid valve are controlled by an electronic timer circuit giving a duty cycle cooling period of 10 seconds for every 5 minute intervals. Initial trials for spraying were 20 seconds every 10 minutes. However, this spray interval was not effectively keeping the temperature from decreasing, as the 20 second cooling effects drops off significantly after 5 minutes. Around 20 litres of water is used per day when the sprayers are activated for 10 seconds every 5 minutes. The electronic timer circuit runs from 09:00 am till 3:00 pm, this time period was considered because the PV module's surface temperature climbs considerably as revealed in the pilot study [15].

To investigate the performance of the PV cell under different cooling interventions, the system was operated for 15 weeks from December 2013 to March 2014, as this is the hottest period of the year in Vanderbijlpark [18]. This design is employed to minimize the consumption of water which is crucial to the objective of the project. The water tank is enclosed by a cardboard box to avoid heating by solar irradiation. Having the precise pump selection is important in making this system viable. The right pump needs to be:

- Low powered;
- Sufficient flow rate (at the head pressure); and
- Reliable

**Forced Air cooling setup**

The second cooling technique involves attaching eight DC brushless fans (SUNON 12 VDC 10 W, 120·120·38 mm) to the back of a PV module to improve the electrical performance of the module. Figure 4 shows the back of a PV module with the fans attached. The brushless fans are powered from a 12 VDC source and are connected in

parallel. The fans disperse the heat generated by the PV module to the surrounding areas. The experiments were conducted from 9:00 am until 3:00 pm. The fans are controlled by an electronic timer circuit giving a duty cycle cooling period of four minutes on and one minute off. Initial trial runs of air cooling set the interval to be 20 seconds for every 10 minutes. However, the results showed no significant temperature drop as compared to the PV module without cooling (reference system). Figure 4 shows the forced air cooling setup and the angle changing mechanism.

Figure 4: The back of a PV module with attached cooling fans

## IV. EXPERIMENTAL RESULTS

The obtained results are presented to check the performance of the PV module under different cooling interventions. Figure 5 presents the relationship between the PV module surface temperature and its output voltage with the different cooling interventions applied. The difference between the PV module's surface temperature and output voltage for a particular day is plotted.

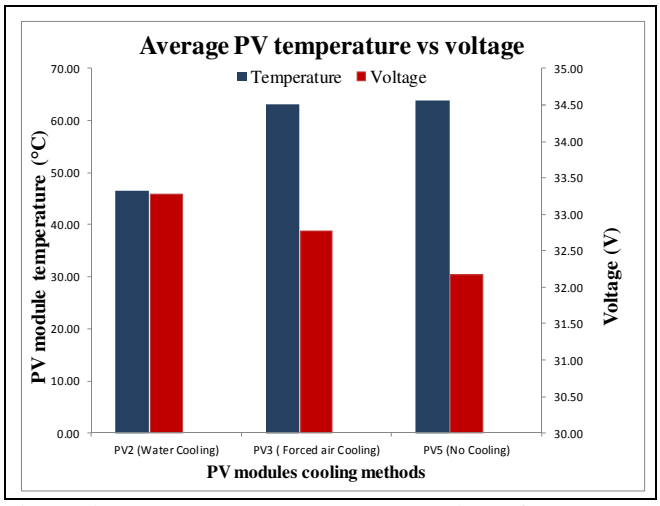



Figure 5: Average PV temperature vs voltage for 31st Dec 2013.

It shows that the output voltage of a PV module with cooling is slightly higher than the PV module without cooling. The results shows a direct correlation is observed between temperatures rise and voltage decrease. It also shows the advantage of using the water as a coolant since it can sustain a lower PV module surface temperature, providing better temperature uniformity along the PV module surface than forced air. Table 2 shows the statistical data for a period of one full week, which indicates that there is a statistical significant relationship (p-value < 0.001) between the temperature rise and output voltage decrease (shown by the negative correlation value).

Table 2: Average, mode, median, correlation and probability-value (p-value) statistics for temperature and voltage for different techniques

| **Week 5 Data** | **December 29th - January 4th** | | | | | |
|---|---|---|---|---|---|---|
| | PV2 | | PV3 | | PV5 | |
| | Water cooling | | Forced Air cooling | | No cooling | |
| | Temp. (°C) | Vol. (V) | Temp. (°C) | Vol. (V) | Temp. (°C) | Vol. (V) |
| Average | 47.57 | 32.91 | 63.38 | 32.43 | 63.91 | 31.73 |
| Mode | 46.13 | 33.08 | 63.32 | 32.55 | 61.00 | 31.90 |
| Median | 48.08 | 32.98 | 64.43 | 32.50 | 64.44 | 31.83 |
| Correlation | -0.44 | | -0.40 | | -0.50 | |
| P-value | 0.00 | | 0.00 | | 0.00 | |

Result also shows that with forced air cooling, there is a substantial temperature rise along the PV module due to the low heat capacity of air. The thermal properties of air make it less effectual as a coolant medium than water [19]. More power fans will therefore be needed to achieve the same cooling performance. The water cooling system resulted in an average higher output voltage of 1.18 V and an average temperature reduction of 16.34°C (or 25.57%) when compared to a no cooling system. The water cooling system was able to maintain an average PV module surface temperature of below 50°C.

Table 3 shows the results for the entire 15 week period, indicating that the water cooling system resulted in an average higher output voltage of 780 mV.

Table 3: Average voltage, correlation and p-value for 15weeks

| **Data** | **15 weeks period** | | | | | |
|---|---|---|---|---|---|---|
| | PV2 | | PV3 | | PV5 | |
| | Water cooling | | Forced Air cooling | | No cooling | |
| | Temp. (°C) | Vol. (V) | Temp. (°C) | Vol. (V) | Temp. (°C) | Vol. (V) |
| Average | 39.22 | 31.30 | 48.39 | 31.16 | 49.10 | 30.59 |
| Correlation | 0.11 | | 0.09 | | 0.07 | |
| P-value | 0.00 | | 0.00 | | 0.00 | |

The effect of ambient temperature on the output power of a PV module can be clearly discerned in Figure 6. It shows the different combinations of temperature and power that can be produced by a given PV module under different cooling conditions. The operating temperature plays a central role in the PV conversion process. Both the electrical efficiency and hence, the power output of a PV module depend linearly on the operating temperature. The results indicate a higher current value of 65.30% (water cooling system) when compared to the reference system current (no

cooling system) which results in a 49.6% power difference between these two systems.

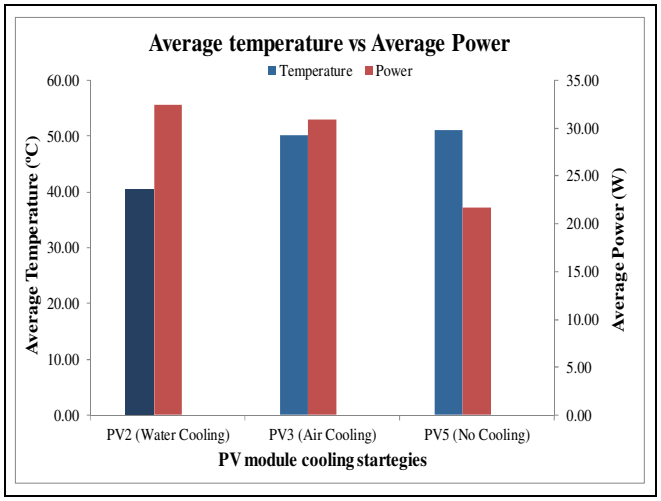



Figure 6: Average PV temperature vs power for a period of 15 weeks

## V. CONCLUSION

The purpose of this paper was to optimize the available output power from a PV module by comparing the electrical performances of PV modules with and without cooling techniques. The variation of temperature between these cooling and no cooling techniques can be as high as 16ºC. Results indicate a direct correlation between PV surface temperature rise and voltage decrease. It further reveals that water cooling is more effect than air cooling, with a water cooling system producing 1.18 V more than a system with no cooling in the last week of December 2013. Weekly period percentages indicate a higher voltage value of 3.58% and 2.16% (water and forced air cooling respectively) when compared to a no cooling technique. However, water cooling of a PV module did reveal significant decolouration in the protective tempered glass layer. Future work exists to investigate ionized or deionized (distilled) water as a better option.

## V. REFERENCES


[1] D. Redfield, "Solar energy: Its status and prospects," *CSIT Newsletter, IEEE,* vol. 4, pp. 15-19, 1976.

[2] Y. Dajiang and Y. Huiming, "Energy Conversion Efficiency of a Novel Hybrid Solar System for Photovoltaic, Thermoelectric, and Heat Utilization," *Energy Conversion, IEEE Transactions on,* vol. 26, pp. 662-670, 2011.

[3] M. Hoffman, "PV solar electricity industry: market growth and perspective.," *Solar Energy Mater. Solar Cells,* vol. 90, pp. 3285-3311, 2006.

[4] IEA. [Online]. Available: http://www.iea.org/aboutus/faqs/renewableenergy/

[5] A. Jäger-Waldau, "1.09 - Overview of the Global PV Industry," in *Comprehensive Renewable Energy*, S. Editor-in-Chief: Ali, Ed., ed Oxford: Elsevier, 2012, pp. 161-177.

[6] S. Dubey, J. N. Sarvaiya, and B. Seshadri, "Temperature Dependent Photovoltaic (PV) Efficiency and Its Effect on PV Production in the World – A Review," *Energy Procedia,* vol. 33, pp. 311-321, 2013.

[7] Z. Farhana, Y. M. Irwan, R. M. N. Azimmi, A. R. N. Razliana, and N. Gomesh, "Experimental investigation of photovoltaic modules cooling system," in *Computers & Informatics (ISCI), 2012 IEEE Symposium on*, 2012, pp. 165-169.

[8] S. Odeh and M. Behnia, "Improving Photovoltaic Module Efficiency Using Water Cooling," *Heat Transfer Engineering,* vol. 30, pp. 499-505, 2009.

[9] H. G. Teo, P. S. Lee, and M. N. A. Hawlader, "An active cooling system for photovoltaic modules," *Applied Energy,* vol. 90, pp. 309-315, 2012.

[10] M. Chegaar and P. Mialhe, "Effect of atmospheric parameters on the silicon solar cells performance," *Journal of Electron Devices,* vol. 6, pp. 173-176, 2008.

[11] A. Luque and S. Hededus, *Handbook of Photovoltaic Science and Engineering*. West Sussex: John Wiley & Sons Ltd, 2003.

[12] C. S. Solanki, *"Solar Photovoltaics: Fundamentals, technology and applications"*, 2nd ed. New Delhi: PHI learning private Ltd., 2011.

[13] T. P. Chang, "Study on the Optimal Tilt Angle of Solar Collector According to Different Radiation Types," *International Journal of Applied Science and Engineering,* vol. 6, pp. 151-161, 2008.

[14] S. Krauter and R. Hanitsch, "Actual optical and thermal performance of PV-modules," *Solar Energy Materials and Solar Cells,* vol. 41–42, pp. 557-574, 1996.

[15] A. Ozemoya, A. J. Swart, C. Pienaar, and R. M. Schoeman, "Factors impacting on the surface temperature of a PV panel," presented at the Southern Africa Telecommunication Networks and Applications Conference (SATNAC) 2013 16th International Conference, Stellenbosch, 2013.

[16] D. L. King, W. E. Boyson, and J. A. Kratochvil, "Photovoltaic array performance model," Sandia National Laboratories, Albuquerque, New Mexico2004.

[17] O. Asowata, J. Swart, and C. Pienaar, "Optimum Tilt and Orientation Angles for Photovoltaic Panels in the Vaal Triangle," in *Power and Energy Engineering Conference (APPEEC), 2012 Asia-Pacific*, 2012, pp. 1-5.

[18] A. J. Swart, R. M. Schoeman, and H. C. Pienaar, "Assessing the effect of variable atmospheric conditions on the performance of photovoltaic panels: A case study from the Vaal Triangle," in *Energy Effciency Convention (SAEEC), 2011 Southern African*, 2011, pp. 1-6.



Ozemoya Augustine received his undergraduate degree in 2011 from Vaal of Technology University, South Africa. And he is presently studying towards his Master degree in electrical engineering at the Vaal University of technology. His research interests include how to improve the efficiency of photovoltaic modules by controlling the ambient temperature.

# Quantifying the effect of varying percentages of full uniform shading on the output power of a PV module in a controlled environment

Arthur J Swart and Pierre E Hertzog
Department of Electrical, Electronics and Computer Engineering
Central University of Technology, Private BagX20539, Bloemfontein, 9300
Tel: +27 51 507 3907, Fax: +27 86 607 1786
Email: drjamesswart@gmail.com

**Abstract-Partial shading of a PV module has received much attention over the past few years as it results in uneven cell power generation, compromising total power production. Research relating to the exact effect that varying percentages of full uniform shading has on the output power of a PV module is somewhat lacking. The purpose of this paper is to quantify the percentage of full uniform shading of a given PV module within a controlled environment, correlating it to the output power of the module. The percentage of full uniform shading equates to the percentage of light intensity of the direct beam component which is not proportional to the output power of the PV module (e.g. a 35% full uniform shading net allows 65% of light to pass which results in a 50% reduction of PV output power).**

**Index Terms—full uniform shading, controlled-environment, efficiency, alternative energy**

## I. INTRODUCTION

"Almost every way we make electricity today, except for the emerging renewables and nuclear, puts out Carbon Dioxide. And so, what we're going to have to do at a global scale is create a new system. And so, we need energy miracles" [1] . Energy miracles, as referred to by Bill Gates, require on-going research into developing and understanding new energy systems. Renewable energy systems, such as photovoltaic (PV) systems, still require much research and development in order to improve efficiency and reduce overall manufacturing costs. In fact literature states that an ever-increasing need to improve the efficiency of energy production still exists today [2], especially in view of its ever increasing global capacity (see Figure 1).

The efficiency of PV cells varies from 3% to 31%, depending on the technology, light spectrum, atmospheric conditions, temperature, design and material used [3]. Interrupting direct beam radiation lowers the output voltage of a PV module significantly, influencing the amount of output power available for driving an alternative energy system, which may lead to system downtime or even component failure over a period of time [4]. This interruption is usually due to cloud movement or shading of the PV module by natural or man-made causes. Partial shading of a PV module has received much attention over the past few years as it results in uneven cell power generation, compromising total power production [5]. A Google Scholar Search of the terms "partial shading" and "solar panel" revealed some 685 hits, while the words "full shading" and "solar panels" revealed only 17 hits.

Research relating to the exact effect that varying percentages of full uniform shading exert on the output power of a PV module is somewhat lacking. For example, Giaffreda et al. [6] contrasted full uniform shading to partial shading of a PV cell and proved that its cell temperatures increase dramatically when shaded. However, no percentages of varying shade were reported, nor the effect on the output power.

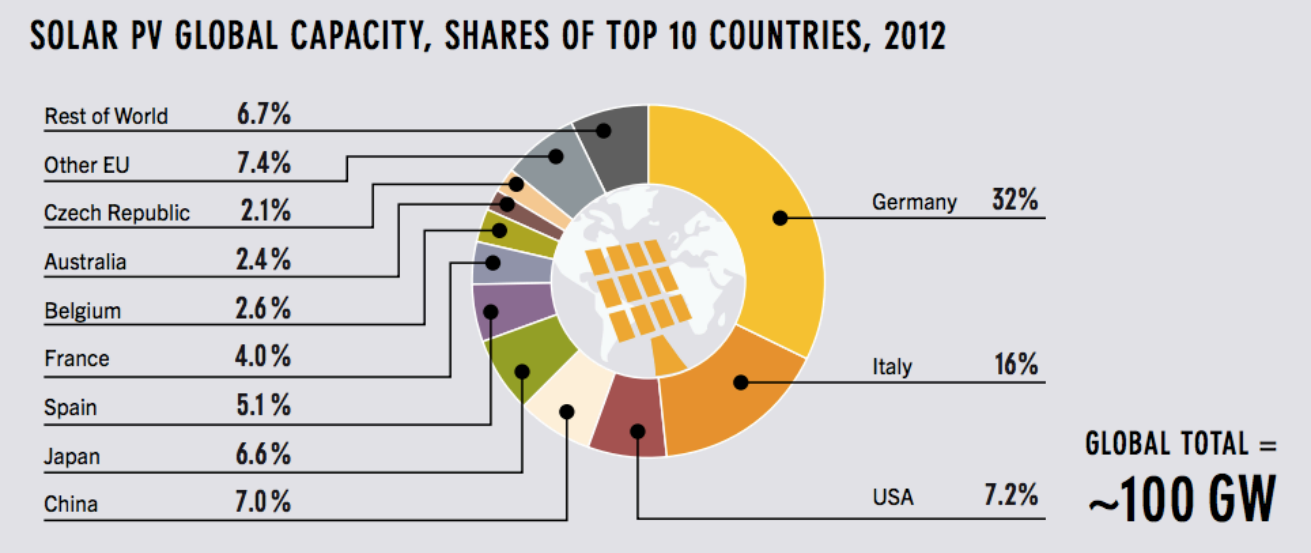



Figure 1: PV Global Capacity in 2012 [7]

Results given by Christy regarding shading suggested that the reduction in current is not proportional to the amount of shading on the PV panel [8]. Again, no varying percentages of full uniform shading were reported on. Gummeson et al. [9] reported that a fully-shaded 11.4 $cm^2$ PV module produced 29 times less power than the same PV module under bright indoor lighting conditions. Again, no percentages of full shade were mentioned, although a significant power reduction was given. Johnson [10] used Blue Hawk 4 mm thick, heavy-duty plastic sheeting to cover the top of PV modules to provide varying levels of insolation, which included un-shaded, partial shaded and full shaded modules. Power reduction was found to be around 33% for full uniform shading. Again, no varying percentages of shading were used.

The purpose of this paper is to quantify the percentage of full uniform shading of a given PV module within a controlled environment, correlating it to the output power of the module. The percentage of full uniform shading equates to the percentage of light intensity of the direct beam component. Literature pertaining to the importance of direct beam radiation will be covered, with varying percentages of full uniform shading being equated to diffuse radiation. The methodology and practical setup will then follow. The results feature a number of graphs and photographs followed by succinct conclusions.

## II. Direct and Diffused Beam Radiation

PV modules receive direct (beam), diffused and reflected ground radiation during varying atmospheric conditions [11]. Direct radiation is the part which travels unimpeded through space and the atmosphere to the surface of the earth, while diffused radiation is the part scattered by atmospheric constituents such as molecules, aerosols and clouds [12]. Figure 2 illustrates direct, diffused and reflected ground radiation received by a specific PV module used in research conducted at the Vaal University of Technology (VUT) [13].

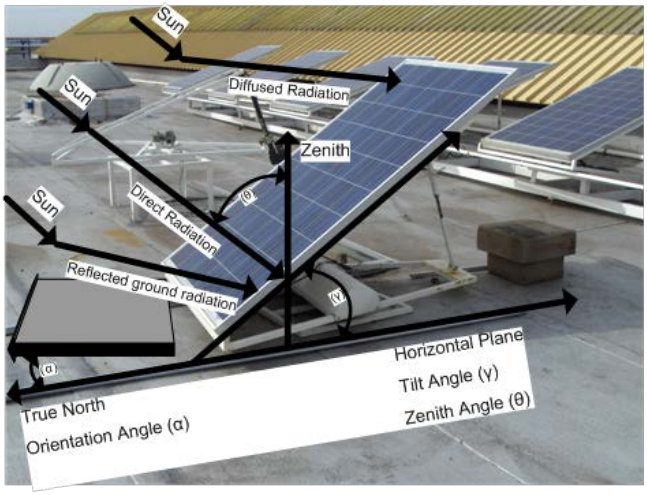



Figure 2: Different radiation beams [13]

Wenham et al. [14] reported that on a cloudy day, all the incoming radiation is assumed to be diffused, with intensity approximately equal to 20% of the direct beam component Cloudy conditions, as well as air pollution, therefore inhibit direct radiation, giving rise to diffuse radiation which is not conducive to optimum PV performance. Diffuse radiation on a PV module could take many forms, including shading from:

- a tree in summer (all leaves present);
- a tree in winter (no leaves present);
- thin clouds; and
- thick clouds

These forms of diffused radiation do not all exhibit the same percentage of shading. For example, evergreen trees (trees that have leaves throughout the year) provide a higher percentage of shading than deciduous trees (trees that shed their leaves in winter) [15]. Furthermore, research has shown that appropriate positioning of large trees near buildings could save approximately 4.7% in cooling demand and 3.3% on electricity usage [16]. These varying type and size of trees all exhibit different percentages of shading which could impact differently on the output power of a PV module that has been installed within its reach.

Thin clouds are relatively transparent to sunlight [17], thereby providing a lower percentage of shading on the earth's surface when compared to thick clouds. These thin clouds are hard to identify on satellite images because they either reflect too little solar radiation or block too little terrestrial emission [18], while thick clouds are easier to detect [19]. Thick clouds are easy to distinguish because they are obviously brighter, and usually gathered together, forming a large block [20], resulting in a larger percentage of shading on the earth's surface. Cloud movement in SA is well documented, resulting in PV modules being exposed to both thin and thick clouds with varying percentages of shading.

These varying percentages of shading have the potential to negatively influence PV based systems, such as PV-electrolyzer hydrogen generators. These types of generator systems could be used as a standby power source for remote telecommunication sites [21]. A key requirement for these types of systems is direct sunlight, which could translate into an improvement factor of 1.2 in the hydrogen-production rate per unit area of the PV cells [22]. Direct beam radiation as opposed to diffuse beam radiation is therefore a must for the optimal operation of these types of energy generator systems. Quantifying the percentage of full uniform shading of a given PV module to its output power may assist in further analyzing the hydrogen-production rate for varying degrees of direct sunlight.

## III. Research Methodology

The first aim of this research is to verify the amount of light that passes through different shade nets that will be used in the full uniform shading of the PV modules. This is done using a constant light source, light sensor and two black cylinders (see next section for setup). A control and an experimental system are used.

The second aim of this research is to verify the influence of different percentages of full uniform shading on the performance of a PV module (termed PV Module 2). Each shade net is used to cover PV Module 2 for 10 minutes (experimental system), with 10 minute intervals where no shading was used. This gave PV Module 2 the opportunity to return to normal operation. PV Module 1 remained completely unshaded (control system) for the duration of the tests which occurred between 07:30 and 08:30 in the morning. A reason for choosing this time slot is that full

uniform shading from trees is more likely to occur during the rising and setting of the sun, when long shadows may be cast due to the sun's hour angle.

## IV. Practical Setup Shading

In order to establish an environment with minimum influence of external light, a constant light source, light sensor and cylinders were used. Two black cylinders (each 525 cm in length with a diameter of 27 cm) were staked upon each other (see Figure 2). The top cylinder was sealed from external light and a 12 V 3 W (230 lm) LED lamp (light source) was securely mounted in the center of the top cylinder. In the control system, the LED lamp was switched on and the light intensity was measured with using a PHYWE light intensity meter (light sensor mounted at the bottom of the bottom cylinder).

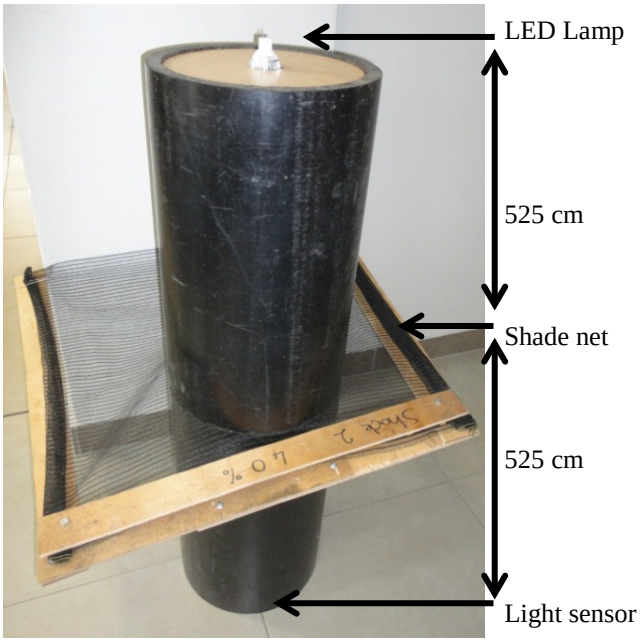


Figure 3: Practical setup of the full uniform shading experiment

In the experimental system, three different percentages of shade net were placed between the top and bottom cylinders (525 cm from the light source and light sensor). In this instance, light from the light source would have to travel through the shade net to reach the light sensor. This light intensity was measured to enable the exact calculations of the shading percentages. These shade nets were then used in the practical setup of the PV system.

## V. Practical Setup PV System

NO batteries were included in the practical setup of the two PV systems (exact duplicate of each other) due to uncertain variations which may exist between batteries from the same manufacturer and with the same model number. In fact, battery-to-battery variations in e.m.f at a given state of charge may be in the order of 50 mV due to variations in the manufacturing process, ageing and charge-discharge cycling of a single 2.25 V cell [23]. A 60 LED lamp (12 V, 3 W) was therefore chosen as the load resistance which was connected directly to the PV module via a 22 Ω resistor. The purpose of the series resistor is to raise the threshold operating voltage and ensure that the voltage across the LED never exceeds 12 V as maximum power point voltage of the module is 16.5 V. The threshold operating voltage is the point at which the LED starts emitting light, albeit it very faint. A data logging interface (DLI) was included between the PV module and the LED lamp, which serves to condition the voltage and current from the PV module to enable logging via a PC. This DLI circuit is based on previous research done by Swart [24] and Asowata [25, 26]. A PICOLOG 2016 is used as a data logger with its conversions parameters set to that prescribed by Swart [24], with readings taken every 10 seconds for 3 separate days in March 2014. Three different shade nets (with exact measured and calculated shading percentages) were placed over PV Module 2 (becoming the experimental system), while PV Module 1 remained completely unshaded (becoming the reference or control system).

An aluminum frame was constructed to securely mount the two identical 10 W PV modules. The modules were mounted at the same tilt angle of 39º, equating to 10º plus the latitude value of 29º for the Central University of Technology (CUT). Values of latitude plus 10º for PV module tilt angles in South Africa were suggested by Chinnery [27] and substantiated by Asowata [28]. The area is located in the semi-arid part of South Africa that enjoys 55% of its annual rainfall between January and April [29]. The practical setup was done inside an air-conditioned room where the temperature was kept constant at $26^0$C. This was in order to prevent excess temperature degradation which has a significant effect on the output voltage of a PV module [30]. The current through the LED lamp as well as the voltage across it was logged for both the control and the experimental system. The power dissipation for both systems was then calculated in MS Excel.

## VI. Results and Discussion

Figure 4 presents the results of two LED's which were considered for the load resistance in the practical setup (a 60 LED lamp and a 3 LED lamp operating at 12 V). The 60 LED lamp with NO series resistor (solid grey line) revealed a threshold operating voltage of 3.72 V (DC current must be more than 200 mA). Inserting a 22 Ω series resistance with this 60 LED lamp moves the threshold operating voltage to approximately 8.29 V. The 3 LED lamp's threshold operating voltage was 8.04 V with NO series resistance and 8.09 V with a 22 Ω series resistance (DC operating current is 5 mA). The maximum power point current value of 610 mA is shown using a thick dotted line. Choosing the 60 LED lamp in series with a 22 Ω resistor for the load would mean that the threshold operating voltage would have to be at least 8.29 V (this being 50% of the PV modules maximum power point voltage). The current at this point would have to be at least 200 mA (this being 33% of the PV modules maximum power point current). This would ensure that the PV module would be operating at an output power of more than 1.5 W (being 15% of its peak output power which equates to the efficiency of this PV module under STC).

The results of the shading experiment are shown in Figure 5. Using both the SUN and two different LEDs (3 W and 4 W) reveals similar results using the different shade nets. A 35% shade net allows 65% of light to pass while a 61% shade net allows 49% of light to pass.

Figure 6 shows the power reduction for a PV module (PV Module 2) which has been exposed to three nets with

different shading percentages. The obvious rise in output power is due to the rapid movement of the SUN between 07:30 and 08:30. Power reduction is just as evident, with roughly 0.6 W being produced by PV Module 2 (61% shade net used) around 08:20 when PV Module 1 is producing close to 3 W within the same controlled environment.

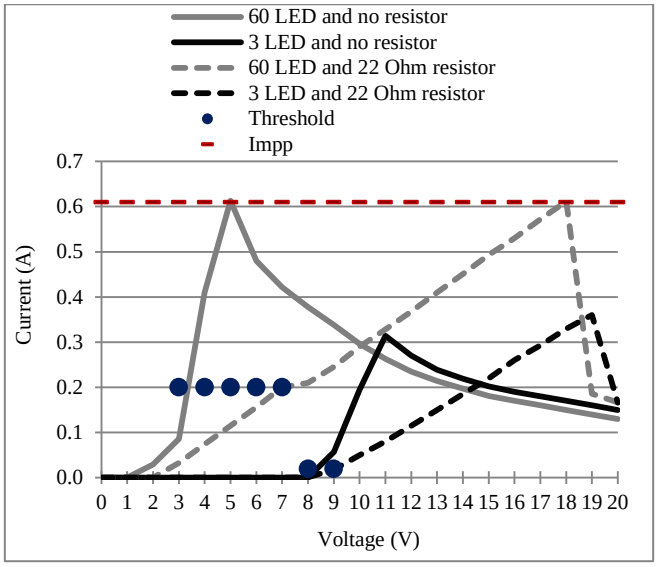



Figure 4: LED comparisons for the load resistance

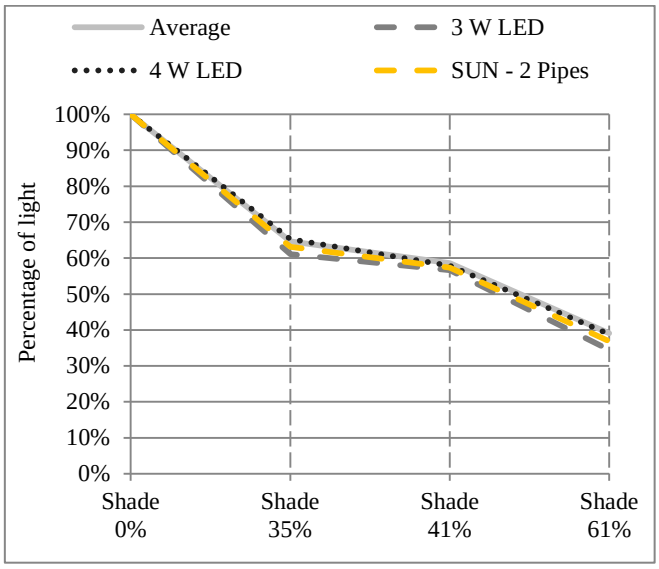



Figure 5: Full uniform shading results using LED's and the SUN

Figure 7 shows the calculated power difference between PV Module 1 and 2 for two days in March (24 and 27). An average calculation of these two results is used in the correlation to shade net percentages.

Figure 8 reveals the correlation of output power to the percentage of full uniform shading of a given PV module within a controlled environment. The percentage of full uniform shading equates to the percentage of light intensity of the direct beam component which is not proportional to the output power of the PV module (e.g. 35% shading allows 65% of light to pass which results in 50% of output power).

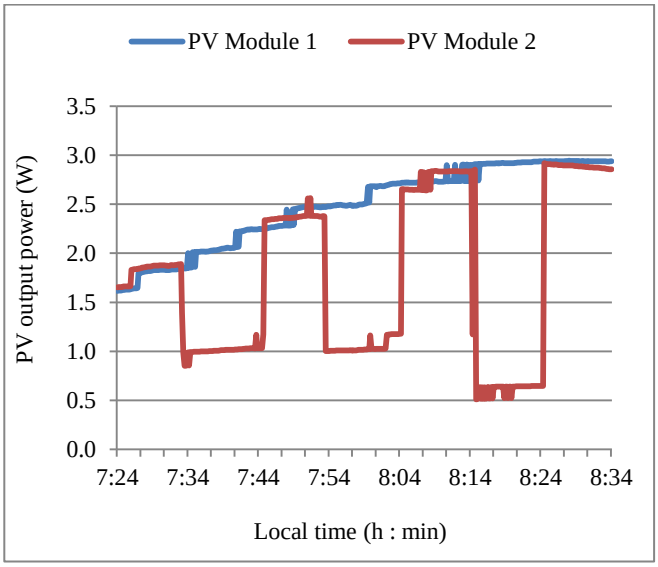



Figure 6: Output power of both modules – identical systems

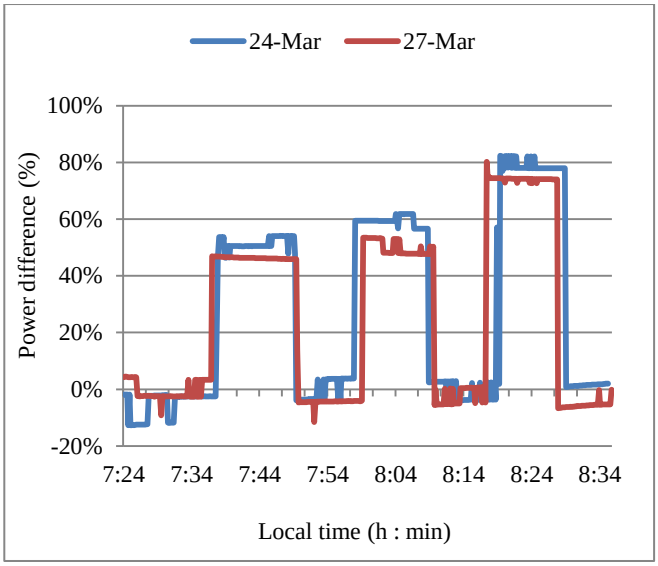



Figure 7: Percentage power difference between the systems

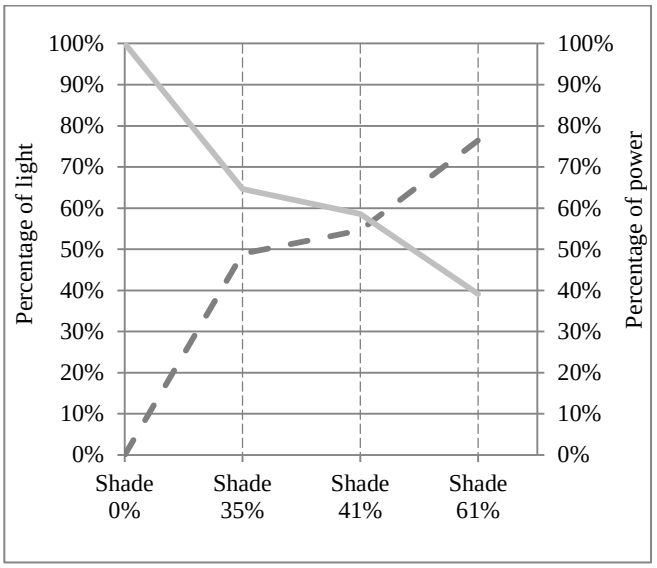



Figure 8: Power reduction to full uniform shading percentages

## VII. CONCLUSIONS

The purpose of this paper was to quantify the percentage of full uniform shading of a given PV module within a relatively pollution-free environment, correlating it to the output power of the module. Three different shade nets were quantified using the shading experiment encompassing two black cylinders, a constant light source and light sensor. Two identical PV systems were used, where PV Module 1 was the control system and PV Module 2 was the experimental system. Results reveal that a 35% full uniform shading net will result in a 50% output power reduction, while a 61% full uniform shading net will give a 76% power reduction.

## VIII. REFERENCES


[1] Brainy Quote. (2013, 18 March). Available: http://www.brainyquote.com/quotes/

[2] G. A. Karim, Fuels, energy, and the environment. Boca Raton: CRC Press, 2012.

[3] F. A. Farret and M. G. Simoes, Integration of Alternative Sources of Energy. Hoboken: John Wiley Press, 2006.

[4] A. J. Swart, R. M. Schoeman, and H. C. Pienaar, "Ensuring sustainability of PV systems for a given climate region in South Africa," presented at the AFRICON 2013, Mauritius, 2013.

[5] K. K.A., "Voltage-offset resistive control for photovoltaics," Master of Science, Electrical and Computer Engineering, University of Illinois, Illinois, 2011.

[6] D. Giaffreda, M. Omana, D. Rossi, and C. Metra, "Model for Thermal Behavior of Shaded Photovoltaic Cells under Hot-Spot Condition," in Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT), 2011 IEEE International Symposium on, 2011, pp. 252-258.

[7] REN21. (2014). *Homepage*. Available: http://www.ren21.net/

[8] D. W. Christy, "An Experimental Evaluation of the Performance of the Amorphous Silicon PV Array on the NCSU AFV Garage," Masters of Science, Mechanical Engineering, North Carolina State University, Raleigh, 2007.

[9] J. Gummeson, S. S. Clark, K. Fu, and D. Ganesan, "On the limits of effective hybrid micro-energy harvesting on mobile crfid sensors," in Proceedings of the 8th international conference on Mobile systems, applications, and services, 2010, pp. 195-208.

[10] Z. S. Johnson, "Module-level power converters for parallel connected photovoltaic arrays," Master of Science, Missouri University of Science and Technology, Missouri, 2012.

[11] A. A. El-Sebaii, F. S. Al-Hazmi, A. A. Al-Ghamdi, and S. J. Yaghmour, "Global, direct and diffuse solar radiation on horizontal and tilted surfaces in Jeddah, Saudi Arabia," Applied Energy, vol. 87, pp. 568-576, 2010.

[12] T. V. Ramachandra and B. V. Shruthi, "Spatial mapping of renewable energy potential," Renewable and Sustainable Energy Reviews, vol. 11, pp. 1460-1480, 2007.

[13] A. J. Swart, H. C. Pienaar, and R. M. Schoeman, "Assessing the effect of variable atmospheric conditions on the performance of photovoltaic panels: A case study from the Vaal Triangle," presented at the SAEEC 20111, Emperor's Palace Convention Centre, Johannesburg, 2011.

[14] S. R. Wenham, M. A. Green, M. E. Watt, and R. Corkish, Applied photovoltaics, 2nd ed. Cornwall: TJ International Ltd, 2007.

[15] D. S. Gardner and R. M. Goss, "Management of turfgrass in shade," Turfgrass: Biology, Use, and Management, pp. 219-247, 2013.

[16] J. Shultz, L. Witmer, J. E. Rey, and S. Brownson, "Impact of shade on HVAC energy consumption in buildings: a residential case study," The Pennsylvania State University, 2011.

[17] P. Y. Groisman, T. R. Karl, and R. W. Knight, "Observed impact of snow cover on the heat balance and the rise of continental spring temperatures," Science, vol. 263, pp. 198-200, 1994.

[18] W. Sun, G. Videen, S. Kato, B. Lin, C. Lukashin, and Y. Hu, "A study of subvisual clouds and their radiation effect with a synergy of CERES, MODIS, CALIPSO, and AIRS data," Journal of Geophysical Research: Atmospheres (1984–2012), vol. 116, 2011.

[19] A. Ahmad, B. M. Aboobaider, M. Ghani, K. Abdul, S. Razali, S. M. Isa, and N. M. Hashim, "A Localised Cloud Detection and Masking Method Using Spectral Analysis," Australian Journal of Basic & Applied Sciences, vol. 7, 2013.

[20] Y. D. Yu, X. B. Yang, S. J. Xiao, and J. L. Lin, "Automated ship detection from optical remote sensing images," Key Engineering Materials, vol. 500, pp. 785-791, 2012.

[21] A. O. Amoo, W. J. Bekker, and v. H. C. Pienaar, "Solar Driven Hydrogen Generation for a Telecommunications Fuel Cell Power Plant," in SATNAC 2010, Spier Wine Estate, Stellenbosch, 2010.

[22] N. A. Kelly and T. L. Gibson, "Hydrogen generator photovoltaic electrolysis reactor system," ed: Google Patents, 2008.

[23] M. L. Gopikanth and S. Sathyanarayana, "Impedance parameters and the state-of-charge. II. Lead-acid battery," Journal of Applied Electrochemistry, vol. 9, pp. 369-379, 1979/05/01 1979.

[24] A. J. Swart, H. C. Pienaar, and R. M. Schoeman, "Cost-effective energy monitoring of domestic off-grid PV systems," presented at the APPEEC 2013, Beijing Yanshan Hotel, Beijing, China, 2013.

[25] O. Asowata, A. J. Swart, and H. C. Pienaar, "Evaluating the effect of a stationary PV panel on the charging rate of Deep-Cycle Valve-Regulated Lead-Acid Batteries," presented at the AFRICON 2013, Mauritius, 2013.

[26] O. Asowata, A. J. Swart, and H. C. Pienaar, "Optimum tilt and orientation angles for photovoltaic panels in the Vaal Triangle," presented at the APPEEC 2012, Grand Mercure Baolong Hotel, Shanghai, China, 2012.

[27] D. N. W. Chinnery, "Solar heating in South Africa," Pretoria CSIR-Research Report 248, 1981.

[28] O. Asowata, A. J. Swart, H. C. Pienaar, and R. M. Schoeman, "Optimizing the output power of a stationary PV panel," presented at the SATNAC 2013, 2013.

[29] H. Snyman, W. Venter, W. Van Rensburg, and D. Opperman, "Ranking of grass species according to visible wilting order and rate of recovery in the Central Orange Free State," Journal of the Grassland Society of southern Africa, vol. 4, pp. 78-81, 1987.

[30] A. Ozemoya, J. Swart, and C. Pienaar, "Controlling the ambient temperature of a PV panel to maintain high conversion efficiency," presented at the SATNAC 2012, George, South Africa, 2012.



**James Swart** received his DTech: Electrical: Engineering degree in 2011 from the Vaal University of Technology. His research interests include engineering education and alternative energy.

**Pierre Hertzog** received his DTech: Electrical: Engineering degree in 2004 from the Central University of Technology. His research interests include alternative energy and automation & manufacturing systems.

# WORK IN PROGRESS:

# ACCESS NETWORK TECHNOLOGIES

# Incremental FTTH deployment planning

J Laureles •, M.J. Grobler •, S.E. Terblanche °

TeleNet Research Group
•School for Electrical, Electronic and Computer Engineering
°Centre for Business Mathematics and Informatics
North-West University, Potchefstroom Campus
Email: {21640653, leenta.grobler, fanie.terblanche}@nwu.ac.za

***Abstract*—An increase in the demand for network availability has been experienced in growing urban areas. However, expanding a network when the need arises, with no planning in advance for possible future expansions, can be very expensive. Therefore, if the initial planning takes possible future expansions into account, it may allow for cheaper/more effective incremental deployments. This research focuses on developing a mathematical model that takes into account relevant aspects for determining the optimal solution for deploying a Fibre-to-the-home (FTTH) Passive optical network (PON), subject to incremental demands. Success of this research means that service providers will be able to design and deploy a network in a shorter time than by using manual planning.**

***Index Terms*—FTTH, Incremental network deployment, Optical fibres, Optical network planning, PON**

## I. Introduction

WITH society growing at an alarming rate, geographical areas that were once uninhabited are slowly growing in population, extending urban areas [1] in various directions. The demand for network availability has, as a result, increased rapidly, thus forcing service providers to meet these demands. To upgrade and extend already deployed networks has been found to be very expensive [2] which is why a more cost effective solution is desired.

Network plans are currently manually designed [3] whereby an employee iterates through the possible layouts attempting to find the optimal one. Manual-iteration is a time consuming process with results that, when compared to a more automated process, are solely dependent on the employee's knowledge on network deployment. A minor mistake can end up costing the service provider a lot of money. Although this method has been implemented and used successfully, it still contains many risks.

Related work using PON design include *Single PON network design with unconstrained splitting stages* [4] and *Optimization of passive optical network planning for fibre-to-the-home applications* [3], both very recent and successful studies. Although both of these studies include PON network design, neither specifically address the incremental network plan.

With this being said, this research aims on developing a solution whereby an optimal FTTH PON can be designed with the capabilities of future expansions. The main focus is therefore on minimising deployment costs and manual labour as well as addressing the incremental demand problem.

The remainder of this paper is organised as follows: Section II briefly discusses the background behind network planning and what it entails. Section III compares two types of network plans, namely the Just-In-Time and Incremental Plan, and makes use of them to clarify the purpose of the research. Section IV presents the research proposal, objectives and methodology. Section V discusses the expected results, should the research be successful and lastly Section VI ends the paper off with a conclusion.

## II. Background

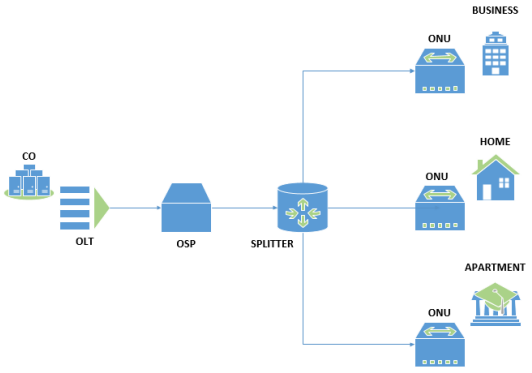



Fig. 1. A basic example of a network layout

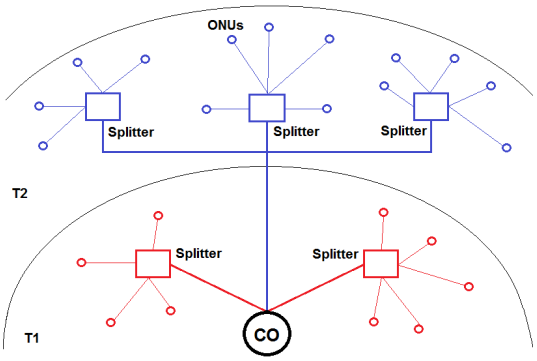
As depicted in Figure 1, a typical access network consists of a Central office (CO) of a Service Provider (SP), splitters and Optical network unit (ONU)s. The basic aspects connecting the CO to the subscriber, be it a household, apartment or business, originates from the Optical line terminal (OLT), located at the CO, through cables capable of transferring digital data to an Outside plant (OSP). A splitter located at the OSP then creates the necessary splits to the various destinations, namely the ONUs located at the various subscriber's premises [5].

Deploying an entire network requires accurate information on the geographical area, population distribution and the location of the present infrastructures and components [5]. These factors are the first data sets required and form the basis of planning a network.

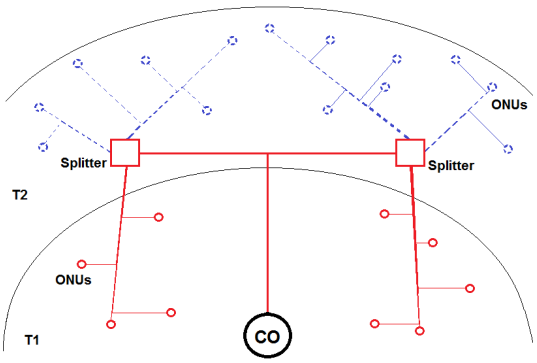
## III. Research Purpose

Figure 2(a) depicts the Just-In-Time Plan that is currently used by many service providers. The plan operates on a "rise-of-demands" basis, whereby the network is laid out according to the current demand. As the demand increases, so too, does the network expand accordingly.

An Incremental Plan on the other hand, takes future demands into account when laying out a network. For example, the two splitters depicted in Figure 2(b), are positioned to




(a) Example of a Just-In-Time Plan




(b) Example of an Incremental Plan

Fig. 2. Just-In-Time- vs Incremental- Plan.

accommodate the demands not only for the first time period, T1, but for the second time period, T2, as well.

A decrease in the amount of splitters required is immediately noticeable when comparing Figure 2(a) to Figure 2(b). Fewer splitters indicates fewer trenches that have to be dug up. The fewer trenches and splitters needed to build a network over multiple time periods, lowers the overall cost of implementing and expanding a network. With this research focused on determining an optimal solution for current and future demands simultaneously, service providers could end up saving greatly for each new and upgraded network.

## IV. Proposed Research

This research is aimed at developing a mathematical model and solution approaches for the problem of deploying an FTTH PON taking incremental demands into account.

### A. Research Objectives

The primary objective is to model and optimize the deployment of an incremental, time-based FTTH .

The secondary objectives are as follows:

- To implement theoretical data referenced from a real life PON deployment problem
- To design a mathematical model that, when used, can assist in determining the optimal network deployment solution
- To solve the model through the aid of CPLEX [6]
- To implement heuristics in obtaining a close to optimal solution

### B. Research Methodology

The following outlines the steps that will be performed to obtain a valid model:

- To thoroughly understand a single network and the purpose of each component,
- To design a mathematical model based on theoretical data,
- To identify possible constraints, such as the location of roads, houses, etc. to accurately model the problem,
- To design a small scale scenario for the model and test whether it can accurately determine results in a reasonable amount of time,
- To enlarge the deployment problem to test the accuracy of the model with more data points and lastly,
- To perform validation and verification tests to ensure the accuracy of the model.

## V. Expected Results

Focus has, thus far, been placed on getting acquainted with network planning and design. From what has been observed, the multiple data points are expected to increase the level of difficulty when determining the optimal solution. After the research has concluded however, a close to optimal, not necessarily fully optimal, solution is expected.

## VI. Conclusion

The successful completion of this research has the potential of solving the growing network availability problem. Not only will service providers be able to eliminate having to manually design an entire network, but they will also be able to extend networks years after initial deployment. This will not only prevent unnecessary expenses from taking place, but also decrease the duration it currently takes to design an optimal network. With less time needed to design a network, service providers will be able to design and possibly deploy more networks in the same amount of time than it currently takes. In doing so, service providers will be able to keep up with the incremental demand rate, satisfying both old and new clients, whilst avoiding unnecessary expenses, an advantage for both parties.

## References

[1] S. Bakhshi and C. Dovrolis, "The price of evolution in incremental network design: The case of mesh networks," in *IFIP Networking Conference, 2013*. IEEE, 2013, pp. 1–9.

[2] S. Verbrugge, K. Casier, B. Lannoo, J. Van Ooteghem, R. Meersman, D. Colle, and P. Demeester, "Ftth deployment and its impact on network maintenance and repair costs," in *Transparent Optical Networks, 2008. ICTON 2008. 10th Anniversary International Conference on*, vol. 3. IEEE, 2008, pp. 2–5.

[3] S. van Loggerenberg, "Optimization of passive optical network planning for fiber-to-the-home applications," Master's thesis, North West Universtiy - Potchefstroom Campus, 2013.

[4] L. Gouveia, M. J. Lopes, and A. de Sousa, "Single pon network design with unconstrained splitting stages," *European Journal of Operational Research*, 2014.

[5] J. Segarra, V. Sales, and J. Prat, "Planning and designing ftth networks: elements, tools and practical issues," in *Transparent Optical Networks (ICTON), 2012 14th International Conference on*. IEEE, 2012, pp. 1–6.

[6] I. ILOG, "Cplex optimization studio," *online] http://www-01. ibm. com/software/integration/optimization/cplex-optimization-studio/, last accessed Oct*, vol. 26, 2012.

**Jonabelle Laureles** is a Telkom CoE student who received her undergraduate degree B.Eng in Computer and Electronic Engineering at the North West University Potchefstroom Campus in 2013. She is currently working towards an M.Eng degree in Computer Engineering at the same institution. Her research interests include mathematical modelling, network modelling and positioning.

# An Efficient Sum-Product Decoding Algorithm for Quasi-Cyclic LDPC Codes

Yuval Genga, Jaco Versfeld
School of Electrical & Information Engineering, University of the Witwatersrand,
Private Bag 3, 2050, Johannesburg, South Africa
Email: Yuval.Genga@students.wits.ac.za; jaco.versfeld@wits.ac.za

***Abstract*—Since the rediscovery of Low Density Parity Check Codes (LPDC) in the late 1990s, this class of codes have become increasingly popular due to their good error correction capabilities. The Quasi-cyclic (QC) LDPC codes have been shown to perform very well over the AWGN channel with iterative decoding. Decoding algorithms such as the Fast Fourier Transform q-ary sum product algorithm (FFT-QSPA) and the Min Sum Algorithm (MSA) yield a high BER performance but at the expense of computational complexity. However, a decoding scheme called the circulant permutation matrix revolving iterative decoding scheme (CPM-RID) was recently devised by Liu, Lin and Abdel-Ghafar and is believed to reduce the complexity of any LDPC updating algorithm when implemented properly. This research will set out to attempt to implement the CPM-RID decoding scheme with the FFT-QSPA in an attempt to reduce the computational complexity of the decoding algorithm without any BER performance degradation**

***Keywords*—Circulant permutation structure, LDPC, quasi cyclic structure,revolving iterative decoding.**

## I. Introduction

Low Density Parity Check (LDPC) codes were discovered by Gallager in his PhD thesis [1] in 1963 but due to their computational and implementation complexity at that time they were mostly ignored. It wasn't until the late 1990's when this class of codes were rediscovered by Mackey and Neil [2], who proved a near Shannon limit performance when decoding using probabilistic soft decision decoding algorithms.

Nonbinary LDPC codes, commonly referred to as q-ary LDPC codes, were first investigated by Davey and MacKey [3]. The research by Davey and Mackey involved iterative decoding using the Sum Product Algorithm (SPA). Davey and Mackey applied the SPA for binary LDPC codes to decoding q-ary LDPC codes and referred to this extension as q-ary SPA (QSPA). Despite the good decoding performance, computational complexity of the QSPA was still considerably high. To counter this issue Barnault and Declercq devised a Fast Fourier transform of the q-ary SPA and referred to it as FFT-QSPA [4]. The FFT-QSPA proved to be more effective than QSPA as it reduced the computational complexity without performance degradation [5].

Even though FFT-QSPA significantly reduces the computational complexity of the QSPA, it is still rather complex as it incorporates a high number of multiplications in the probability domain for both check node and variable node processing. As a result the development of a logarithmic domain approach was applied to approximate the QSPA, called the Min Sum Algorithm (MSA). The MSA proved to reduce the computational complexity of the decoding process, but this was at the cost of worsened BER. Liu, Lin and Abdel-Ghaffar [6] proposed a decoding scheme called the revolving Iterative decoding (RID) scheme to counter the issue of complexity versus performance. In their paper they indicated that the RID scheme can be incorporated with any updating algorithm and can reduce the computational complexity of the said algorithm. They went on to prove it with MSA. For decoding quasi cyclic (QC)-LDPC codes with a circulant permutation matrix (CPM) structure, a modification of the RID called the CPM-RID is used [7]. The CPM-RID scheme proposed in [7] was applied to the MSA decoding algorithm. This research will therefore look at the implementation of the CPM-RID scheme for QC-LPDC codes using the FFT-QSPA decoding algorithm. This will be done so as to see if the CPM-RID scheme can reduce the computational complexity of the FFT-QSPA algorithm and at the same time give the same BER performance.

## II. The CPM-RID Decoding Scheme

The Revolving Iterative Decoding (RID) scheme is an effective and reduced complexity algorithm for decoding algebraic cyclic and QC-LDPC codes. It decodes well for both binary and non-binary cases [6]. The process of decoding is carried out based on a single small submatrix of the parity check matrix of the code in a revolving manner [6]. The RID proposed by Liu, Lin and Abdel-Ghaffar is shown to reduce the hardware implementation complexity and the size of memory required to store the information. An indicator of the hardware complexity of the QC-LDPC decoder is the number of check node (CN) message processing units and the number of connections from the variable nodes (VN) messages processing units and the CN processing units [7]. This decoding algorithm of the RID is only applicable when it is based on the *block cyclic structure* of the parity check matrix of QC-LDPC codes [6]. However most QC-LDPC codes are defined by parity check matrices that have a *circulant permutation matrix* (CPM) structure. To resolve this issue a modification to the RID called the CPM-RID was proposed by Liu, Lin and Abdel-Ghaffar [7]. The CPM-RID scheme is devised based on the CPM structure of the parity check array of the H matrix of a QC-LDPC code without column and row permutations [7]. The advantage of the CPM-RID scheme is that it retains the reduced hardware decoding complexity of the RID decoding scheme and also maintains the simple wire routing due to the CPM structure [8]. For the CPM structure to be maintained on the H matrix a *section wise-cyclic shift* must be performed. Furthermore, if an algorithm to update the reliabilities of the decoded symbols is properly chosen, the computational complexity can be reduced as well.

### A. Section wise cyclc shift

To understand what a section-wise cyclic-shift is let us consider a parity check matrix H. The matrix H has a $m \times n$ array of circular permutation matrices (CPMs) and/or zero matrices (ZMs) of size $(q-1) \times (q-1)$. Let the $m$ row blocks of H be denoted as $H_0$, $H_1$,....$H_{m-1}$. Each $m$ row block consists of $n$ CPMs and/or ZMs of size $(q-1) \times (q-1)$. For $0 \leq i < m$ and for $0 \leq k < q-1$, let $H_{i,k}$ = ($h_{i,k,0}$, $h_{i,k,1}$.....$h_{i,k,n-1}$) be the $k$-th row in the $i$-th row block $H_i$ which consists of $n$ sections, $h_{i,k,0}$, $h_{i,k,1}$.....$h_{i,k,n-1}$, each containing $q-1$ components. Each section of the row block $H_i$ is the k-th row of either a CPM or a ZM. If all the $n$ sections of $h_{i,k}$ are cyclically shifted simultaneously one place to the right within the sections, we obtain a $(k+1)$-th row $h_{i,k+1}$ of $H_i$. This cyclic-shift within each section is referred to as the section-wise cyclic-shifting of the rows $h_{i,k}$. For the case of $k = q-2$, the section wise cyclic shift of $h_{i,q-2}$ results in the 0-th row $h_{i,0}$ of $H_i$. Thus all the rows of the $i$-th row-block $H_i$ can be obtained by section-wise cyclically shifting the 0-th row $h_{i,0}$ $q-2$ times. This section-wise cyclic shifting of the rows of H maintains the CPM-structure of each row block H [7].

Let us consider $H_0^*$ to be an $m \times n(q-1)$ matrix which consists of the first rows $h_{0,0}$, $h_{1,0}$,..., $h_{m-1,0}$ of the $m$ row-blocks $H_0$, $H_1$,..., $H_m - 1$ of the parity-check array H. Thus from the section-wise cyclic structure of H, we can obtain the entire array of H by section-wise cyclically shifting $H_0^*$ $(q-2)$ times. Due to the section-wise cyclic structure we can decode the QC-LDPC codes given by the parity check array H based on the submatrix $H_0^*$ alone in almost the same way as the RID scheme proposed in [6][7].

Each decoding iteration based on $H_0^*$ is called the *decoding sub-iteration*. The reliabilities of the received symbols are updated, at the end of each decoding sub-iteration, with a chosen reliability updating algorithm . Therefore, the reliability vector ($n(q-1)$ components in $n$ sections) and the received sequence ($n(q-1)$ symbols in $n$ sections) are section-wise cyclically shifted to the left by one position and used as the input information, to carry out the next decoding sub-iteration based on $H_0^*$ [7]. It can be seen that any $q-1$ decoding sub-iterations performed based on $H_0^*$ are equivalent to one decoding iteration based on the entire parity-check array H. At the end of each decoding sub-iteration, the syndrome **s** of the hard-decision of the received sequence (after section-wise cyclically shifted one position to the left) is computed based on the entire parity-check array H. If s = 0, we stop the decoding process; otherwise, we continue the decoding process until a preset maximum number of decoding sub-iterations are reached [7].

This is a revolving iterative decoding (RID) scheme as it revolves around $H_0^*$ iteratively for the decoding process of QC-LDPC codes. This variation of the RID scheme proposed in [6] is referred to as the CPM-RID scheme. The submatrix $H_0^*$ is referred to as the *decoding matrix* [7].

The CPM-RID can be incorporated with any known updating reliability scheme to decode a QC-LDPC code with a CPM-structure. For this research FFT-QSPA decoding algorithm will be incorporated with the CPM-RID scheme. The FFT-QSPA has been selected because of its good BER performance and its high computational complexity. The CPM-RID has been said to reduce the complexity of the updating algorithm [6]. Hence this research will attempt to reduce the computational complexity of the FFT-QSPA while decoding QC-LDPC codes.

## III. Research Question

What performance and computational complexity does the FFT-QSPA attain when incorporated with the CPM-RID?

## IV. Methodology

In order to perform the research, simulations will be run in MATLAB to compare the performance of the FFT-QSPA and the CPM-RID modification of the FFT-QSPA. A non-binary QC-LDPC code will be constructed over GF($q$) whose parity check array consists of $\alpha$-multiplied CPMs where $\alpha$ is a primitive element of GF($q$) and $q$ is a prime or a power of a prime. The performance of the decoding algorithms on the QC-LDPC code will then be analyzed over a AWGN channel using 16-QAM as a modulation scheme. This will all be done via simulation. The performance of both the FFT-QSPA and the CPM-RID FFT-QSPA will compared by using Monte Carlo simulations. If time allows, the algorithm will be implemented on a FPGA.

## V. Conclusion

Liu, Lin and Abdel-Ghaffar [7] showed that it is possible to reduce the complexity of an updating algorithm and maintain the same BER performance with the CPM-RID decoding scheme. They, however, proved this with the MSA and implemented it on binary QC-LDPC codes. This research aims to attain similar results with the better performing but more complex FFT-QSPA, when applied to non-binary QC-LDPC codes.

## References


[1] R.G. Gallager,"Low Density Parity Check Codes", *IRE Trans. Inform Theory*,. vol IT-8, no.1, pp. 21-28, Jan 1962.

[2] D. J.MacKey and R. M. Neil, "Near Shannon limit performance of low density parity check codes", *Electro. Lett*,. vol. 32, pp. 1645-1646, Aug. 1996.

[3] M. Davey and D. J.MacKey,"Low density parity check codes over GF(q)", *IEEE commun. Lett*, vol 2. no. 6, pp. 165-167, June 1998.

[4] L.Barnault and D. Declercq,"Fast Decoding Algorithm for LDPC over GF(2q)" *Proceeding of ITW2003*, 3rd ed. pp. 70-73, Paris, France, March 31- April 4,2003.

[5] S. Lin, S. Song L.Lan, L.Zeng and Y. Tai,"Construction Of Nonbinary Quasi-Cyclic LDPC Codes: A Finite Field Approach" *IEEE Trans. Commun,*, vol. 56, no. 4, pp.545-554, April 2008.

[6] K.Liu, S. Lin and K Abdel-Ghafar, "A Revolving Iterative Algorithm for Decoding Algebraic Cyclic and Quasi-Cyclic LDPC Codes", *IEEE Commun. Lett,*, vol. 61, no. 12, pp.4816-4827, Dec 2013

[7] K.Liu, S. Lin and K Abdel-Ghafar, "Decoding of Quasi-Cyclic LDPC Codes With Section-Wise Cyclic Structure"*Information Theory and Applications Workshop*, pp.1-10, Feb 2014

[8] Y.Chen and K. Parhi, "Overlapped Message Passing for Quasi-Cyclic Low Density Parity-Check Codes",*IEEE Trans. circuits and I systems,*, vol. 51, no. 6, pp.1106-1113, Jun 14, 2004.



**Yuval Genga** - Is a Masters student at the University of the Witwatersrand. He is currently doing his research in the field of telecommunications. His research interests include forward error correction.

# QoS Aware scheduler algorithm for guaranteed throughput

E. M. Mthimunye[1]. K. Djouani[2] and A. Kurien[3]
Department of Electrical Engineering Tshwane University of Technology[1], PBag X680, Pretoria 0001
Tel: +27 12 382 4191, Cell: 0814354199 Fax: +27 12 382 5294
Email: Tsatsi.Mthimunye@gmail.com, Mthimue@telkom.co.za[1]

***Abstract***- Long Term Evolution (LTE) adopts shared-channel transmissions in which time and frequency resources are dynamically shared among user equipment (UE's). As such an optimal utilisation of shared network resources is a key factor for broadband wireless access systems (BWAS). Packet scheduling is one of the Radio Resource Management (RRM) algorithms which manage the allocation of shared resources to users. This study proposes a resource scheduling algorithm that incorporates the admission and congestion control algorithm to guarantee QoS under impact of interference. The proposed scheduler implemented on the Medium Access Control (MAC) considers two main types of bearers: Guaranteed Bit Rate (GBR) and Non-Guaranteed Bit Rate (non-GBR). GBR users are classified as users who are willing to pay more for wider bandwidth and better network access, therefore their services require higher priority handling. Non-GBR service requires no priority handling [10]. The proposed scheduler algorithm with admission and congestion control algorithm controls the admission of GBR and non-GBR users, maximises the system capacity while satisfying the QoS requirements of admitted services and manages the load. The algorithm will prove that the system can guarantee QoS, improve spectral efficiency and good throughput for delay sensitive users while maintaining acceptable interference levels.

Index Terms—resource scheduling, GBR, Non-GBR, LTE

## I. INTRODUCTION

Future broadband wireless systems are expected to support the rapidly growing traffic demands from end-users. This is evident by the increasing development of various low-cost smart phones, tablets, laptops and other mobile computing devices. These devices provide an excess of applications such as voice, image, video and multimedia services. LTE is one of the technologies chosen to deliver on these expectations and it uses Orthogonal Frequency Division Multiple Access (OFDMA) on the downlink and Single Carrier – Frequency Division Multiple Access (SC-FDMA) on the uplink [10]. LTE adopts shared-channel transmissions in which time and frequency resources are dynamically shared among user equipment (UE's). Downlink packet scheduling in LTE is performed at the MAC layer and it allocates resources on the physical downlink shared channel (PDSCH) to UEs and selects appropriate MCSs for the transmission of system information and user data [10]. Since the users experience different Signal to Interference Noise Ratio (SINR), the scheduler decides which users to schedule and the number of Physical Resource Blocks (PRB's) to allocate. As the system load increases the LTE system is expected to maintain the QoS of admitted users and manage the interference.

3GPP specifies the RRM related signals [10], but the actual algorithms in the network are not defined and this study will cover some of the studied algorithms e.g. Proportional Fair (PF), Modified Largest Weighted Delay First (M-LWDF), Exponential Proportional Fair (EXP/PF) and Inter-Cell Interference Coordination (ICIC) techniques. The M-LWDF algorithm considers the QoS of users, channel quality and adds the weight and Head of Line (HOL) packet delay [2]. The only drawback is, the algorithm does not consider the average HOL packet delay and interference. Subsequent to M-LWDF an improved algorithm EXP/PF was proposed to cater for both real-time and non-real time traffic [3]. The algorithm does consider fairness and average HOL packet delay, however the impact of external interference was not considered. Some of the authors have researched the inter cell interference co-ordination algorithm, however the algorithm compromised the cell edge throughput [8].

Based on the limitations mentioned above we propose the optimal QoS scheduler which incorporates the admission and congestion control algorithm to control admission of users, maximize the system capacity, guarantee QoS under deteriorating radio conditions. The scheduler will be implemented on MATLAB software for verification.

## II. PROBLEM STATEMENT

The overall goal of the packet scheduler is to maximize the cell capacity, while making sure that the minimum QoS requirements of the EPS bearers are met and that there are sufficient resources for best effort bearers. Since circuit switched connections are not supported in LTE there is a high demand for a well performing QoS system and interference management mechanism.

This leads to the following research questions:

*(i) How can the proposed optimal QoS scheduler algorithm be applied on an LTE network in order to guarantee QoS for delay sensitive traffic under high load and* deteriorating *radio conditions?*

*(ii) How to formulate a model targeted at minimising interference without compromising on throughput?*

*(iii) How to formulate a model aimed at maximising the spectral efficiency?*

## III. RESEARCH OBJECTIVE

The main objective is to show how we can offer QoS under deteriorating radio conditions or under impact of interference margin.

## IV. ADMISSION CONTROL AND SCHEDULER TECHNIQUE

LTE scheduling is the responsibility of the eNB in LTE, however some of the additional aspects of scheduling and QoS handling could take place in the Evolved Packet Core (EPC). Historically the radio interface is considered to be the weak link in the overall end-to-end mobile service systems.

This was practically due to limited physical radio resources i.e. limited bandwidth or channels in the system. As a result the RRM principle is one of the major functionalities of the LTE system, whereby all UE's with Radio Resources Control (RRC) connection in a cell share the channels and resources in the cell.

The LTE packet schedulers function is to allocate time-frequency resources to UEs in each subframe for both uplink and downlink transmission. The LTE Resource Block (RB) has a duration of 0.5ms and a bandwidth of 180 kHz (12 subcarriers). The minimum scheduling unit consists of 12 subcarriers and 1 subframe with a duration of 1ms [10].

The admission control algorithm maximizes system capacity while satisfying QoS requirements of admitted services stability by admitting or rejecting services as shown on fig1. Admission control uses the QoS satisfaction rate of GBR services to reflect QoS conditions of admitted GBR services in the cell.
Congestion control reduces congestion caused by an insufficient of radio resources.

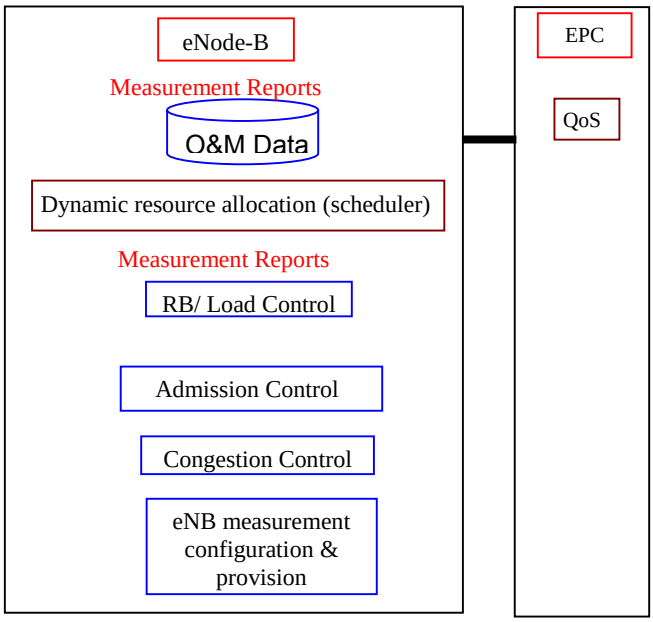


Fig1: 3GPP Algorithm model

The proposed scheduler algorithm will be evaluated through the use of the following formulas:

User throughput will be computed using the following formula:

$$m_{i,k} = \frac{d_k^i(t)}{R^i(t-1)} \times exp\left(\frac{\alpha_i D_i - x}{1+\sqrt{x}}\right) \times i_o \quad (1)$$

$$\alpha_i = -\frac{log\delta_i}{\tau_i} \quad (2)$$

$$x = \frac{1}{N_{rt}} \sum_{i=1}^{N_{rt}} \alpha_i D_i \quad (3)$$

$$R_i(t) = \left(1 - \frac{1}{t_c}\right) \times R_i(t-1) + \frac{1}{t_c} \times r_i(t-1) \quad (4)$$

$$d_k^i(t) = R(t) \times \frac{n\_bits_{i,k}(t)}{symbol} \times \frac{n\_symbol}{subcarrier} \times \frac{n\_subcarrier}{RB} \quad (5)$$

Where $\delta i$ is the acceptable packet loss rate of the i-th user, $\tau i$ is the delay threshold of the i-th user and $Nrt$ is the number of real-time users. Furthermore, let $Di$ be the delay of the first packet in the queue of the i-th user. Io is the inter-cell interference factor.

## V. Research Methodology

As system load increases the system quality degrades due to interference. This research project aims at investigating and developing an optimal solution to guarantee QoS under deteriorating radio conditions. The following approach will be adopted to achieve this goal:

*Phase 1*: Literature Review
Various latest studies and developments on 3GPP's LTE scheduling techniques will be thoroughly surveyed. Thereafter, useful information and facts necessary to formulate the final research problem will be identified and established.

*Phase 2*: Mathematical Modelling and System Design
The proposed scheduler algorithm will be mathematically developed based on the formulated problem. Standard scheduling algorithms will be analysed and compared to the proposed QoS scheduler which guarantees QoS.

*Phase 3*: Implementation and simulation studies
The proposed algorithm will be simulated in an extensive network simulator using ~~a~~ MATLAB.

*Phase 4*: Analysis of results and thesis report write up
The obtained simulation results will be thoroughly analysed and thereafter published.

**Evah Mmatsatsi Mthimunye** received her undergraduate degree in 2006 from the Tshwane University of Technology and is presently studying towards her Master of Technology degree at the same institution. Her research interests include QoS aware scheduler in LTE, interference and Spectral efficiency in LTE.

References:

[1] Proebster, Magnus, Mueller, Christian M., Bakker, Hajo, "Adaptive fairness control for a proportional fair LTE scheduler," IEEE PIMRC, pp. 1504-1509, 2010
[2] Xian, Yong-Ju; Tian, Feng-Chun; Xu, Chang-Biao; Yang, Yue; "Analysis of M-LWDF fairness and an enhanced M-LWDF packet scheduling mechanism," Journal of China Universities of Posts and Telecommunications, vol. 18, no. 4, pp. 82-88, August 2011
[3] W. Fu, Q. Kong, W. Tian, C. Wang, L. Ma, "A QoS-Aware Scheduling Algorithm Based on Service Type for LTE Downlink" Xidian University, Xi'an, China. 2013
[4] 3GPP TS36.211. V8.0.0, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation," www.3gpp.org. Oct, 2010
[5] 3GPP TS36.300. v10.l.0. "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN)," www.3gpp.org, Oct, 2010
[6] L. Hentilä, P. Kyösti, M. Käske, M. Narandzic , and M. Alatossava. (2007, December.) MATLAB implementation of the WINNER Phase II Channel Model ver1.1 [Online]. Available: http://projects.celtic-initiative.org/winner+/phase_2_model.htm
[7] S. Zhang, \Inter-cell interference coordination in indoor LTE systems," Masters Thesis, Royal Institute of Technology, Sweden, 2011.
[9] [Online]. Available: http://www.nt.tuwien.ac.at/ltesimulator/
[10] H.Holma,A.Toskala 'LTE for UMTS OFDMA nd SC-FDMA based radio access"2009

# WORK IN PROGRESS:

# CONVERGED SERVICES

# A Single-Queue Priority Scheduler for Video Transmission in WLANs

Joshua Adeleke[1], Mqhele E. Dlodlo[1] and Clement Onime[2]
Department of Electrical Engineering,
University of Cape Town[1]
Tel: +27 216504853
and Abdus Salam International Centre for Theoretical Physics,
Trieste, Italy[2]
email: {joshua.adeleke, mqhele.dlodlo}@uct.ac.za[1]; onime@ictp.it[2]

**Abstract- Wireless networks are faced with an increasing demand for transporting and delivering time-sensitive video based resources to end-users or consumers. With the wide-spread use of wireless devices, delays and congestions are common and various research work have already been carried out to minimize delay and enhance data rate in wireless networks. In this paper, a scheduling algorithm is used to increase the quality of service (QoS) perceived by the end-user. The Single-Queue Priority Scheduling Algorithm (SQPS) is used at point of entry into a wireless network to reduce the loss rate and also service different priorities of video frames. The SQPS algorithm relies on metrics such as deadline, priority and cost to reduce the delay and increase the data rate for faster transmission of Motion Pictures Expert Group (MPEG) video frames in the Wireless Local Area Network (WLAN).**

**Index Terms—SQPS, WLAN, QoS**

## I. INTRODUCTION

The world we live in today is increasingly dependent on wireless communication technology. Wireless technology is used in the transmission of broadcast radio and television signals. Thanks to mobile telecommunications networks, wireless technology is used by everyone who owns a mobile phone. The Internet is a world wide global information network of connected computers. Wireless technology is also used for connecting computers together using the IEEE 802.11x technology also known as Wireless Local Area Network (WLAN).

Video is fast becoming the predominant service transported by the Internet as demonstrated by the popularity of various on-line video based resources such as YouTube and educational resources like iTunes University and massive open online courses (MOOC). The Cisco Visual Networking Index (VNI) reported that for the first time in 2012 that mobile video traffic was 51% of the total traffic and estimates that mobile video growth rate will account for 66% of the total mobile traffic by 2017[1]. The potential for real-time mobile video is huge provided it is offered seamlessly or delivered on-time to the end-users, that is with a good quality of service (QoS). Real-time video streaming is highly delay sensitive and QoS transmission related errors such as delays, packet loss and/or re-ordering can cause poor rendering of video as perceived by the end-user.

The Internet transmits data packets on a "best effort model" where all types of traffic are routed as soon as possible over the most cost effective combination of links. The best effort model assigns the same priority to all classes of traffic regardless of QoS requirements. The heterogeneity of the Internet backbone and unreliable end-to-end QoS support may be addressed using adequate scheduling algorithms. Scheduling of video packets before they are transmitted can increase the way the packets are prioritized and delivered to the receiver (consumers).

## II. BACKGROUND & RELATED WORK

The authors in [2] developed an algorithm that is able to allow arbitrary priority levels with a single queue for all priority requests whether low or high. The results obtained show that the algorithm peformed better than previous schemes in terms of disk utilization[2].

A selective scheme for sequence of frames transmitted over links with limited capacity is presented in [3]. They showed using dynamic programming that optimal schedulers can largely improve playback quality in video streaming[3].

Also, [4] developed a dynamic buffer management algorithm to cater for buffer underflow. The tradeoff between buffer and QoS

is studied using a dynamic programming framework[4].

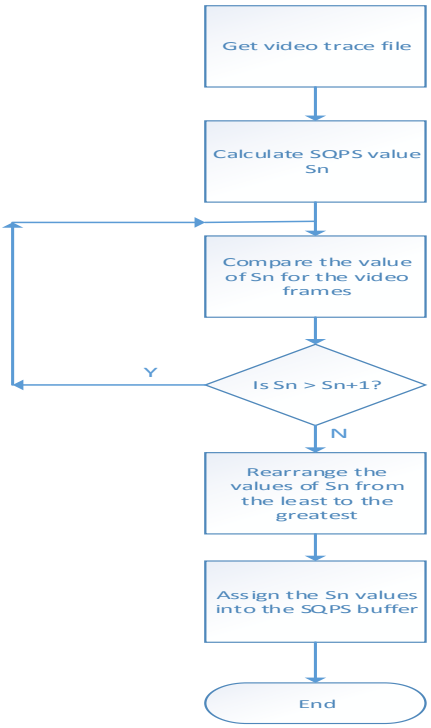



**Figure 1:** *A Flowchart of the Single Queue Priority Scheduler (SQPS)*

## III. SINGLE QUEUE PRIORITY SCHEDULING ALGORITHM (SQPS)

**Figure 1** shows a flowchart for the proposed implementation of SQPS algorithm, the major tasks to be performed by SQPS is to rearrange the video frames, compare the deadlines and decide on the video frames to transmit or drop for minimal loss and congestion avoidance in the buffers. The SQPS value is used to arrange the frames from the least to the greatest.

## IV. IMPLEMENTATION

A wireless network is simulated using the OPNET 14.5 simulator. The network has six wireless nodes (end-user terminals), two access points (AP). The APs are inter-connected through a network switch device, while each AP has 3 wireless nodes connected to it. Video frames are transmitted between wireless nodes across the APs and network switch. A simulation run with both APs under equal traffic load and nodes, (with monitoring for buffer loss and throughout at the APs) show that for 33Mb transmitted in 300s, buffer overflow events resulted in about 60% loss (see **Figure 2**). The SQPS algorithm is to be implemented as a component of the AP receiving the packets for transmission.

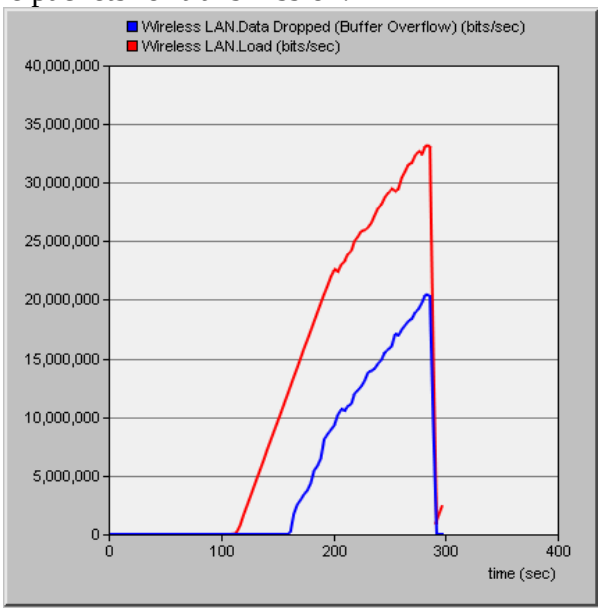



**Figure 2:** *Buffer Overflow vs Load*

## V. EXPECTED RESULTS

The results from the SQPS is expected to enhance the transmission of video frames with minimal loss in WLAN environments. This will reinforce the research efforts towards a seamless transmission of mobile video in wireless networks.

## VI. REFERENCES


[1] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2013–2018." [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html. [Accessed: 04-Jun-2014].

[2] S. Ghandeharizadeh, L. Huang, and I. Kamel, "A Cost Driven Disk Scheduling Algorithm for Multimedia Object Retrieval," vol. 5, no. 2, pp. 186–196, 2003.

[3] S. Mehdian and B. Liang, "Optimal Frame Transmission for Scalable Video with Hierarchical Prediction Structure," pp. 1–16, 2013.

[4] A. Dua and N. Bambos, "Buffer Management for Wireless Media Streaming," *IEEE GLOBECOM 2007-2007 IEEE Glob. Telecommun. Conf.*, pp. 5226–5230, Nov. 2007.



**Joshua Adeleke** received his undergraduate degree in 2004 from Babcock University, Nigeria and is presently studying towards his Masters degree in Electrical Engineering at the University of Cape Town. His research interests include wireless multimedia, wireless communication, e-health.

# A Connection Management System to Enable the Wireless Transmission of MIDI Messages

Brent Shaw and Richard Foss
Department of Computer Science
Rhodes University, Grahamstown, 6139
Tel: +27 46 603 8294

g09s2665@campus.ru.ac.za, R.Foss@ru.ac.za

***Abstract*—This paper proposes a wireless system for the distribution of MIDI messages for use in the live and studio environment. The system will make use of the MIDI and MIDINet protocols, creating wireless nodes that will enable the transmission of MIDI between devices with connection management capabilities through the use of mobile devices. The paper describes the configuration of the system, providing an architecture of the hardware and software designed for these wireless MIDI nodes.**

***Index Terms*—MIDI, wireless, embedded systems, web server.**

## I. Introduction

MIDI (Musical Instrument Digital Interface) is a protocol originally designed to enable the transmission of note messages between keyboard controllers and synthesisers [1]. Its use has been expanded and is now used in many other contexts for example control of mixing consoles and for show control. MIDI messages are transmitted over a cable that is limited in length to 15 meters [2].

There have previously been solutions that allow for MIDI messages to either be sent wirelessly [3] in a point-to-point manner and implementations that allow for MIDI to be sent over Ethernet networks [4], [5].

MIDINet is a protocol that enables network transmission of MIDI messages [6], [7]. The protocol allows for MIDI inputs or outputs on MIDINet nodes (currently workstations) to be connected so that MIDI messages may be routed over an Ethernet network.

This paper describes a system that utilises the MIDINet protocol to enable the one-to-many transmission of MIDI messages wirelessly and allows for connection management through the use of embedded web servers. It describes the configuration of such a system and the hardware and software design of nodes that enable the wireless routing of the MIDI messages. These nodes will be referred to as MIDI ConMan nodes.

## II. Proposed System Configuration

The system configuration in Figure 1 shows devices with MIDI connectivity such as a keyboard controller, synthesiser, digital mixing console and desktop computer, each connected to a MIDI ConMan device via a standard MIDI cable. The system will allow for MIDI messages to be routed between these devices over a wireless network.

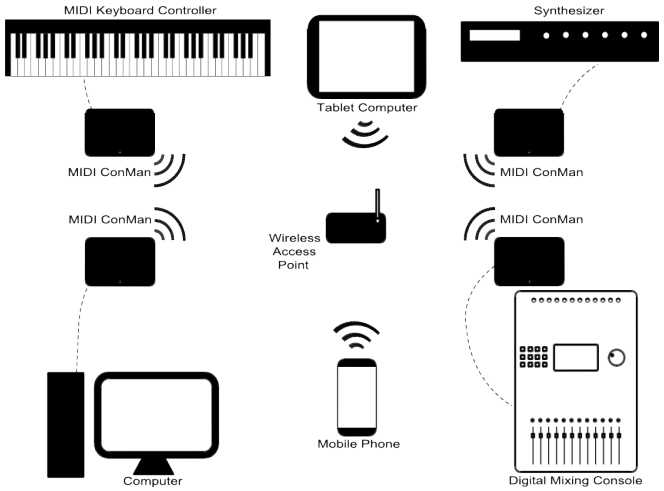



Fig. 1: Example configuration of MIDI ConMan system.

The connections in the system and the MIDI ConMan devices themselves will be managed through standard web browsers, using a computer, mobile phone or tablet.

As an example a MIDI keyboard controller connected to a MIDI ConMan device can be used to send MIDI messages to other MIDI controllable devices such as a remote synthesiser.

A computer running MIDI sequencing software would be able to send MIDI control messages to MIDI devices, allowing for automated control from digital audio workstation software over synthesisers, effect processors and digital mixing consoles.

## III. Hardware Specification

The MIDI ConMan hardware will need to feature a wireless networking module and a set of MIDI in and out ports. This project will make use of XMOS's multi-core microcontrollers and development kits [8]. Although there is no single development board currently available that provides the necessary connectivity, XMOS's SliceKIT provides a neat prototyping solution [9].

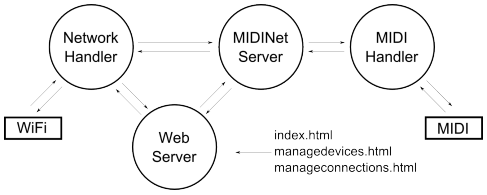



Fig. 2: High-level MIDI ConMan software components.

The XMOS SliceKIT features a 16 core microcontroller aimed at parallel configurations, mounted on a board that allows for the connection of 4 separate modules, termed 'slices' [9], [10]. The project will make use of the WiFi Slice and the Audio Slice, which will be attached to the SliceKIT [11].

## IV. Software Specification

The MIDI ConMan software will consist of 4 concurrent components: The network handler, MIDI handler, Web Server and the MIDINet server.

The MIDI handler provides a driver for the MIDI ports on the Audio Slice, while providing a set of functions that the MIDINet Server will use to send and receive local MIDI messages.

The Network handler sends and receives both TCP messages to the Web Server and UDP multicast messages for the MIDINet Server. It also handles the devices' wireless configuration and the assignment of IP addresses for the MIDI ConMan device.

The Web Server component provides an interface in the form of web pages that allow the user to manage connections and device settings for the MIDI ConMan devices remotely from any web browser.

The MIDINet Server allows for connections to be made between remote MIDI devices and for MIDI messages to be routed in a one-to-many fashion over the wireless network.

## V. Project Status

Thus far a prototype interface has been created for the connection management and device management web applications. A virtual machine running a web server was created and scripts were put in place so that the web application could be prototyped in an environment that closely resembles that of the proposed MIDI ConMan software specification. This has resulted in a web application that simulates the user experience that would be had while managing devices and connections within the MIDI ConMan system.

## VI. Conclusion

It is possible to route MIDI messages over standard Ethernet networks, allowing for virtual connections to be created so that a network of MIDI devices may be used together. Currently a distributed MIDI routing and management system requires a high overhead in terms of space and monetary cost.

Through the creation of smaller nodes, the current requirement of desktop computing devices to run the MIDINet protocol on an Ethernet network can be removed, enhancing the usability of the protocol by providing distributed connection management and MIDI routing over a wireless network.

The next phase in this research project includes the creation of a prototype system that will enable testing of the system on embedded hardware. The research will ultimately culminate in the creation of a single printed circuit board and associated firmware that will provide the necessary hardware features for the MIDI ConMan system.

## References

[1] R. A. Moog, "Midi: Musical instrument digital interface," *J. Audio Eng. Soc*, vol. 34, no. 5, pp. 394–404, 1986. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=5267

[2] MIDI Manufacturers Association. (1999/2008) Complete midi 1.0 detailed specification. [Online]. Available: http://www.midi.org/techspecs/gm.php

[3] Starr Labs. AirPower2 Wireless MIDI: Wireless MIDI to USB + MIDI Stand-alone Solution. [Online]. Available: http://www.starrlabs.com/index.php?route=product/product&product_id=79

[4] D. Schmitt. ipMIDI. [Online]. Available: http://www.nerds.de/en/ipmidi.html

[5] T. Erichsen. rtpmidi. [Online]. Available: http://www.tobias-erichsen.de/software/rtpmidi.html

[6] R. Foss and T. Mosala, "Routing midi messages over ethernet," *Journal of the Audio Engineering Society*, vol. 44, no. 5, pp. 406–415, 1996. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=7898

[7] R. Foss, "Object oriented design part 3: Midinet complete design and implementation," 2013, Rhodes University Computer Science III course notes.

[8] XMOS. xcore multicore microcontrollers. XMOS. [Online]. Available: https://www.xmos.com/download/public/xCORE-Multicore-Microcontrollers-Overview(1.2).pdf

[9] ——. Slicekit hardware manual. XMOS. [Online]. Available: https://www.xmos.com/download/public/sliceKIT-Hardware-Manual(1.0).pdf

[10] ——. Xs1-l16a-128-qf124 datasheet. XMOS. [Online]. Available: https://www.xmos.com/download/public/XS1-L16A-128-QF124-Datasheet(X8006A).pdf

[11] ——. slicekit selector guide. XMOS. [Online]. Available: https://www.xmos.com/download/public/sliceKIT-Selector-Sheet(X6225D).pdf

**Brent Shaw** Brent Shaw is a masters student studying at Rhodes University[1]. His research is in the field of audio networking, with a focus on remote control over audio devices.

**Richard Foss** Richard Foss is an associate professor in Rhodes University Department of Computer Science. He leads a research group in the field of audio networking.

[1]This work was undertaken in the Distributed Multimedia CoE at Rhodes University, with financial support from Telkom SA, Tellabs, Genband, Easttel, Bright Ideas 39, THRIP and NRF SA (TP13070820716). The authors acknowledge that opinions, findings and conclusions or recommendations expressed here are those of the author(s) and that none of the above mentioned sponsors accept liability whatsoever in this regard.

# Real-time Background Subtraction Under Sudden Illumination Changes

C.J.F. Reyneke, P.E. Robinson and A.L. Nel
HyperVision Research Laboratory
Department of Electrical and Electronic Engineering Science
Faculty of Engineering and the Built Environment
University of Johannesburg
P.O. Box 524, Auckland Park, Johannesburg, 2006
Tel/Fax: +27 (0)11-559-2147
email: corius.reyneke@gmail.com; {philipr, andren}@uj.ac.za

***Abstract*—This paper investigates three background modelling techniques that are robust against sudden and gradual illumination changes for a single, stationary camera. The first makes use of a novel feature extraction operator that considers both spatial texture and colour information.The second uses a combination of a frame-based Gaussianity Test and a pixel-based Shading Model to handle sudden illumination changes. The third solution implements non-parametric Kernel Density Estimation (KDE) and estimation of global intensity level changes by mean-shift.**

**Our aim is to eventually identify the best-performing solution, improve upon it, and implement it on a GPU for real-time application.**

***Index Terms*—background subtraction, sudden illumination changes, real-time.**

## I. Introduction

Background subtraction techniques have traditionally been applied to object detection in computer vision systems and have since become a fundamental component for many applications ranging from human pose estimation to video surveillance. The goal is to remove the background in a scene so that only the interesting objects remain for further analysis or tracking. Techniques such as these are especially useful when they can identify object regions without prior information and when they can perform in real-time.

Real-life scenes often contain dynamic backgrounds such as swaying trees, rippling water, illumination changes and noise. While a number of techniques are effective at handling these, sudden illumination changes such as a light source switching on/off or curtains opening/closing continue to be a challenging problem for background subtraction[1]. In recent years a number of new segmentation techniques have been developed that are robust to sudden illumination changes but only for certain scenes. Our aim is to eventually identify the best-performing solution, improve upon it, and implement it on a GPU for real-time application.

## II. Related Work

A number of texture-based methods have developed to solve the problem of sudden illumination changes. Heikkila *et al.*, Xie *et al.* and Pilet *et al.* make use of robust texture features [3]. Heikkila *et al.* make use of local binary pattern histograms as background statistics. Xie *et al.* assumes that pixel order values in local neighbourhoods are preserved in the presence of sudden illumination changes. They provide an output image by classifying each pixel by its probability of order consistency. Pilet *et al.* make use of texture and colour ratios to model the background and segment the foreground using an expectation-maximization framework. Texture-based features work well, but only in scenes with sufficient texture; untextured objects prove to be a difficulty.

Another way of dealing with sudden illumination changes is to maintain a representative set of background models [3]. These record the appearance of the background under different lighting conditions and alternate between these models depending on observation. The techniques that make use of this approach mostly differ in their method of deciding which model should be used for the current observation. Brumitt *et al.*, implement the Wallflower system which chooses the model as the one that produces the lowest number of foreground pixels. This proves to be an unreliable criterion for real-world scenes. Stenger *et al.* make use of hidden Markov models but in most cases, sharp changes occur without any discernible pattern. Also, Stenger *et al.* and Brumitt *et al.* require off-line training procedures and consequently cannot incorporate new real-world scenes into their models during run-time [4]. Sun *et al.* implement a hierarchical GMM in a top-down pyramid structure [3]. At each scale-level a mean pixel intensity is extracted and is matched to the best model of its upper-level GMM. While mean pixel intensity is useful for the detection of illumination changes, it is also sensitive to changes caused by the foreground. Additionally, the Hierarchical GMM does not exploit any spatial relationships among pixels which can output incoherent segmentation. Dong *et al.* employ principle component analysis to build a number subspaces where each represent a single background appearance. The foreground is segmented by selecting the subspaces which produces minimum reconstruction error. However, their work does not discuss how the system reacts to repetitive background movements.

More recently, Zhou *et al.* [5], Ng *et al.* [1] and Hwang *et al.* [6] have developed techniques that claim to be robust to sudden illumination changes. These will be discussed in more detail.

## III. Proposed Solutions

### A. Background Modeling using Spatial-Colour Binary Patterns (SCBP)

This approach makes used of a novel feature extraction operator, the Spatial-Colour Binary Pattern (SCBP), which takes spatial texture and colour information into consideration. It is an extension of a local binary pattern which is adapted to be centre-symmetrical and to consider only two colour channels for the sake of computational efficiency.

A histogram of the SCBP feature is extracted and a model, consisting of several histograms, is built for each pixel. For each new frame, a pixel is labeled as foreground or background based on a proximity measure between its SCBP histogram and its model, which is then updated. Furthermore, object contours are refined using a statistical operator to reduce false positives. Adaptive thresholds are implemented to improve both the tolerance of dynamic regions and the sensitivity of static regions [5].

### B. Background Modeling using a Shading Model and a Gaussianity Test

This approach implements a hierarchical framework that uses a combination of a pixel-based Shading Model and a block-based Gaussianity Test.

The Gaussianity Test is based upon the assumption that camera noise is spatially Gaussian and temporally uncorrelated. Therefore, when a difference frame is calculated, only foreground objects and Gaussian noise remains. The test is implemented to distinguish between these two types of pixels [1].

However, the assumption that background regions are Gaussian distributed does not hold true in the presence of sudden illumination changes because there is photometric distortion. A Shading Model is combined with the Gaussianity Test in order to compensate for this [1].

The shading model assumes that if there is no physical change between two frames, such as a moving object, then the ratio of pixel intensities will be constant and independent of the shading coefficients of the frames. When there are no foreground objects in the scene the ratio of pixel intensities will be Gaussian distributed. Now, the Gaussianity Test can be implemented to distinguish between foreground and background pixels in the presence of sudden illumination changes [1].

### C. Background Modeling using Non-parametric Kernel Density Estimation

This solution implements non-parametric Kernel Density Estimation (KDE) and estimation of global intensity level changes by mean-shift. Instead of modeling the background using a several Gaussian distributions as is typical of Mixture-of-Gaussian methods, the technique is non-parametric. This means that each single sample of the $N$ samples considered is a Gaussian distribution. This allows the density function to be estimated more accurately by only depending on recent information from the sequence [6].

Sudden illumination changes are overcome using a mean-shift algorithm that considers intensity level changes as a ratio between two successive frames. A problem that may arise from this approach is the misclassification of saturated areas caused by illumination effects or a wide aperture. The kernel function is modified to handle this[6].

## IV. Conclusion & Future Work

Our aim is to eventually identify the best-performing solution, improve upon it, and implement it on a GPU for real-time application.

The SCBP histogram feature approach distinguishes itself from other texture-based methods because it also considers colour information. This is promising for scenes that contain untextured objects which would otherwise be problematic. It is possible to improve upon this method by optimizing the block size of the local SCBP histograms.

The Shading Model and Gaussianity Test approach is a novel method for handling sudden illumination changes. It is promising because it makes use of both pixel and block-based processing. It is possible to improve upon this method by optimizing the block size of the Gaussianity Test.

By compensating for sudden illumination changes these methods have become too computationally expensive for real-time application. We hope to implement one of these techniques on a GPU in order to overcome this.

## References


[1] K. K. Ng, S. Srivastava, and E. Delp, "Foreground segmentation with sudden illumination changes using a shading model and a gaussianity test," in *Image and Signal Processing and Analysis (ISPA), 2011 7th International Symposium on*, September 2011, pp. 236–240.

[2] M. Heikkila and M. Petikainen, "A texture-based method for modelling the background and detecting moving objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 657–662, 2006.

[3] J. Li and Z. Miao, "Foreground segmentation for dynamic scenes with sudden illumination changes," *Image Processing, IET*, vol. 6, no. 5, pp. 606–615, July 2012.

[4] X. Zhao, W. He, S. Luo, and L. Zhang, "Mrf-based adaptive approach for foreground segmentation under sudden illumination change," in *Information, Communications Signal Processing, 2007 6th International Conference on*, Dec 2007, pp. 1–4.

[5] W. Zhou, Y. Liu, W. Zhang, L. Zhuang, and N. Yu, "Dynamic background subtraction using spatial-color binary patterns," in *Image and Graphics (ICIG), 2011 Sixth International Conference on*, Aug 2011, pp. 314–319.

[6] Y. Hwang, K. Sung, J. Chae, Y. Park, and I.-S. Kweon, "Robust background maintenance by estimating global intensity level changes for dynamic scenes," *Intelligent Service Robotics*, vol. 2, no. 3, pp. 187–194, 2009.



**Cornelius Reyneke** received his undergraduate B.Eng in Electrical and Electronic Engineering with endorsement Information Technology in 2012 from the University of Johannesburg. He is presently studying towards his Masters at the same institution. His research interests include Depth-from-Focus, Background Subtraction and GPU implementation.

# Autonomous Facial Expression Recognition Using the Facial Action Coding System

Nathan de la Cruz,[1] Mehrdad Ghaziasgar[1], James Connan[2] and Reg Dodds[1]
[1]Department of Computer Science
University of the Western Cape, Private Bag X17, 7535 Bellville, South Africa
Telephone: +(27) 21 959–3010, Fax: +(27) 21 959–3006
and [2]Department of Computer Science
Rhodes University, P O Box 94, Grahamstown, 6140
Tel: + (27) 46 603-8291, Fax: + (27) 46 636-1915
Email: [1]3033421@myuwc.ac.za; [1]mghaziasgar@uwc.ac.za; [2]j.connan@ru.ac.za; [1]rdodds@uwc.ac.za

***Abstract*—Automated facial expression recognition is a well-researched field. The majority of research in this field focuses on recognizing individual Facial Action Coding System action units or whole-facial expressions. This research aims to develop an autonomous action unit recognition system that additionally uses action units to recognize six prototypical facial expressions, namely, Happy, Sad, Angry, Disgust, Fear and Surprise, as well as the neutral expression.**

***Index Terms*—Facial Action Coding System, Action Units, Dense Optical Flow, Facial Expression, Recognition**

## I. Introduction

Facial expressions are universally similar: facial expressions are innate human similarities that can be analyzed and recognized. This idea is supported by and prevalent in the research of Ekman and Friesen which models muscle movements in the face and uses these movements as the atomic units of all facial expressions. The model is called the Facial Action Coding System (FACS) and individual distinct muscle behaviour is called an action unit (AUs) [1].

The FACS defines 44 unique AUs. Of these, 30 AUs are related to the contraction of muscles in the face, of which 12 muscles are located in the upper face and 18 are located in the lower face. These AUs can occur both alone and in combination with other AUs. Over 7000 distinct AU combinations have been observed [2].

A large body of research has focused on recognizing AUs automatically using Computer Vision. The motivations for and applications of such systems are varied and include deception detection [3], [4], emotion recognition [5], and sign language recognition [6]. The majority of such systems fall in two classes. The first class of systems aims to recognize individual AUs, without collating the recognized AUs to infer whole-facial expressions such as the six prototypical expressions: Happy, Sad, Angry, Disgust, Fear and Surprise. The second class of systems recognizes groups of AUs as single units, to infer limited sets of whole-facial expressions, without recognizing the constituent AUs.

This research proposes an autonomous AU recognition system, and additionally proposes first to recognize individual AUs, and subsequently combine these units towards recognizing whole facial expressions, in this case the six prototypical expressions, as well as the neutral expression. This is carried out in the context of the research at the SASL research group at the University of the Western Cape towards creating an automatic machine-translation system that can translate fully-fledged South African Sign Language (SASL) phrases into English and vice versa. A major component of the system is the ability to recognize facial expressions in SASL videos [6].

The proposed method for autonomously tracking and recognizing AUs is similar to the combination of dense optical flow and Hidden Markov Models (HMMs) proposed by Lien *et al.* [7]. Our research additionally focuses on recognizing whole-facial expressions.

The rest of the paper is organised as follows: Section II discusses the related work; Section III discusses the research methodology; the paper is concluded in Section IV.

## II. Related work

Two studies are discussed below. The first, by Lien *et al.* [7], investigates strategies to recognize individual AUs. The second study, by Kapoor *et al.* [8], attempts to recognize groups of AUs. The scope of a work-in-progress paper limits this review; for a wider coverage refer to [9].

Lien *et al.* compared four AU recognition strategies, namely: dense flow extraction with Hidden Markov Models (HMMs); facial-feature tracking with discriminant analysis; facial-feature tracking with HMMs; and high-gradient component detection with HMMs. The comparison aimed to determine the strategy that achieves the highest AU recognition accuracy. Three AUs occurring in the brow region, three in the eye region and six in the mouth region were considered and used to compare the recognition rates of the strategies.

Videos of 100 adults performing a series of facial expressions starting from the neutral expression were recorded and used as the data set. Subjects sat directly in front of a camera and performed a series of facial expressions, as instructed by the researcher. The results revealed that the two best performing strategies in the brow region were dense

flow extraction with HMMs and facial feature tracking with discriminative analysis, with recognition accuracies of 92% and 91% respectively.

The two best performing strategies in the mouth region were dense flow extraction with HMMs and facial-feature tracking with HMMs, with recognition accuracies of 92% and 88% respectively. It is clear that dense flow extraction with HMMs had the best overall AU recognition accuracy. As such, it is selected as the AU recognition strategy in our research.

Kapoor *et al.* used an infrared camera with IR LEDs to capture videos [8]. The pupils were tracked using the red eye effect [10]. Two templates consisting of eight points plotted around the eyes and three points plotted on the eye brows are super-imposed onto the facial images, aligning the centre of the templates with the centre of the tracked pupils. The points on the templates are used to track the facial skin in order to track select regions of interest that contain AUs of interest.

Support Vector Machines (SVMs) were used to recognize 10 whole-facial expressions which were pre-defined combinations of six AUs around the eye and brow regions. Some combinations consisted of only a single AU.

An independently sourced database was used to evaluate the system. The database consisted of videos of 8 children in a real-life learning situation. The children were asked to play a game called FRIPPLE PLACE [11]. The game consists of a variety of puzzles that require mathematical reasoning. Each child worked on the puzzles for twenty minutes, and two cameras recorded the facial expressions of the child during this time. A trained FACS expert labelled the videos, indicating the AUs that were present in each frame. Eighty (80) frames were manually chosen to test the system.

The system achieved an average 61.25% accuracy in recognizing the 10 expression groups.

## III. RESEARCH METHODOLOGY

Our research methodology has three aspects: *first*, image pre-processing techniques will be applied to segment and normalize the face in an image of a subject. The method proposed by Mushfieldt *et al.* [6] has been shown to be particularly effective in this regard. *Secondly*, dense flow extraction will be used to track facial features and HMMs will be used to recognize a set of AUs given those features and *thirdly*, two classification techniques, HMMs and SVMs, will be used to recognize the six prototypical expressions, as well as the neutral expression. The two methods will be compared. The following subsections describe the use of each technique.

### A. Hidden Markov Models

HMMs predict the probability that an input sequence matches a set of states, based on a pre-defined state model. The input consists of the AUs observed over a series of frames, and the output is a series of probabilities that the input AUs match each pre-trained whole-expression. The whole-expression with the highest probability is assumed to be the observed one. This method takes temporal information into account, but can be complex to model and train.

### B. Support Vector Machines

SVMs are binary classifiers that determine one of two classes to which an input data point belongs, based on a pre-trained data model. One of several multi-class decision strategies can be used to cater for a number of classes. Similar to Kapoor *et al.*, the input consists of the AUs observed in a single frame, and the output is the whole-expression that the set of AUs may most likely represent in the context of this research. This method does not take temporal information into account, but it is a simpler model and it is easier to train. A stabilization strategy, such as the one proposed by Li *et al.* [12], can be used to prevent continuous changes to the recognized expression.

## IV. CONCLUSION

This paper proposes a facial expression recognition system that first recognizes individual AUs and subsequently groups these units to recognize whole-facial expressions. A combination of dense flow extraction and HMMs was shown to recognize accurately AUs. Two techniques—HMMs and SVMs—are used to recognize whole-facial expressions.

## REFERENCES


[1] P. Ekman and W. Friesen, *The Facial Action Coding System: A Technique for the Measurement of Facial Movement*. San Francisco: Consulting Psychologists Press, 1978.

[2] K. Scherer and P. Ekman, *Handbook of Methods in Nonverbal Behavior Research*. Cambridge University Press, 1982.

[3] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikainen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–6.

[4] T. Pfister, X. Li, G. Zhao, and M. Pietikainen, "Recognising spontaneous facial micro-expressions," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1449–1456.

[5] I. A. Essa and A. P. Pentland, "Coding, analysis, interpretation, and recognition of facial expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 7, pp. 757–763, 1997.

[6] D. Mushfieldt, M. Ghaziasgar, and J. Connan, "Robust facial expression recognition in the presence of rotation and partial occlusion," in *Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference*. ACM, 2013, pp. 186–193.

[7] J. Lien, T. Kanade, J. Cohn, and C. Li, "Detection, tracking, and classification of action units in facicial expression," *Journal of Robotics and Autonomous Systems*, vol. 31, no. 3, pp. 131–146, 2000.

[8] A. Kapoor, "Automatic facial action analysis," Master's thesis, Massachusetts Institute of Technology, Cambridge., 2002.

[9] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern Recognition*, vol. 36, no. 1, pp. 259–275, 2003.

[10] J. Van de Kraats and D. van Norren, "Directional and nondirectional spectral reflection from the human fovea," *Journal of Biomedical Optics*, vol. 13, no. 2, pp. 1–13, 2008.

[11] Edmark, "Fripple place," 1991, http://www.riverdeep.net/edconnect/softwareactivities/criticalthinking/fripple place.jhtml.

[12] P. Li, M. Ghaziasgar, and J. Connan, "Hand shape estimation by 2D appearance and 3D animation for sign-language," in *Proceedings South African Telecommunication Networks and Applications Conference (SATNAC 2011)*, East London, Eastern Cape, South Africa, 2011, pp. 409–414.


**Nathan de la Cruz** is currently a M.Sc. student at the University of the Western Cape. His interests are image processing, Internet programming and machine learning.

# CUDA Optimized dynamic programming search for automatic speech recognition on a GPU platform

Babedi B. Letswamotse[1], Naison Gasela[1] and Zenzo P. Ncube[2]
Department of Computer Science
North-West University, Mafikeng Campus[1], Private Bag X2046, Mmabatho 2745
Tel: +27 18 3892811, Fax: +27 18 3929775
and Department of Information and Communication Technology
Sol Plaatjie University[2], Private Bag X5008, Kimberley 8300
email: {18040969, Naison.Gasela}@nwu.ac.za[1]; Zenzo.Ncube@spu.ac.za[2]

**Abstract- In a typical recognition process, there are substantial parallelization challenges in concurrently assessing thousands of alternative interpretations of a speech utterance to find the most probable interpretation. During this process, input signals are converted into feature vectors, thus decoding these feature vectors to produce relevant output is a computationally expensive task. Many time critical applications are unable to use Automatic Speech Recognition (ASR) due to the heavy latency in processing the speech with a large vocabulary size. This work proposes to optimize the performance of dynamic programming search. Optimizing dynamic programming search requires a certain level of parallelism since search is a parallel process. We find that a better way to optimize speech recognition search is by the use of parallel architectures such as graphic processing units (GPUs). GPUs provide large computational power at a very low expense which positions them as global accelerators. These savings encourage using GPUs as hardware accelerators to support computationally intensive applications.**

**Index Terms— GPU, Decoding, Dynamic Programming, Automatic Speech Recognition, CUDA**

## I. INTRODUCTION

Speech is the most effective form of communication for human to human interactions, so people expect the same when it comes to human-machine (computer) interactions. They expect the speech recognition systems in which the computer speaks and recognizes any human language. For these expectations to be met speech recognition has to be put into practice. Speech Recognition is the process of recognizing spoken input and converting it into written text through a speech recognition system.

Speech recognition systems are also used in a wide range of applications such as air traffic control, embedded telecommunication systems, robotics and, computer and video games. Many systems in the real world make use of speech recognition, these kinds of systems assist in so many ways. They assist with online shopping and voice activated passwords, security systems etc. Optimizing the performance of automatic speech recognition systems will help improve the services provided by these systems both commercially, socially and educationally.

Most of the modern speech recognition systems are usually based on statistic models such as Hidden Markov Models (HMMs). According to [1], HMMs are popular due to their simplicity and computational feasibility, and their parameters that can be estimated automatically from a large amount of data.

Speech recognition is divided into three phases: feature extraction, classification and decoding. The classification stage is a collection of segmented words and sub-words into different classes based on some properties [2]. Classification is composed of acoustic models which are files that are generated by taking audio recordings of speech and their transcriptions and then compile them into statistical representations of the words sounds. Each of these statistical representations is assigned a label called a phoneme [3], pronunciation dictionary which is a machine-readable dictionary that contains a collection of words and their transcriptions and the language model which is a probability distribution P(s) over words S that attempts to reveal how frequently a string S occurs as a sentence. Language models are often used for dictation applications. In any speech recognition systems there are two vital metrics to contemplate: the elapsed time between the acquisition of the speech signal and the recognized word, and the accuracy.

In a typical recognition process, there are substantial parallelization challenges in concurrently assessing thousands of alternative interpretations of a speech utterance to find the most probable interpretation. Most time critical applications are unable to use Automatic Speech Recognition (ASR) due to the heavy latency in processing the speech with a large vocabulary size.

In this research the authors focus on the decoding process of speech recognition. The decoding process (which is often referred to as search) in a speech recognizer's operation is to find a sequence of words whose corresponding acoustic and language models best match the input feature vector sequence [1]. Search is a computationally expensive task since it handles irregular graph structures with data parallel operations. There are algorithms that were developed specifically for this task but many search algorithms were developed prior to the existence of parallelism. Dynamic programming has a complexity of O ($n^2$), which can cause unreasonable demands on both the processing time and system memory. Parallelizing the Dynamic Programming search will help improve the efficiency of the recognition process hence reducing the latency in processing speech for large vocabulary systems.

## II. RELATED WORK/BACKGROUND

Rehman et al. [4] implemented a dynamic programming algorithm (Viterbi) on NVidia graphics processing unit using CUDA and concluded that it has been accelerated from 3 to 6 times as compared to the serial execution on

central processing unit.

Hachkar et al. [5] used two algorithms to implement a system of Automatic Recognition of isolated Arabic Digits: Dynamic Time Warping (DTW) and Discrete Hidden Markov Model (DHMM). DTW-based system recognition leads to recognition accuracy of 77%. The better recognition accuracy of about 92% was obtained with DHMM-based system. They found that the recognition performances for the two ASR systems are worse in noisy environment, but the pattern recognition using HMM is better than the pattern using DTW.

Wei and Weisheng [6] attempted on improving the recognition efficiency without compromising the recognition accuracy. They analysed the traditional Viterbi-Beam search algorithm and proposed an improved adaptive Viterbi- Beam search algorithm by analysing the voice activity model of different stages. The method combining Viterbi algorithm with Beam pruning technique is useful to compress the search space, which reduces the computational complexity. The experimental results show that the search space is compressed effectively without affecting the recognition accuracy and an improvement on search efficiency of 35.77% is observed.

## III. RESEARCH METHODOLOGY

This work follows three processes which form the overall research methodology. The first being the dynamic programming search method survey, which will be done to look at what recent speech processing improvements have been achieved using this approach particularly on CUDA GPUs. Secondly is the experimental setup, which will be focusing on the actual experimental tools and how they will be setup. Thirdly as proof of concept, performance comparisons between the then CUDA GPU optimized dynamic programming search algorithm and the original algorithm will be done.

HTK assimilated CMU sphinx 4 recognizer will be used for recognition on both the Linux based workstation with Intel core i7-3770 CPU and the GPU based system running on GTX 8800. The Hidden Markov Models will be connected together in a sequence to enable continuous speech. A speech recognition Viterbi search algorithm will be implemented on both systems. The Viterbi search algorithm will be optimised by loop unrolling to improve the optimality of the search process and thus improving the efficiency of the recognition process.

CUDA will be used to implement the optimised version of the speech recognition Viterbi search algorithm on the GPU based system using the already existing language corpus. We will make use of the shared memory only for communication intensive processes due to the memory limitations associated with shared memory. For this implementation we will synchronize the threads to ensure that the parallel threads cooperate in order to yield correct results and avoid deadlocks. We will use a barrier synchronization primitive called _syncthreads ( ) which is provided by CUDA. Performance results of the implementations will also be thoroughly analyzed and evaluated. Tools to be used include the following:

- HTK version 3.4.1
- CMU (Carnegie Mellon University) Sphinx 4
- GTX 8800 GPU
- NVidia CUDA 5.5 Toolkit
- Linux based workstation with Intel core i7-3370 CPU only
- CMU US BDL Arctic 0, 95 Speech corpus

## IV. CONCLUSIONS

The aim of this research is to optimize the performance of the Speech Recognition Dynamic Programming search, therefore we are going to implement a dynamic programming based search algorithm (Viterbi) and optimize it using CUDA.

## REFERENCES


[1] N. Indurkhya and F. J. Damerau, "An Overview of Modern Speech Recognition, "in *Handbook of natural language processing*. London: CRC Press, 2009, Ch. 15, pp. 339-366.

[2] M. Rahman,F. Khan, and A. Bhuiyan, "Continuous Bangla speech segmentation, classification and feature extraction., "*International Journal of Computer Science Issues, (IJCSI)*, vol. 9, no. 2, pp. 67-75, March (2012).

[3] Acoustic Modelling. Microsoft Research. [Online], http://research.microsoft.com/en-us/projects/acoustic-modeling/ (Accessed: 8 March 2014).

[4] M. K. Rehman, M. U. Sarwar, M. R. Talib, M. S. Mansoor and M.B. Sarwar, "Parallel Implementation of Dynamic Programming Algorithm Using Graphics Processing Unit," International Journal of Computer Science and Management Research, vol. 2, no. 4, pp. 2097-2107, April 2013.

[5] Z.Hachkar, A. Farchi , B.Mounir and J. EL Abbadi "A Comparison of DHMM and DTW for Isolated Digits Recognition System of Arabic Language," International Journal on Computer Science and Engineering ,(IJCSE), vol. 3, no. 3, pp. 1002- 1008, March 2011.

[6] H. Weisheng and L. Wei, "Improved Viterbi Algorithm in Continuous Speech Recognition ," in International Conference on Computer Application and System Modeling (ICCASM 2010), Taiyuan, 2010, pp. v7-207- v7-209.



**Babedi Betty Letswamotse** received her undergraduate degree (Computer Science and Mathematics) in 2011 and her Honours degree (Computer Science) in 2012 from the North West University (Mafikeng Campus) and is presently studying towards her Master of Science degree in Computer Science at the same institution. Her research interests include GPU general purpose computing and automatic speech recognition.

# WORK IN PROGRESS:

# CORE NETWORK TECHNOLOGIES

# Towards Investigating Transmission Penalties in a Flexible Spectrum Optical Network

D. Kiboi Boiyo, E.K. Rotich Kipnoo, R.R.G. Gamatham, A.W.R. Leitch and T.B. Gibbon
Physics Department, Nelson Mandela Metropolitan University, P.O. Box 77000, Port Elizabeth 6031, South Africa. Tel: +27 41-504-2141; Fax: +27 41-504-2573.
Email: Duncan.Boiyo@nmmu.ac.za

Abstract- we investigate a multichannel transmission over elastic spectrum for efficient and scalable signal transmission in high capacity telecommunication systems. The flexible spectrum is sliced into multiple slots containing multiple channels but incur crosstalk penalties. We provide simulated analysis of the transmission penalty resulting from nonlinear effects for two closely spaced channels, a scenario that could occur in a flexible spectrum network. The penalty is investigated at 20 Gb/s bit rate as a function of both channel spacing and transmission power and is found to be within 9.5-0.9 dB for 28-50 GHz spacing respectively for a 10 dBm transmitter. This work is of interest to future network applications within South Africa such as the SKA Project which requires efficient, cost effective and high network data transmission technologies.

Index Terms— Flexible Spectrum, ROADMs, SKA, Optical fibre network

## I. INTRODUCTION

Transport networks are experiencing high bandwidth demand due to high definition video distribution services and requirements for faster broadband by mobile and fixed terminals. This demand has drastically increased the network traffic. Flexible spectra provide effective and efficient solution due to the sliced wide bandwidth [1].

Flexible spectra provide dynamic adjustment of the wavelengths, transmission formats and data rates by employing Reconfigurable Optical Add/Drop Multiplexers (ROADMs) and switches which elastically adjust to the demand of data while maintaining the quality of the transmission and service. In slicing the spectrum for wavelength assignment, effective and low penalty channel spacing is selected to reduce the effects of interchannel interactions like crosstalk [2].

For South Africa, a future big data driver is the Square Kilometre Array (SKA) Project, which is destined to build the world's largest radio telescope. The square kilometre area coverage for data collection will connect its dishes and aperture arrays by optical fibre, which will offer 110 times the current global internet traffic [3]. The large volume of data will also be shared between their central office and worldwide consumers, relying on technologies such as flexible networks.

The proposed network investigated in this paper is a 20 Gb/s transmission scheme over a 12 km length of optical fibre. In this work, we theoretically demonstrate a sliced bandwidth with channel spacing of 28-50 GHz accommodating multiple channels.

## II. THEORY

In wavelength division multiplexed (WDM) systems, several optical frequencies are coupled onto a single fibre transmission span. Depending on how these channels are spaced, there arises coupling and crosstalk between the adjacent multiplexed signal channels. Crosstalk is as a result of nonlinear interaction in the fibre that limits the performance of the WDM system [4]. Such interchannel interference reduces the sensitivity of the detectors.

The nature of the signal degradation is quantified as a power penalty (in decibel, dB) at a fixed channel spacing is given by;

$$C = 10\log_{10}(I / MDP) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

where I= Power of the interfered channel

MDP= Minimum Detectable Power for a bit error rate in the absence of crosstalk (Back-to-Back) [5].

## III. RESEARCH DESIGN

Figure 1 shows the simulation of a two channel transmission using VPI transmission maker software. A 1552.52 nm source modulated using a 20 Gb/s Non-Return-to-Zero (NRZ)-Pseudo- Random Binary Sequence (PRBS) is coupled into the fibre simultaneously with a similar source of varied wavelength and power. The signals are transmitted over a 12 km G.655 Non-Zero Dispersion Shifted Fibre. The power and spacing of the interfering channel are then varied from 0-25 dBm and 28-50 GHz respectively, as the power penalties are measured at$10^{-9}$ BER.

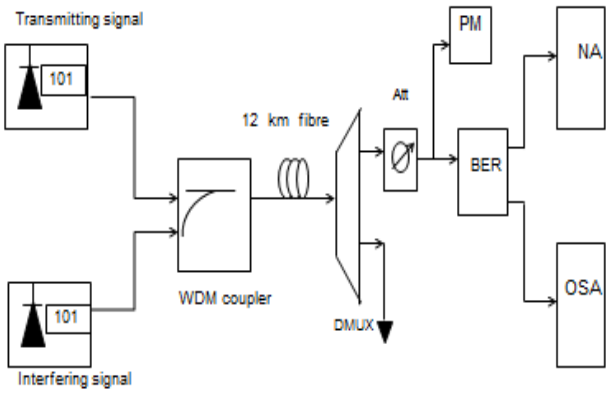



Figure 1 Two channel simulation set up

## IV. RESULTS AND DISCUSSION

In figure 2, both the signal and interfering channels were powered at 10 dBm but with varying channel spacing between them. Figure 2 shows a back to back (BB) received signal without interference (with interfering channel OFF). For 50 GHz, the crosstalk penalty was 0.9dB while 28 GHz had a 9.5 dB penalty. Further reduction in the spacing caused more errors and an error floor representing a highly distorted unrecovered signal.

In fig. 3, the spacing of the signal and interfering channel is kept constant at 37.5 GHz and power of the signal is 10 dBm, while varying the power of the interfering channel from 0 to 25 dBm. The increased interference from the high power interfering channels results in too many bit errors and therefore increased BER values in the transmitted channel.

Further increase in the interfering channel power results in an error floor. The error floor is as a result of receiver overload since receiver distortion overrides BER error mechanism of the receiver.

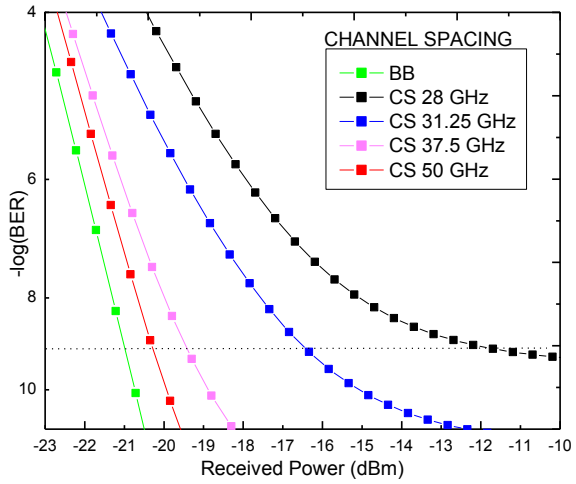



Figure 2 BER against received power for different channel spacing (28-50 GHz) at fixed channel powers.

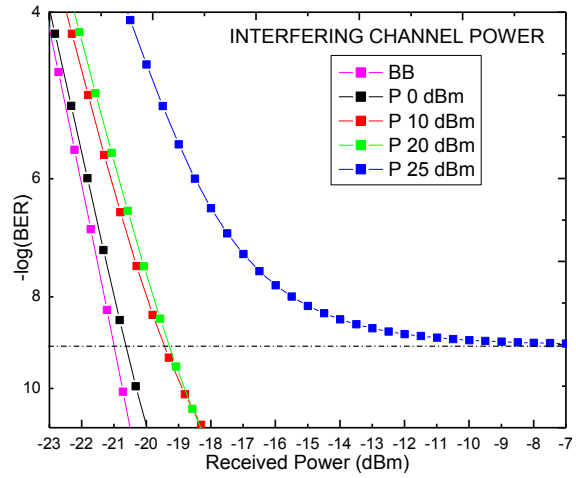



Figure 3 the effect of increased received signal power level on the bit errors at fixed channel spacing

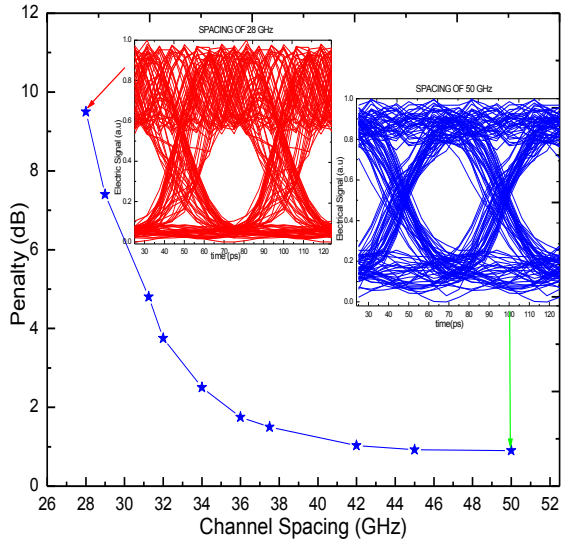



Figure 4 the penalty incurred due to increased channel spacing between channels

The penalties for spacing 28-50 GHz are shown in fig 4. Penalties reduce with increase in the channel spacing. The inset of fig 4 shows the eye diagram patterns for spacing 28 and 50 GHz. The 28 GHz spacing shows significant eye closure due to crosstalk, while 50 GHz has a more open eye implying less signal distortion.

## V. CONCLUSION

In spectrum assignment, high efficiency is realized when more channels are slotted in the spectrum. If the channel spacing is too close, then crosstalk penalties are incurred. The simulation results have shown that the effective channel spacing with low transmission penalty can be used to accommodate more channels. Therefore, with a bandwidth of 400 GHz and a channel spacing of 31.25 GHz, 12 channels can be assigned to the bandwidth but with a penalty of 4.8 dB. The outcome of this work has important implications for multichannel, single optical fibre transmission applications such as the SKA Project and broadband networks with high spectral usage and effective network performance at a low cost.

## VI. ACKNOWLEDGEMENT

We are grateful for Research Funding and support from: Telkom, Dartcom, Infinera, Ingoma Communication Services, NLC, NRF, THRIP, SKA-SA, Cisco, CSIR and the ALC.

## REFERENCES

1. Miroslaw Klinkowski & Kryysztof Walkowiak, “Routing and Spectrum Assignment in Spectrum Sliced Elastic Optical Path Network”, IEEE Comm. Letters, Vol.15, No.8, August, 2011.
2. Masahiko Jinno, et al, “Spectrum- Efficient and Scalable Elastic Optical Path Network: Architecture, Benefits, and Enabling Technologies”, IEEE Comm. Magazine, 2009.
3. http:// www.ska.ac.za/index.php
4. Agrawal G. P., Nonlinear Fibre Optics, Academic press, 4th edition, 2007.
5. Kaminov I.P., et al, Optical Fibre Telecommunications IVB Systems and Impairment, Academic Press, 4th ed. San Diego, 2002.

Duncan Kiboi Boiyo was born in 1984 in Mt. Elgon, Kenya, obtained his BSc. Physics/ Mathematics Honors degree in 2009 from Moi University, Kenya and MSc Physics in 2013 at University of Eldoret, Kenya. He’s currently a PhD student at NMMU, working on SKA optical fibre network; communication, networking and network impairments.

# WORK IN PROGRESS:

# DATA CENTRE & CLOUD

# Micro Data Centres for Multi-Service Access Nodes and their effect on latency and services

David van Wyk*, Jacques van Wyk#
Department of Electrical, Electronic and Computer Engineering, University of Pretoria
E-mail: vanwyk.dvw@gmail.com*, jhvanwyk@up.ac.za#

**Abstract – Large data centres are usually used to maximize efficiency and lower CAPEX and OPEX, but this may introduce larger latencies for users not in close proximity of a data centre or for those that do not have direct fibre connections. Our proposed solution is to implement micro data centres within the MSAN infrastructure to reduce latency experienced to large data centres, as well as other services like caching and media delivery.**

**Index Terms – latency, caching, backup and replication, MSAN, VDSL.**

## I. INTRODUCTION

As computer processing speed has increased over the last three decades, the number of applications for which computers could potentially be used has expanded vastly. One of these applications, and possibly the one that had the most significant impact on the world as a whole, is the Internet. Ever since the Internet was introduced in the late 1960s and early 1970s the main buzzword that accompanied any mention of the Internet was the word *latency*.

In the beginning, latency might not have been as much of an issue as it is today. This is mostly due to the fact that the amount of data and information that was in circulation back then was not nearly as much, or as frequent as what is currently going on. While its significance is often underplayed, latency plays a central role in the transmission of data. What is also very important to note is that as applications have become more sophisticated, latency has become a much larger issue and a real problem.

As can be seen in Table 1, modern technologies have increased at a fairly rapid rate, with some increasing more rapidly than others. One of the characteristics that have not kept pace is the latency of telecommunications networks.

| | 1980s | 2011 | Improvement |
|---|---|---|---|
| **CPU speed** | <10MHz | 4x2GHz | ≈1200x |
| **Storage Space** | <20MB | >2TB | ≈100,000x |
| **Memory** | <4MB | 8GB | 2000x |
| **Home Bandwidth available** | <1Mbps | 40Mbps | >40x |
| **Net bandwidth** | 3Mbps | 10Gbps | ≈3000x |

**Table 1: The table shows how computer specifications have improved over the last 30 years [1].**

One way to describe this advancement or lack thereof is through the following example. In 1983, a remote procedure call (RPC) of 16KB would have a round-trip time (RTT) of approximately 2.5ms. In 2011 the same RPC packet size in a pair of Linux servers was found to have a RTT of just over 80µs over a 10Gbps Ethernet connection. This time is a very small improvement (when compared to how much other technologies have evolved) over the speeds achieved in 1983. In fact, this improvement in time was just more than 30x faster than the 1983 speeds. When compared with the fact that core network bandwidth has improved by well over 3000x since 1983, it becomes evident that last mile solutions do not keep up with improvement trends, leading to higher latency [1].

There is a non-existent to low service provisioning of backup and replication services in the broader residential and small, medium and micro enterprises (SMMEs) sectors. Usually only larger companies can afford a proper data backup solution. Although there are a few cloud storage solutions out there, none of them are catering for complete system backup solutions (imaging) of systems in case of complete system failures, such as hard drive crashes.

Current Asymmetric Digital Subscriber Line (ADSL) technology is asymmetric in nature, with upstream rates of less than 3.3 Mbps for ADSL2+M (but on average less than 1.1 Mbps for ADSL2+ service offers in South Africa) [2]. This will mean that a typical 250 GB hard drive will idealistically take 63 hours to transfer, which is way too long. With the introduction of 20 and 40 Mbps Very-high-bit rate DSL (VDSL) technology [3] by Telkom in March 2013 [4], upstream speeds of up to 16 Mbps are possible [4], which now makes backup and replication services viable. One problem that still remains is the latency within the network as a whole. Large data centres maximize efficiency and lower CAPEX and OPEX, but this may introduce large latencies for users not in close proximity of a data centre or for those that do not have direct fibre connections. SpeedTest.net collects large amount of user test data and an average latency of 14 ms and 45 ms are typically experienced from Pretoria to the Telkom Internet test point in Centurion on ADSL and Telkom Mobile respectively. Similarly, 15 ms and 49 ms are experienced to the Telkom SAIX test point in Randburg. When considering next generation technologies, a latency of 1 ms is proposed [5].

Our proposed solution is to implement micro data centres within the Multi-Service Access Node (MSAN) infrastructure to reduce latency experienced to large data centres, as well as other services like caching and media delivery. Caching and Storage Arrays can be implemented as a Phase 1 solution. During a second phase this can be extended to incorporate processing capability to extend

offered services to other Cloud-based services, i.e. Applications as a Service (AaaS). MSANs can also provide the capability to provide "Virtualized databases" closer to the user.

## II. Typical Applications for Micro Data Centres

Since 2010, Telkom implemented a fibre to the curb/cabinet (FTTC) strategy, upgrading the copper feeder cables from exchanges with fibre optic cables (also known as Digital Loop Carriers (DLCs)) and moving DSL Access Multiplexing (DSLAM) equipment closer to customers. DLCs have been implemented in the form of MSANs. The main reason behind this implementation is that the fibre optic connections can achieve much greater speeds, have lower attenuation per unit length, lower latency and no Electromagnetic Interference (EMI) properties compared to copper wire. This reduces the overall latency, mainly contributed by the copper part of the link.

Another reason for the reduction in latency is that the user is connected directly to the MSAN instead of connecting to a main data center facility through the core Public Switched Telephone Network (PSTN) network. Additional delays are mostly due to the fact that the signal has to pass through multiple switches and network nodes, whereas with the MSAN principle, the signal would be passing through only one set of network nodes.

By implementing a micro data centre within the MSAN, that could act as a caching/file/media/backup server. Caching could significantly reduce the delay experienced by first having to retrieve the data from the Internet. It may also be used for frequently accessed content within a community. Examples of such content would be a very popular YouTube video and gaming platforms such as Steam and Origin, which often provide online updates to games, to cache updates to more popular games on these servers. Telkom already offers a bolt-on bundle for R49 [6] a month where any data expended on downloading updates through these platforms is treated as uncapped and does therefore not have an effect on the user's data cap. Although this offer was a limited time offer, the implementation of micro data centres could make applications like these more viable.

A final example would be a video-on-demand type of service where popular videos or movies can be rented from a BoxOffice-type of platform and streamed directly to the user's television or computer. Telkom also made a similar offer to the gaming bolt-on bundle where it allowed soft-capped users to use DStv's online services without the data being deducted from their bundle [6].

Most users do not realize the importance or value of their personal data, till it's gone after a hard drive crash or even theft. By providing a full-backup service as an additional service to line rental, user's data can be protected and be available with little delay due to close proximity of mini-data centres. This would help alleviate congestion on certain links and thus alleviate congestion on exchanges. Backup data is treated locally - only Internet traffic affect data caps. Many Internet users in South Africa cannot afford to pay for premium Internet connections [7]. This causes by far the biggest challenge that comes with implementing this type of infrastructure. The solution to this problem however lies with the fact that MSANs will be enabled with a VDSL connection to all its end users, even if the user is not paying for that particular service. This would enable the user to use the full speed of the connection to access any or all of the above-mentioned services, while Internet access may be throttled if desired.

## III. Conclusion

Micro data centres may provide a viable solution to reduce latency experienced by end users, especially i.t.o. backup and replication services, but also provides a caching platform for recently-used Internet data, multi-media and other files.


## ACKNOWLEDGEMENT

This work is based on Research supported in part by the National Research Foundation of South Africa (Grant specific UID 83913) and our industry partners Telkom, Bytes Universal Systems, Tellumat and EMC. The grantholder acknowledges that opinions, findings and conclusions or recommendations expressed in any publication generated by the NRF supported research are that of the author(s), and that the NRF and our industry partners accept no liability whatsoever in this regard.


## REFERENCES


[1] Diego Ongaro, Stephen M Rumble, Ryan Stutsman, Mendel Rosenblum, and John K Ousterhout, "It's Time for Low Latency," in *Procedings of the 13th USENIX conference on hot topics in operating systems*, Napa, 9-11 May 2011, pp. 11-16.

[2] ITU. (2009, January) G.992.5: Asymmetric digital subscriber line 2 transceivers(ADSL2)-Extended ADSL2(ADSL2+). [Online]. http://www.itu.int/rec/T-REC-G.992.5/e

[3] ITU. (2004, June) G.993.1: Very high speed digital subscriber line transceivers (VDSL). [Online] Accessed 13 June 2014. http://www.itu.int/rec/T-REC-G.993.1/en

[4] MyBroadband. (2013, January) Will you area get 40Mbps VDSL. [Online]. Accessed 13 June 2014. http://mybroadband.co.za/news/broadband/68378-will-your-area-get-40mbps-vdsl.html

[5] Federico Boccardi, Robert W Heath, Angel Lozano, Thomas L Marzetta, and Petar Popovski, "Five Disruptive Technology Directions for 5G," *IEEE Communications Magazine*, pp. 74-80, February 2014.

[6] Steve Whitford. (2014, June) Uncapped Gaming R49 Monthly. [Online] Accessed 13 June 2014. http://gaming.do.co.za/articles/localnews/uncapped_gaming_r49_monthly.htm

[7] MyBroadband. (2014, May) More depressing broadband news for SA. [Online]. Accessed 13 June 2014 http://mybroadband.co.za/news/broadband/101624-more-depressing-broadband-news-for-sa.html



**David van Wyk** received his B.Eng (Electronic Engineering) degree in 2013 from the University of Pretoria and is currently studying towards his B.Eng (Hons). His research interests are mainly focused on telecommunications with a specific view on Internet and fixed line infrastructure.

**Jacques H. van Wyk** obtained his B.Eng and M.Eng degree in Electronic Engineering from the University of Pretoria in 1997 and 1999 respectively. He is currently a Senior Lecturer in the Department of Electrical, Electronic and Computer Engineering and Director of the Centre for Telecommunication Engineering for the Information Society (CeTEIS). He is a registered professional engineer.

# WORK IN PROGRESS:

# INTERNET SERVICES AND APPLICATIONS

# Using a Natural User Interface to Support Information Sharing Among Co-Located Mobile Devices

Timothy Lee Son, Janet Wesson and Dieter Vogts
Department of Computing Sciences
Nelson Mandela Metropolitan University, P. O. Box 77000, Port Elizabeth 6031
Tel: +27 41 5042323, Fax: +27 41 5042831
email: {Timothy.LeeSon, Janet.Wesson, Dieter.Vogts}@nmmu.ac.za

**Abstract - As the number of mobile devices continues to increase, these devices need to interact with one another. The communication among these devices enables mobile users to share information with each other more easily and frequently. However, existing information sharing methods between co-located mobile devices are clumsy and inefficient. A need exists for a more intuitive and efficient solution to support information sharing. The current process requires the user to repeat several steps in order to share information with multiple, co-located users. This research will identify the problems with existing information sharing methods and investigate whether a natural user interface (NUI) can be designed to support information sharing among co-located mobile devices.**

**Index Terms - Information sharing, natural user interface, proxemics, mobile computing**

## I. Introduction

The processing and storage capabilities of mobile devices has grown rapidly. Mobile users and their communication devices, such as smartphones, tablets, and notebook computers, need to exchange information on a regular basis [1]. Current information sharing methods in mobile devices require users to share information through a series of repetitive manual steps, which are time-consuming and ineffective.

The emergence of new computing interfaces has led to the development of many new interaction techniques, such as scrolling, flipping, dragging, and clicking [2]. The new interfaces use natural and intuitive interaction techniques, which make user actions easier to perform and require less time. For the purpose of this research, intuitive is taken to mean the ease of understanding or knowing something without any conscious reasoning [3].

The problem with existing information sharing methods among co-located mobile devices is that the process is clumsy and tedious. None of the existing methods use a natural user interface (NUI) to support information sharing.

NUIs are typically present in technologies that allow users to perform natural movements and gestures to control the application or manipulation of on-screen content. NUIs show great promise for defining new ways of interactive computing [4]. Thus, they could potentially be used to support information sharing among co-located mobile devices.

This paper commences with a discussion of related work in Section II regarding information sharing and NUIs. Section III describes the research objectives. Conclusions and future work are discussed in Section IV.

## II. Related Work

### A. Information Sharing

Information sharing is an activity through which information is exchanged among organisations and people, using different technologies. Information sharing is one of the most basic activities of coordination [5]. The sharing of information occurs in every facet of our daily lives, because of the constant communication between people and the need to coordinate events.

The technological progression of mobile devices, such as smartphones, tablets, laptops, and other handheld devices, have allowed these devices to store and share information. This information can include user location, audio, video, and orientation, which is made possible through the increasing number of sensors on mobile devices [1]. The types of information most commonly shared among mobile device users are documents, videos, audios, and images [1].

### B. Natural User Interfaces

Recent developments in input technologies are changing the way we interact with computing devices [6]. The traditional input devices such as the mouse and keyboard are being replaced by touch and motion-based interfaces. These interfaces are typically referred to as NUIs.

The primary goal of NUIs is to provide users with an intuitive way to interact with applications so that no learning or training is required to perform specific actions. There are several types of NUIs, namely touch screen interfaces, gesture recognition systems, speech-based interfaces, gaze-tracking interfaces, and proximity-based interfaces.

NUIs possess several benefits [6]. The two most notable benefits are that NUI novices are able to quickly learn basic functionality by a simple demonstration of the action. NUI novices are also able to move rapidly and progress with relative ease to become an expert. A NUI can include any the following interaction techniques:

*Touch:* Multi-touch is the ability of the interface, typically a touchscreen, to recognise multiple points of contact with the interface. Typically, users of mobile devices use their fingers as an input device to directly interact with the information on their screens [7].

*Proximity:* Proximity-based interaction is the interaction of a system with different entities (people, digital devices, and non-digital objects) based on their distance, orientation, movement, identity, and location [8].

*Gestures:* Gestures refer to any motion involving physical movements of the user's hands or fingers. Gestures provide a natural, direct, and intuitive way of interacting with a computing device, allowing easier human-computer interaction for all types of users. There are two types of

gestures, namely touch and in-air gestures. Touch gestures are mainly found in touch screen interfaces where the user performs a specific gesture to achieve a certain system response. In-air gestures are any movements performed by the user that are recognised by the system without touching the screen [9].

## III. Research Objectives

The main research question of this research is: *How can a NUI be designed to intuitively and efficiently support information sharing among co-located mobile devices?*

### A. Methodology

This research will use the Design Science Research Methodology (DSRM) that focuses on the development and performance of designed artefacts. The DSRM consists of several activities to be performed: 1) identify and understand the problem, 2) identify and outline the artefact, 3) design and develop the artefact, 4) demonstrate the artefact, and 5) evaluate the artefact.

### B. NUI Techniques

The information sharing process typically consists of selecting the file(s), selecting the recipient(s), initiating the transfer, and accepting the transfer. One of the main objectives of this research is to investigate different NUI interaction techniques to design appropriate, easy to use interaction techniques for sharing information. Some examples of existing (and new) interaction techniques [3] for the different phases of sharing information are the following:

Selecting the appropriate file(s) could be achieved by performing a touch gesture resembling a grab action. A single recipient could be selected by using the mobile device as a pointing device and pointing in the general direction of the recipient's device. Similarly, multiple recipients could be selected by waving the device in the general direction of the recipients' devices.

Initiating the transfer of the file(s) could be performed by moving the mobile device in the direction of the recipient's device, in a motion resembling handing over a physical object. Similarly, accepting the transfer can be initiated by the recipient(s) performing an in-air gesture resembling receiving a physical object.

A number of the above interaction techniques require the devices to be proximity-aware. Therefore, research will be conducted into making co-located mobile devices proximity-aware of each another.

### C. Prototype

A number of software prototypes will be developed to evaluate whether a NUI, implementing the proposed NUI interaction techniques, can intuitively and efficiently support information sharing among co-located mobile devices.

### D. Evaluation

The intuitiveness and efficiency of the proposed NUI techniques will be determined by means of a user study. Data will be collected by means of observation and logging, interviews, and questionnaires.

## IV. Conclusions and Future Work

This paper has identified the problems of existing information sharing methods used by mobile devices and highlighted the potential use of a NUI as a possible solution. The envisaged research contribution is to determine how a NUI can be designed to provide an intuitive and efficient solution to information sharing among co-located mobile devices.

The next phase of this research will involve the identification of the relevant NUI techniques to be used with mobile devices to support information sharing. A prototype will then be developed incorporating a NUI and evaluated to determine the intuitiveness and efficiency of the prototype in supporting information sharing among co-located mobile devices.

## V. Acknowledgements

The author would like to acknowledge the financial assistance of the NMMU/Telkom Centre of Excellence, without which this research would not be possible.

## VI. References

[1] T. Zhang, S. Madhani, and E. Van Den Berg, "Information for Mobile Users and Devices," pp. 7–11, 2005.

[2] J. Oh, H. Robinson, and J. Lee, "Page flipping vs. clicking: The impact of naturally mapped interaction technique on user learning and attitudes," *Comput. Human Behav.*, vol. 29, no. 4, pp. 1334–1341, Jul. 2013.

[3] A. Britton, R. Setchi, and A. Marsh, "Intuitive interaction with multifunctional mobile interfaces," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 25, no. 2, pp. 187–196, Jul. 2013.

[4] S. Seow, D. Wixon, A. Morrison, and G. Jacucci, "Natural user interfaces: the prospect and challenge of touch and gestural computing," *CHI '10 Ext. Abstr. Hum. Factors Comput. Syst.*, pp. 4453–4455, 2010.

[5] D. Saab and E. Maldonado, "Building global bridges: Coordination bodies for improved information sharing among humanitarian relief agencies," *Proc. 5th Int. ISCRAM Conf.*, no. May, pp. 471–483, 2008.

[6] J. Blake, *Natural User Interfaces in .NET*. Greenwich, CT: Manning Publications Co., 2012.

[7] J. Yao, T. Fernando, and H. Wang, "A multi-touch natural user interface framework," *2012 Int. Conf. Syst. Informatics*, no. Icsai, pp. 499–504, May 2012.

[8] N. Marquardt, "Proxemic interactions with and around digital surfaces," *Proc. 2013 ACM Int. Conf. Interact. tabletops surfaces - ITS '13*, vol. 2, pp. 493–494, 2013.

[9] S. Agrawal, I. Constandache, S. Gaonkar, R. Roy Choudhury, K. Caves, and F. DeRuyter, "Using mobile phones to write in air," *Proc. 9th Int. Conf. Mob. Syst. Appl. Serv. - MobiSys '11*, p. 15, 2011.

**Timothy Lee Son** received his BCom Honours degree (Cum Laude) in 2013 from the Nelson Mandela Metropolitan University and is presently studying towards his Master of Commerce degree in Computer Science and Information Systems at the same institution.

# Preliminary Thoughts on Services without Servers

Philip Machanick and Kieran Hunt
Department of Computer Science
Rhodes University, Grahamstown
Tel: +27 46 603 8635, Fax: +27 46 603 7608
email: {p.machanick,kieran.hunt}@ru.ac.za

**Abstract-Warehouse-scale computing supports cloud-based services such as shared disk space, computation services and social networks. Although warehouse-scale computing is inexpensive per user, the cost to entry is high, and the pressures to generate revenues to cover costs leads service providers to pursue monetizing services aggressively. In this paper, we explore some ideas for removing the need for central servers by exploiting peer-to-peer technologies.**

**Index Terms—distributed systems, cloud, peer-to-peer**

## I. INTRODUCTION

Large-scale services such as shared file systems (Google Drive, Apple iCloud, Dropbox and Microsoft's OneDrive) and on-demand computation resources (such as Amazon's Elastic Compute Cloud, or EC2) have started to proliferate with the generic "cloud" label. Such services build on infrastructure originally created to support large-scale services such as Google search, Amazon's bookstore and Facebook. Many of these large-scale services – Google's search, Facebook and Twitter to name a few – are free to use, but have a commercial aspect in that their creators use user traffic to generate revenue streams such as advertising.

Despite impressive gains in implementation of such services [Mishra *et al*. 2010], they fall far short of the promise of distributed computing. They lack a transparent namespace – most such services still look more like networked services with names that appear to relate to a particular server, even if there is some virtualization behind the scenes. Scalability is implemented by large-scale resources in a small number of places, rather than by placing resources near the users. Cost is not shared over the users, except in the indirect sense that a large user base makes for a more attractive target for advertisers. The last point also points to one of the weaknesses of this sort of service from the user point of view: if *you* are seen as the product, as was famously said of Facebook [8], "your" service provider constantly is under pressure to monetize you.

While some services inherently are salable – e.g., Amazon's EC2 generates revenues directly [12] – providing services that users do not expect to pay for should be based on shared cost-sharing, rather than on free services paid for by advertising. Otherwise, the temptation to monetize invasion of privacy is too high.

In this paper, we explore the extent to which an existing shared-cost model, peer-to-peer (P2P) file-sharing, can adapt to a wider range of services. We start with specific services, then generalize to wider possibilities.

The remainder of this paper is structured as follows. Section II summarizes previous work, including approaches to scalability and distributed computing, as well as P2P technologies. Section III outlines how some simple services can be implemented using P2P, including existing work and our own ideas, and Section IV contains conclusions.

## II. BACKGROUND

In this section we briefly review the relationship between distributed computing and the cloud, which is a poor approximation to the intent of distributed services, and the broader concept of scalable services.

### A. Distributed Computing and the Cloud

Distributed computing in its general form implies a number of properties [14]:

- *location-transparent naming* – a name of an entity should be related to its logical purpose or relationship to other entities, not where it is situated
- *locality-independent resources* – whether a resource is local, on a local network or in a more remote location is a performance detail, and should not be an inherent property of any resource
- *decentralized scalable infrastructure* – a system should be able to work over a wide geographic region, which also implies an appropriate level of security.

Cloud-based services violate basic properties of distributed computing. To the user, it is clear that there is an external server, and hence, a distinction between purely local resources and cloud-based services. Names are therefore not full location-transparent. Further, cloud services require connection to a server (even if limited offline activity may be allowed), making network connection essential rather than a performance detail. Scalability is achieved by concentrating resources in warehouses of computers [3], rather than by distributing resources widely.

### B. Scalable Services

In general, how can services be scaled up? Some of the scalability problem is in scaling up large-scale computation; the general case is hard because some problems are not partitionable [1]. Here we constrain ourselves to services where computation is not large-scale; even so we have problems of scaling up naming. Traditionally, name-scaling has been a function of middleware [2]. We can however isolate scalable naming as a single concept as, for example, in distributed hash tables (DHT) [13, 9], which are widely used in P2P systems (though some have argued that sharing

in P2P systems is inherently scalable [5] without DHTs).

In general, scalable services should not depend on the number of users to be viable and – even better – should become more viable as the number of users increases.

## III. Simple P2P-Based Shared Services

A number of services layered on P2P already exist. For example, Skype voice over IP (VoIP) is layered on a proprietary P2P protocol [4]. BASS is a scalable video on demand service based on Bittorrent [6].

More recently, the Bittorrent Sync application programming interface (BTS API) has been released[1], allowing applications to be layered on top of the Bittorrent Sync service [7], which provides secure P2P file sharing. Applications that use the BTS API include Vole[2], a twitter-like service that shares the underlying files using BTS rather than a central server, and SyncNet[3], which implements a web server by distributing the files across clients.

These services indicate the general possibility of breaking dependence on central servers.

Our work builds on these foundations. We are investigating the extent to which you can completely break away from networked services and implement true distributed services based on P2P protocols. Our starting point is implementing a twitter-like service on top of the BTS API, using Java for portability – thereby eliminating the web browser. We will follow with a service more like Facebook, then look into how services like email can be made to work in a purely distributed fashion [10].

Implementation of a twitter-like application built on the BTS API should be very simple. Users wanting to share updates exchange their encryption keys (these could be mapped to usernames or handles) to allow other users to request and download their updates. A simple GUI sets such a program apart from traditional, web-based applications. Otherwise, BitTorrent Sync handles most of the operation.

## IV. Conclusions

Distributed computing is a powerful idea that has somehow got lost in a network-centric world. Warehouse-scale computing uses distributed computing concepts internally, including highly scalable distributed file systems, yet the interface presented to the user uses network-like names, even if the actual resource named may be disguised.

The proposal presented here is to implement true distributed services without servers, based on P2P technology. The extent to which such services can be implemented is part of our investigation; if we can implement a significant range of such services, we can reduce the need for central resources, and hence the pressure to monetize services even when it is inappropriate to do so.

Further, if these ideas work, not only can they scale very well, but they have a very low barrier to entry.

## Acknowledgements

This work was undertaken in the Distributed Multimedia CoE at Rhodes` University, with financial support from Telkom SA, Tellabs, Genband, Easttel, Bright Ideas 39, THRIP and NRF SA (TP13070820716). The authors acknowledge that opinions, findings and conclusions or recommendations expressed here are those of the author(s) and that none of the above mentioned sponsors accept liability whatsoever in this regard.

[1] http://www.bittorrent.com/sync/developers/api

[2] http://vole.cc/

[3] http://jack.minardi.org/software/syncnet-a-decentralized-web-browser/

## References

1. G.M. Amdahl. “Validity of the single processor approach to achieving large scale computing capabilities”, *Proc. Spring joint computer conference, AFIPS ’67*, 1967 pp. 483–485.
2. G. Ballintijn, M van Steen and AS Tanenbaum. “Scalable naming in global middleware” *Proc. 13th Int’l Conf. on Parallel and Distributed Computing Systems*, 1999 pp. 624-631.
3. L.A. Barroso and U. Hölzle. “The datacenter as a computer: An introduction to the design of warehouse-scale machines”, *Synthesis lectures on computer architecture*, 2009. doi:10.2200/S00193ED1V01Y200905CAC006
4. S.A. Baset and H. Schulzrinne. "An analysis of the Skype peer-to-peer internet telephony protocol." *arXiv preprint cs/0412017*, 2004.
5. Y. Chawathe, S. Ratnasamy, L. Breslau, N. Lanham, and S. Shenker. “Making gnutella-like P2P systems scalable.” *Proc. 2003 conference on Applications, technologies, architectures, and protocols for computer communications*, 2003, pp. 407-418.
6. C. Dana, D. Li, D. Harrison and C.-N. Chuah. “BASS: BitTorrent assisted streaming system for video-on-demand.” *IEEE 7th Workshop on Multimedia Signal Processing, 2005* pp. 1-4.
7. J. Farina, M. Scanlon, and M. Kechadi. “BitTorrent Sync: First Impressions and Digital Forensic Implications.” *Digital Investigation* vol. 11, pp S77-S86, 2014
8. N. Friesen. “Education and the social Web: Connective learning and the commercial imperative.” *First Monday* vol. 15, no. 12, 2010.
9. D. Korzun and A. Gurtov. “Survey on hierarchical routing schemes in ‘flat’ distributed hash tables.” *Peer-to-Peer Networking and Applications* vol. 4 no 4, pp 346–375, 2011.
10. P. Machanick. "A distributed systems approach to secure Internet mail." *Computers & Security* vol. 24 no. 6 pp 492–499 2005.
11. A.K. Mishra, J.L. Hellerstein, W. Cirne, and C.R. Das. 2010. “Towards characterizing cloud backend workloads: insights from Google compute clusters”, *SIGMETRICS Perform. Eval. Rev*. vol. 37, no. 4, pp 34-41, March 2010.
12. Y. Sangho, A. Andrzejak, and D. Kondo. “Monetary Cost-Aware Checkpointing and Migration on Amazon Cloud Spot Instances”, *IEEE Transactions on Services Computing*, vol. 5, no. 4, pp 512,524, Fourth Quarter 2012.
13. D. Tam, R. Azimi and H.-A. Jacobsen. “Building content-based publish/subscribe systems with distributed hash tables”, in *Databases, Information Systems, and Peer-to-Peer Computing*. Springer Berlin Heidelberg, 2004, pp 138-152.
14. A.S. Tanenbaum and M. van Steen. *Distributed Systems: Principles and Paradigms*. Upper Saddle River, NJ: Prentice Hall, 2002.

**Philip Machanick** received his PhD degree in 1996 from the University of Cape Town. His research interests include distributed computing, computer science education and bioinformatics.

**Kieran Hunt** received his undergraduate degree from Rhodes University in 2013 and is presently studying towards his Honours degree at the same institution. His research interests include distributed computing.

# Efficiency of Mobility Management schemes on virtualized shared Mobile Networks

Ofentse Noah, Neco Ventura
Department of Electrical Engineering
University of Cape Town, Cape Town
email: {ofentse, neco}@crg.ee.ac.za

**Abstract-Network sharing in the mobile network environment has shown to be important in CAPEX and OPEX reduction. Several studies have been conducted and NFV has come out as the main platform for virtualization. This paper briefly elaborates on work in progress which will evaluate mobility management schemes on NFV platforms.**

**Index Terms—Network Functions Virtualization, Radio Access Network, Network Sharing**

## I. INTRODUCTION

Mobile network operators (MNO) face the challenge of consistently deploying new services and technologies in the face of declining average revenue per user (ARPU) [1]. To add to the challenge, MNO services to the subscriber need to be cheaper in order to remain competitive. This leaves for little revenue to fund Capital Expenditure (CAPEX) for deploying new services and technologies.

Currently most MNOs have single Radio Access Network (RAN) ownership model [2]. Each MNO deploys its own RAN to provide service coverage to subscribers. While the single RAN ownership model allows for deployment of services that differentiate from other MNOs, it is associated with extremely high CAPEX and Operating Expenditure (OPEX). This affects established MNOs, it also prevents any new MNOs from entering the telecommunications space.

As a result, MNOs have implemented cost reduction measures. We only consider the technical measures. In some cases, MNOs have outsourced some of their RAN site operations and maintenance. This enables the MNO to focus on service delivery to subscribers. New MNOs usually enter into roaming agreements with established MNOs in order to establish themselves in the market. However, it is clear that these cost reduction measures are not achieving optimal cost savings.

Therefore MNOs like China Mobile, Orange and Verizon to name a few have contributed towards the ETSI Whitepaper on NFV [3] as a means for further cost saving[ 4]. It is expected that they will reduce overall network costs and have faster service deployment and centralized network monitoring and control.

The rest of this paper is arranged as follows: it looks at the related work in the field of network sharing including background. The proposed approach is described. Finally conclusions are presented.

## II. RELATED WORK AND LITERATURE

Network sharing has been used by MNOs for sharing the cost burden of deploying a mobile network. A common example of network sharing is roaming [5], where only the RAN is actively shared. Another common network sharing example is passive sharing where the actual site, the power supply, air conditioning and other site supporting infrastructure are shared. Network sharing has been implemented in many countries and it is said that it could save global network operators 60 Billion United States Dollars (USD) over a period of 5 years [6].

### *A. Network sharing use cases*

Cost saving from network sharing has led to the establishment of network sharing use cases by the 3GPP [7]. For brevity sake, we give an overview.

- MNOs can share common RAN as in 3GPP release 99. This can be done with frequency sharing or not by allowing them direct access to Radio Network Controller (RNC) carrier layer.
- Geographical split of coverage where two or more MNOs with individual frequency licenses with which their respective RANs cover different parts of the country. However, when combined, their coverage covers the whole country.
- One MNO covering a specific region in the country and another MNO is allowed to utilize this coverage for its own subscribers. Outside this region each MNO provides its own coverage.
- One MNO has a frequency license that it shares with other operators. On the other hand, a number of MNOs decide to pool their spectra and share the total pool of spectra.
- Multiple RANs can share a common Core Network. Multiple RANs can belong to different public land mobile networks (PLMNs) and different MNOs. Due to various deployment possibilities, parts of the common Core Network (CN) can be shared.

These use cases have formed a basis for future network sharing standards as detailed in [7].

### *B. Network Functions Virtualization*

Network Functions Virtualization (NFV) is an emerging technology [4] that offers a means to virtualize some or all functions of an MNO. With NFV, equipment costs can be

reduced through using standard x86 hardware [3]. Since virtualization is used, sharing is possible even over heterogeneous x86 hardware.

### C. New Business models

New work points to outsourced RAN as the future operating model. These new business models include having Virtual Network Providers (VNP) that provide network infrastructure and Virtual Network Operators (VNOs) that utilize the Virtual Network (VNet) provided by the VNPs.

VNOs no longer have to worry about the network deployment costs (CAPEX). This responsibility is now taken by the VNP. The VNP also takes responsibility for the OPEX of RAN. Therefore the VNO can now focus on providing services.

The geographical coverage of the RAN can be greatly improved through either a) the VNP deploying one large network or b) several VNPs deploying networks that form parts of the total required coverage in a). The latter provides further cost savings.

Lastly this business model allows for VNOs to purchase only the infrastructure that they need. For example, VNOs focusing on connecting home appliances can purchase a VNet that covers a different foot print to that of a VNO that focuses on connecting Smart Grid devices. This enables even small players to enter the telecommunications market.

### D. Current challenges

Several challenges exist with network sharing through NFV [8]. Since VNPs are bringing together various networks through virtualization, there is a challenge with interference coordination. Load balancing for the virtualized network elements is a challenge that is being investigated for optimal performance of the virtualized platforms. Another challenge is mobility management of VNO traffic between the various VNPs that provide service to a particular VNO.

## III. Proposed Approach

The focus of this study is to evaluate mobility management schemes across VNPs that provide VNOs services. This will be done by creating multiple VNPs that will provide VNets for the VNOs. Mobility Management will be implemented in the virtualized instance and compared to roaming based network sharing.

The comparison between these two cases will give a better understanding of the mobility management performance in virtualized shared networks in relation to the state of the art traditional network.

## IV. Conclusions

This paper introduced the problem of mobility management within the domain of virtualized mobile networks. Related work and literature was presented and future work will work on mobility management on the virtualized network platform where VNets are involved. This should lead to a conclusive comparison between the efficiency of mobility management schemes on virtualized shared networks and traditional roaming shared networks.

## Acknowledgements

This research is supported by Telkom South Africa, Jasco/TeleSciences and the Department of Trade And Industry/National Research Foundation / Technology and Human Resources Programme (DTI/NRF/THRIP).

## V. References

[1] X. Costa-Perez, “Latest trends in telecommunications standards”, ACM SIGCOMM Computer Communication Review, 43:64, Vol. 71, 2013.

[2] A. Khan, W. Kellerer, K. Kozu and M. Yabusaki, “Network Sharing in the Next Mobile Network: TCO Reduction, Management Flexibility, and Operational Independence”, IEEE Communications Magazine, pp 134-142, 2011.

[3] Network Function Virtualization, white paper, SDN and OpenFlow World Congress, Darmstadt, Germany, October 22-24, 2012.

[4] X. Costa-Perez, J. Swetina, T. Guo, R. Mahindra and S Rangarajan, “Radio Access Network Virtualization for Future Mobile Carrier Networks”, IEEE Communications Magazine, pp 27-35, July 2013.

[5] Y. Al-Jarbou and U. Baroudi, “Performance of Heterogenous Traffic in Roaming Based Sharing Multi-Operator WCDMA Networks”, 2nd International Symposium of Wireless Communications Networks, pp 394-398, 2005.

[6] Active ran sharing could save 60 billion usd for operators,. http://www.cellular-news.com/story/36831.php. Last accessed: 10 July 2013.

[7] 3GPP TR 22.852, v13.0.0, Study on Radio Access Network (RAN) Sharing Enhancements Study Item, http://www.3gpp.org/ftp/specs/archive/22_series/22.852/. Last Accessed: 1 Aug 2014.

[8] Y. Zaki, L.Zhao, C.Georg and A. Timm-Giel, “LTE Wireless Virtualization and Spectrum Management,” *Proc IFIP WMNC Conf.*, Oct 2010.

**Ofentse Noah** received his undergraduate degree in 2003 from the University of Cape Town and is presently studying towards his Doctor of Philosophy degree at the same institution.

# A Model of Fraud Detection in Mobile Transaction via Unstructured Supplementary Service Data

Kulani E. Vukeya and Okuthe P. Kogeda
Department of Computer Science
Tshwane University of Technology, Private bag X680, Pretoria 0001
Tel: +27 12 382 9640, Fax: +27 12 3829525
email: kulaniv@gmail.com[1] ; kogedapo@tut.ac.za[2]

Abstract- **Mobile payments have become more popular and are used to perform transactions that carry sensitive information that should only be visible to the person performing that particular transaction. One of the well-known services to perform such transactions is called Unstructured Supplementary Service Data (USSD) and due to the sensitive and important details it performs it has become very attractive target for fraudsters. The problem is that critical threats such as fraudulent transactions, request/response manipulations, weak encryption, and insecure message communications are directly triggering revenue loss for mobile payment service providers, customers and financial institutions. In this paper, we seek to design and implement a model of fraud detection in mobile transaction via USSD. We use Bayesian Network model to define the type of transactions given the customers behavior and EM algorithm to learn the parameters in the model using observed data from transactions via USSD services across the network. The proposed system has been designed.**

**IndexTerms—USSD,Fraud,Meta-learning,Bayesian Networks,Expectation-Maximisation(EM)algorithm, GSM, MNO, MPO**

## I. INTRODUCTION

The number of USSD use in transactions has increased rapidly in the past decade and the use of this technology also include personal and confidential details as well as financial records. We introduce a work in progress that seeks to provide a fraud detection model that can be used by telecommunications industries into managing fraud in transactions via Unstructured Supplementary Service Data **(**USSD). To help identify potentially fraudulent users and their typical usage patterns, detect attempts to gain fraudulent entry to customer accounts, discover unusual patterns which may need special attention, find usage patterns for a set of communication services by customer group, by month, etc. The USSD is a session-based, real-time communication technology for supplementary services. USSD is used in sending messages across a GSM network, between a mobile client and an application server. It uses GSM services and GSM Security, which is known to have inherent flaws in its encryption and authentication algorithms [2]**.** Examples of USSD fraud risk scenario is if staff at an MNO intercept the PIN code used for mobile payment transactions. The transaction data are unencrypted and can easily be read somewhere in the system of the USSD Gateway. MPOs that are not MNOs (e.g. banks, TPPs) typically have limited control over the people accessing the USSD servers. So, this type of fraud scenario is more likely to occur at the MNO level. Preventing fraud in mobile transactions is not easy**,** hence it must be approached with well-defined procedures, policies and guidelines that take into consideration the system boundaries, infrastructure, human behavior, culture, literacy level and market dynamics. The simplicity of the protocol has triggered a lot of fraudulent activities on it because many find it easy to temper with and understand the protocol. Developers of this technology normally look at deployment that enables management of USSD-based applications that can enhance end user loyalty and revenue streams only. Figure 1 below is the typical architecture of the USSD centre in a telecommunication industry. The Architecture normally only involves the elements shown below and no fraud detectors are in place on them. In order to rectify this problem extra work on authentication of the mobile application is required, thus introducing fraud detection system between the mobile application and the network**.**

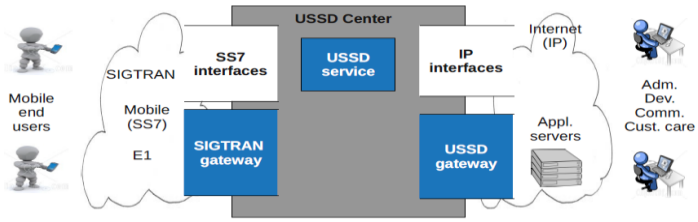



Figure 1: USSD System Architecture

The remainder of this paper is organized as follows: In Section II, we provide background and overview of related work. In Section III, we describe the proposed system design and architecture, In Section IV we provided the model and present conclusion and future work in Section V.

## II. BACKGROUND AND RELATED WORK

Salvatore *et al.* [5] proposed a novel system to address credit card frauds. They describe experiments using meta-learning techniques to learn models that have two key component technologies: local fraud detection agents that learn how to detect fraud and provide intrusion detection services within a single corporate information system, and a secure, integrated meta-learning system that combines the collective knowledge acquired by individual local agents. They used learning algorithms ID3, CART, BAYES, and RIPPER is applied to every partition of the classifiers. These classifiers each attained a True Positive rate of approximately 80% and False Positive rate less than 17%. The results indicated that fraud is detected if the (True Positive) rate is higher than the (False Positive) rate. Our study will apply the BAYES

algorithm to compare the (True Positive) rate of the incomplete data.

Jaakko *at el.* (1996) introduced a Call-Based Fraud Detection in Mobile communications using a Hierarchical regime-switching model [2]. The detection problem is formulated as an inference problem on the regime probabilities. The hierarchical regime-switching model consists of three variables, the first binary variable *Vt* (victimized) ,second binary variable *St*(fraud) is equal to one if the fraudster currently performs, finally, the binary variable *Yt* (call) is equal to one if the mobile phone is being used with state transition matrix:

$pfjk = P(Yt = ilst = j, Yt - l = k); i, j, k = 0,1$ .The dynamics and results found are learned from data using the EM algorithm. Their interest was to estimate the probability that an account was victimized or that fraud is currently occurring based on the call patterns up to the current point in time. It is claimed that the accounts were attacked by fraud when call pattern periods has high traffic from posterior time-evolving probabilities calculations for an account. Therefore, they declare an account to be victimized if the victimized variable at some point exceeds the threshold. Our study will use the EM algorithm to learn the real-time results of the USSD transactions.

## III. System Design and Architecture

This work intends to provide a model that will help minimize fraud in the use USSD service. The FDM (Fraud Detection model).It is data pre-processing techniques for detection, validation, error correction, and filling up of missing or incorrect data. As well calculation of various statistical parameters such as averages, quintiles, performance metrics, probability distributions, etc. For example, the averages may include average length of the session. Again computing user profiles as well as time-series analysis of time-dependent data. Figure 2 below shows the typical Flow of the USSD Transaction highlighted in red, from the Mobile device to the FDM.

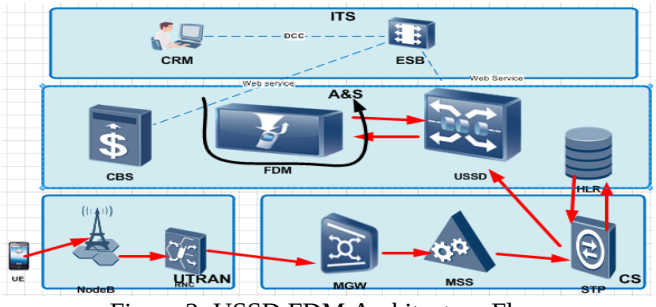


Figure 2: USSD FDM Architecture Flow

## IV. BAYES and EM Algorithms

Maximum likelihood estimation and Bayesian estimation will only work for complete data, i.e. a data set in which each case specifies a value for each of the variables. We will consider the incomplete data set as having been produced from a complete data set by a process that hides some of the data. We can approximate the parameter estimation by the Expectation-Maximization (EM) algorithm. Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors as shown in Equation.1:

$$p(C \mid F_1, \ldots\ldots, F_n) = \frac{p(C)\,p(F, \ldots\ldots, F_n \mid C)}{p(F_1, \ldots\ldots, F_n)} \quad (1)$$

In plain English, using Bayesian terminology, the above equation can be written as

$$posterior = \frac{prior \times likelihood}{evidence} \quad (2)$$

EM is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. M-step the model parameters are optimized using the estimates of the hidden states using the current parameter estimates. The E-step determines the probabilities on the right sides of the equations using the current parameter estimates. EM Algorithm steps are given in Equation. 3 and 4:

E-step: Estimate the complete-data sufficient statistics $t(\times)$ by finding: $t(P) = E(t(\times)Iy, ct(P))$ (3)

M-step: Determine $ct(P+1)$ as the solution of the equations: $E(t(\times)Iy, ct(P))$ (4)

## V. Conclusion and Future Work

The USSD service has its weaknesses from perspectives of security that we have investigated and identified. The service inherits many vulnerabilities due to the security weaknesses inherited in the network flaws. Our paper introduced a FDM model that we are going to use to help minimize fraud encountered in USSD transactions and presented the BAYES and EM algorithms to be used in data calculations and analysis. Next step is to do a critical analysis of the existing security mechanisms that have been identified through literature review and as well as experiments in laboratory. Then we will then evaluate and incorporate the proposed mechanisms and results found into what the FDM built intends to do on the USSD transactions.

## VI. References

[1] Baraka W. Nyamtiga, Anael Sam, Loserian S. Laizer Security Perspectives For USSD Versus SMS In Conducting Mobile Transactions , Vol 1 , No. 1, p.38 – 43, April 2013.

[2] Jaakko Hollmen, Volker Tresp, "Call-based Fraud Detection in Mobile Communication Networks using a Hierarchical Regime-Switching Model" , Vol 2 , No. 1 , p.800 – 896, June 1996.

[3] Pequeno, K A, "Real-Time fraud detection: Telecom's next big step. Telecommunications (America Edition)", Vol 31,No 5, p30-60. November 1997.

[4] Chong, M.K.,"Security of mobile banking: Secure SMS banking",Vol 1,No 2, p3-4, May 2006 Taskin, E, "GSM MSC/VLR Unstructured Supplementary Service Data (USSD) Service", p22-50 , June (2012).

[5] Salvatore J. Stolfo, David W. Fan, Wenke Lee and Andreas L. Prodromidis, "Credit Card Fraud Detection Using Meta-Learning: Issues and Initial Results", Vol 1, No 6, 510-517. July 2013.

**Kulani Vukeya** received her BTECH in Information Technology in 2012 form the Tshwane University of Technology and is presently studying towards her Master of Technology at Tshwane University of Technology.

# WORK IN PROGRESS:

# LIMITED RANGE COMMUNICATIONS

# An arrival-time detection technique for multilateration-based automatic wildlife tracking

Schalk-Willem Krüger and Albert Helberg
TeleNet Research Group
School for Electrical, Electronic and Computer Engineering
North-West University, Potchefstroom Campus
Email: {schalkwillem.kruger, albert.helberg}@nwu.ac.za

***Abstract*—GPS is commonly used to provide localisation capabilities to automatic wildlife tracking systems. However, the energy consumption of GPS-based tags limits the number of position updates retrievable during the tag's battery lifetime, and the GPS module itself adds to the size and mass of a tag. Multilateration tracking systems provide an alternative solution with tags that are lighter, cheaper and longer lasting, but require extra base stations with limited range to be installed. The technique used for arrival-time detection at the base stations has a profound impact on the accuracy and sensitivity of the system. This paper presents work in progress of which the aim is to compare existing detection techniques used in multilateration-based tracking systems, and to propose a technique that improves upon the range of existing techniques. Improving the range will reduce disturbances to the animal and its natural environment and lower the overall cost of the tracking system.**

***Index Terms*—TDoA, multilateration, wildlife tracking, arrival-time detection**

## I. Introduction

Multilateration is a navigation technique commonly used for military surveillance, mobile phone tracking [1], and in aviation for calculating the position of aircraft [2]. The difference in the time at which a signal from an emitter arrives at three or more receivers, called the Time Difference of Arrival (TDoA), is used to locate the position of the emitter. Although the use of multilateration systems is not as common as the use of satellite navigation systems like GPS, there are still many applications where multilateration systems can address the shortcomings of satellite navigation systems.

The technique of multilateration has been studied and used for almost a century [3,4], but recent technological advances allow it to replace conventional systems in more sophisticated applications. One example is automatic wildlife tracking. Existing automatic tags based on GPS are too heavy for some animals, provide a limited number of position updates for the lifetime of the battery, and require an auxiliary RF link for reporting position information back [5].

Improvements to automatic wildlife tracking systems will be to the benefit of researchers. It will, for example, provide more position updates for the same tag lifetime, allow larger sample sizes due to reduced cost, and allow smaller species to be tracked due to decreased tag size. Improvements are also important for feasible solutions that can be used for wildlife protection, such as preventing rhino poaching. Improvements in range can, for example, help to minimise the infrastructure required to deploy a system in a large game reserve.

One crucial aspect of multilateration is arrival-time detection: the ability to detect the arrival and obtain an accurate measurement of the time-of-arrival of a signal sent by an emitter. The detection technique being utilised does not only influence the localisation accuracy, but also the sensitivity of the receiver, which has an impact on the range covered by the receiver.

## II. Background

When designing a wildlife tracking system, it is important to identify the desired characteristics, and to balance that out with the design constraints. In doing so the critical characteristics needs to be retained, while the inessential requirements are kept within acceptable limits.

An example of a sought-after feature is to keep disturbance to the natural environment of the animals to a minimum. This necessitates the amount of installed infrastructure to be minimised, and the communication range of receivers and transmitters to be maximised.

Many of the design constraints are interrelated. The electronic tag should ideally outlive the animal to which it is attached, since capturing and anaesthetising a wild animal to replace a tag is costly and dangerous. Tag replacement also counteracts the objective of minimising disruption to the animal. Tag lifetime can be improved by increasing the capacity of the tag's battery, but that will typically increase the size and mass of the tag, which can be intrusive to the animal. An alternative is to reduce the transmission power, but that would oppose the objective of increasing the communication range.

The relative importance of design constraints is not the same for all animals. For example, tracking accuracy, tag size and tag mass are less of a concern for large animals such as elephants. Although not ideal for research, an accuracy of 50 metres would be acceptable to locate a rhinoceros. Small animals such as birds can be easily captured, but tag size and mass are major constraints. Another appealing feature to consider in a design is the ability to be easily adapted for different animals and different applications.

The cost of a tracking system is a significant design constraint that has to be considered, both for commercial and scientific use. Regulations also impede the design. For instance, the transmission frequency and transmission power levels have to comply with ICASA regulations.

The degree of animal and environmental disruption, communication range, tag lifetime, cost and flexibility are all design constraints that must be carefully balanced when designing a wildlife tracking system.

## III. Related work

MacCurdy *et al.* researched and demonstrated the feasibility of employing multilateration for automatic wildlife tracking [6,7]. They designed a prototype system that uses TDoA measurements to track large numbers of animals over medium ranges up to 5 km. They used small, low-cost, long life transmitters and Digital Signal Processor (DSP) chips for signal processing at the receivers.

The approach used by MacCurdy *et al.* is similar to the acquisition mode of a GPS system [5,8]. Each tag transmits an 11-bit PN sequence chosen from a family of Gold codes at a chip rate of 1 MHz, modulated with BPSK onto a 140 MHz carrier. The receivers continuously seek matches for the 2 MHz wide signal with a real-time matched filter detector implemented with FFT and IFFT operations on a DSP. When a match is detected through the cross correlation between a template and the received signal, the event is timestamped and reported to a server. The server uses the difference in time at which a signal arrived at three or more receivers to determine the position of the emitter.

MarCurdy's system has a few shortcomings. They reported a range of 3–5 km when the tag is placed 1 m above the ground. A spacing of 5 km between receivers (base stations) is prohibitively dense for covering a large geographical region such as a game reserve or national park. Another shortcoming is that the large size and high power consumption of their receiver system can be intrusive to the environment.

## IV. Proposed research

Research is being conducted to compare arrival-time detection techniques that can be used for wildlife tracking in a multilateration system. The primary objective is to design, implement, test and compare a detection technique which improves upon the range of existing techniques. The secondary objectives are to maintain a reasonable level of accuracy, power consumption, cost and flexibility.

## V. Research methodology

Existing literature will be analysed to explore the complications and characteristics involved in the design of a multilateration system, to examine existing detection techniques and compare their performance, and to develop skills and become familiar with technologies needed to design and construct prototypes for experiments.

Numerical simulations will be executed to set a theoretical baseline for the performance of a detection technique in terms of range, accuracy and power consumption, and to evaluate the impact of practical considerations such as clock frequency offsets on the performance.

Laboratory experiments will be performed as initial tests and to establish the feasibility of solutions. An example of a lab experiment is to build a prototype transmitter using a development board or an existing platform, and to capture the signal with a high-end digital storage oscilloscope. The data can then be processed with high-level software on a PC and compared to the simulation results.

Lastly, field tests will be conducted as final evaluation of the performance of detection techniques in physical conditions.

Development kits and existing hardware platforms will be utilised for experiments where possible. Various boards and components are available from the industry partner, YRless International. Utilising their boards has the benefit that existing infrastructure such as base stations and tags can be utilised.

## VI. Conclusion

Research being undertaken to study arrival-time detection techniques was outlined. Related work by MacCurdy *et al.*, which will be used as baseline for the research, was briefly described.

The problem we aim to solve is to compare arrival-time detection techniques for wildlife tracking in a multilateration system, and to design, implement and test a detection technique that improves upon the range of existing systems. A brief background of the research was given in order to substantiate the relevance and actuality of the problem.

Although the research is focused on wildlife tracking, it can be extended to other use cases with similar constraints, such as asset tracking.

## VII. Acknowledgements

Financial assistance towards this research is provided by the National Research Foundation (NRF) through the *Innovation Master's Scholarship* programme and *The Technology and Human Resources for Industry Programme* (THRIP), as well as through the Telkom Centre of Excellence (CoE) at the North-West University. Opinions expressed and conclusions arrived at, are those of the authors and are not necessarily to be attributed to the NRF or the Telkom CoE.

## References

[1] S. Wang, J. Min, and B. K. Yi, "Location based services for mobiles: Technologies and standards," in *IEEE international conference on communication (ICC)*, 2008, pp. 35–38.

[2] C. James, "Multilateration: Radar's replacement?" *Avionics magazine*, vol. 31, no. 4, p. 30, 2007.

[3] H. S. A. Malleson, "The decca navigator system on d-day, 6 june 1944, an acid test," *Naval Science*, vol. 7, 1985.

[4] H. Lee, "A novel procedure for assessing the accuracy of hyperbolic multilateration systems," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. AES-11, no. 1, pp. 2–15, Jan 1975.

[5] R. B. MacCurdy, R. M. Gabrielson, and K. A. Cortopassi, "Automated wildlife radio tracking," in *Handbook of Position Location*, S. A. R. Zekavat and R. M. Buehrer, Eds. Hoboken, NJ, USA: John Wiley & Sons, Inc., Sep. 2011, pp. 1129–1167.

[6] R. B. MacCurdy, R. M. Gabrielson, E. Spaulding, A. Purgue, K. A. Cortopassi, and K. Fristrup, "Real-time, automatic animal tracking using direct sequence spread spectrum," in *Wireless Technology, 2008. EuWiT 2008. European Conference on*. IEEE, 2008, pp. 53–56. [Online]. Available: https://instruct1.cit.cornell.edu/courses/eceprojectsland/PotentialProjectDocs/MicrowaveWeekPaper_Short_Sub.pdf

[7] R. B. MacCurdy, R. M. Gabrielson, E. Spaulding, A. Purgue, K. A. Cortopassi, and K. Fristrup, "Automatic animal tracking using matched filters and time difference of arrival." *Journal of Communications*, vol. 4, no. 7, 2009.

[8] M. Fantino, L. L. Presti, and M. Pini, "Digital signal processing in gnss receivers," in *Handbook of Position Location*, S. A. R. Zekavat and R. M. Buehrer, Eds. Hoboken, NJ, USA: John Wiley & Sons, Inc., Sep. 2011, pp. 975–1022.

**Schalk-Willem Krüger** received his Bachelor of Engineering degree in 2013 from the North-West University and is presently studying towards his Master of Engineering degree at the same institution. His research interests include radio tracking, signal processing and algorithm design.

# AntMAC: A Dynamic Control Channel Selection MAC Protocol Design for Cognitive Radio Ad Hoc Network

Henry O. Ohize[1] and Mqhele E. Dlodlo[2]
[1, 2] Department of Electrical Engineering, Faculty of Engineering and the Built Environment,
University of Cape Town, P. Bag X3, Cape Town 7700
Tel: +27-61-699-6694[1], +27-21-650-3441[2]
Email: ohzhen001@myuct.ac.za[1], mqhele.dlodlo@uct.ac.za[2]

***Abstract*- In this paper, we present an overview of the work-in-progress on the design of a newly proposed MAC protocol called AntMAC for Cognitive Radio (CR) application. Our design is aimed at improving the convergence time associated with selection of a control channel for Ad Hoc CR operation. In particular, we adopt the use of the bio-mimicry concept to develop swarm intelligence based mechanism for the selection of control channel in a heterogeneous, spatial, and time varying spectrum environment, with no pre-existing infrastructure. The long term goal of this work aims at improving Mean Time To Rendezvous (MTTR), throughput, and link utilization as compared to other existing designs.**

**Index Terms— Cognitive Radio, MTTR, Rendezvous,**

## I. INTRODUCTION

With the recent surge in the deployment of wireless communication technology and the advent of newer bandwidth demanding multimedia services, spectrum scarcity is imminent. Yet still, Cisco VNI reported a further explosion of wireless traffic by 2018 [1]. The current static spectrum regulatory framework, wherein, an entire frequency band is exclusively allocated to specific services is inefficient. This static technique has failed to identify underutilized bands thereby resulting in apparent pseudo spectrum scarcity. Several spectrum surveys conducted across the world reveal spectrum underutilization in some licensed bands such as the very-high and ultra-high frequency bands (VHF & UHF) [2]. The trendy solution to spectrum scarcity is the usage of unlicensed devises by Secondary User (SU) in these underutilized bands. Federal Communication Commission (FCC) recently adopted this solution. The new field of research that enables this is the Cognitive Radio (CR) technology.

In CR, Dynamic Spectrum Access (DSA) is made feasible. However, spectrum access by opportunistic user (SU) must not interfere with the incumbent network, also referred to as the primary users (PU). Since DSA is still in its infancy stage, many research challenges remain, ranging from robust broad band sensing issues; access coordination that is, avoiding co-tier and cross-tier interference management issues among SUs and PUs, and Quality of Service issues, just to mention a few.

However, the main motivation for our work is to contribute towards developing a new and efficient MAC protocol for enhancing access coordination in CR DSA based Ad Hoc networks using Swarm intelligence based optimization techniques. By Swarm intelligence, we make reference to the mimicry of pheromone laying behavior of real ants to find the shortest route between their nest and a food source. This technique is popular in literature as Ant Colony Optimization (ACO).

Several MAC protocol designs have been proposed for centralized network architectures such as the IEEE 802.22 and Dynamic Spectrum Access Protocol (DSAP) [3]. However, few have been designed for distributed networks also known as Ad Hoc Networks. The absence of a central coordinator such as an access point or base station and the heterogeneous nature of the spectrum environment with spatial, and time varying availability have contributed to the difficulty in designing such protocols. The introduction of fixed or preselected control channels in MAC protocol design has remained the favourite approach in most works; however, this has generated more challenging issues such as control channel's robustness to PU activities, coverage and control channel security [4]. The focus here is to improve the mean time to rendezvous (MTTR) for a dynamic control channel selection. By rendezvous, we refer to the process of convergence which occurs when each SU seeking to communicate with another must first converge on a common channel to establish a link.

From related works, Horine & Turgut in [5] developed an algorithm for initial network set-up in which nodes seeking to establish connection randomly select available frequency and then emit attention signals on chosen channels. We note that the process of rendezvous was not bounded in time with respect to their analysis. DaSilva & Guerrreiro [6] proposed the use of a predefined non-orthogonal sequence to determine the order for visiting potential channels towards achieving rendezvous. The predefined sequence involved a random permutation of $N$ available channels though no preference for rendezvous on best channel was obtained for MTTR. Similar approaches were adopted in [7, 8]. Also, authors in [6] derived a Modular Clock Algorithm (MCA) and a Modified MCA (MMCA) for solving the rendezvous problem and provided results to show some improvements. Consequently, in our work, we shall bench mark MTTR results of our developed design with that of some techniques mentioned above.

Our proposed MAC protocol design will be done in two phases; network setup phase and data phase. The network setup phase algorithm would involve neighborhood discovery through an intelligent channel hopping sequence. In the data phase, control information is first exchanged on the control channel and data are finally transferred on selected channels by the communicating CR pair.

## II. PROPOSED DESIGN

### A. Basic System Model Description

The entire spectrum band will be mapped into finite band spaces using a linear function. Here we shall consider the finite bands to consist of $n$ non-overlapping orthogonal licensed channels. The PUs are considered holders of $n$ channels indexed uniquely as 1, 2…, $n$. The CR network will consist of $m$ nodes such that $m \geq 2$ nodes share the channels with the PUs in overlay architecture.

### B. Network Setup Phase

The use of ACO methods for coordination and organization of network parameters has been observed in related work [9]. By drawing similarities, we propose the use of such technique in achieving the goal of coordination among CR Ad Hoc Networks (CRAHN) nodes. Towards this, we next describe our proposed algorithm for rendezvous.

Let $A_{ij}$ be sets of available channels observed by $m$ nodes such that $i \in n$ and $j \in m$ and it holds true that a common channel i.e $\bigcap A_{ij} \neq \emptyset$ exist. A system rendezvous occurs if all $m$ nodes seeking to communicate converge on one of the common channel$i \in n$.

Our proposed ACO method deposit pheromone trails on the nodes of quality paths to reinforce the most promising channel in the channel hopping sequence. To achieve this, we present our proposed algorithm in Figure 1.

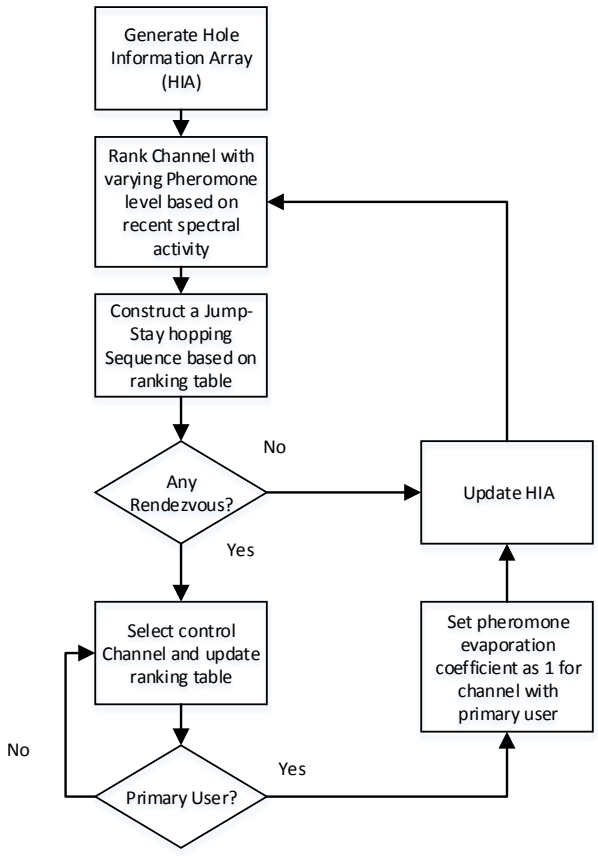



Figure 1: AntMAC Dynamic Hopping Algorithm

### C. Data Phase

Once a common control channel has been established, control information such as request to send RTS, request channel list RCL and clear to send CTS can then be exchanged for the purpose of band aggregation. Data are exchanged over a converged channel.

## III. ANTICIPATED RESULTS

We expect to improve on MTTR, throughput and link utilization. This would be achieved after development of the proposed algorithm and appropriate analysis. We note that PU activity will be modelled as a binomial random process and expect to limit the degree of PU interference based on the proposed designed algorithm.

## IV. CONCLUSION

This paper presents a work-in-progress on the design of a novel AntMAC protocol for CRAHN. The proposed scheme will introduce channel aggregation as a function of slot time, packet size and bandwidth. This will be developed and analyzed using MATLAB and implemented in OPNET. Our design is intended to improve the Mean Time to Rendezvous (MTTR) of an Ad Hoc CR system with respect to other known approaches.

AKNOWLEDGEMENTS

This research is supported by Telkom South Africa, Jasco/TeleSciences, and the Department of Trade and Industry/National Research Foundation/Technology and Human Resources Programme (DTI/NRF/THRIP), and Federal University of Technology, Minna Nigeria.

## REFERENCES

[1] Cisco, 2014. *"Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2013-2018."* Cisco White Paper

[2] Valenta V., Marsalek R., Baudoin G., Villegas M., Martha S., & Fabien R., 2010. "*Survey on Spectrum Utilization in Europe: Measurement, Analyses & Observations."* Proceedings of the 5th international Conference on Cognitive Radio & Oriented Wireless Networks & Communications. 9-11 June 2010. Cannes. Pp 1-5.

[3] Claudia C., Kaushik, R. C. 2009. *"A survey on MAC protocols for Cognitive Radio Networks*." Ad Hoc Network Journal, Elsevier, ISN 1570-8705. Pp 1315-1329

[4] Brando F. Lo (2011). *"A survey of common control channel design in cognitive radio networks."* Physical Communication Journal, Elsevier, ISN 1874-4907. Pp 26-39.

[5] Horine B. & Turgut D., 2007. *"Link Rendezvous Protocol for Cognitive Radio Networks."* Proceedings of the 2nd International symposium on New Frontiers in Dynamic Spectrum Access Network. Pp 444-447 Dublin

[6] Thesis N. C., Thomas R. W. & DaSilva L. A. 2011 *"Rendezvous for Cognitive Radio."* IEEE Transactions on Mobile Computing. 10(2). Pp 216-227.

[7] Lin A., Liu H., Chu X. & Leung Y. W. 2011 *"Jump-Stay Based Channel-hopping Algorithm with Guaranteed Rendzvous for Cognitive Radio Networrks"*. Proceedings of the IEEE INFOCOM. Pp2444-2452. Shanghai.

[8] Chang G., Teng W., Chen H. & Sheu J. 2014 *"Novel Channel-Hopping Schemes for Cognitive Radio Networks"* IEEE Transaction on Mobile Computing. 13(2). Pp407-421.

[9] Qian H., Ping Z., 2012. "Dynamic Channel Assignment using Ant Colony Optimization for Cognitive Radio Networks" Vehicular Technology Conference (VTC Fall), 2012 IEEE Conference. Pp 1-5.

## BIOGRAPHY

**Henry Ohize** received his undergraduate degree in 2004 from Abubarkar Tafawa Balewa University Bauchi, and Master Degree in 2010 from Federal University of Technology, Minna both at Nigeria and He is presently studying towards his PhD at the University of Cape Town. His research interests include Wireless Sensor Networks, Cognitive Radio, Software Defined Radio, and Long Time Evolution Networks.

# The design of a routing protocol for an intermittent wireless ad-hoc mesh network for animal tracking

Charles J. van der Spuy and Albert S. J. Helberg
School of Electrical, Electronic and Computer Engineering
North-West University, Private Bag X6001, Potchefstroom 2520
Tel: +27 18 2991111, Fax: +27 18 2992767
email: {CJ.VanDerSpuy, Albert.Helberg}@nwu.ac.za

**Abstract- Sensors that are used to track wildlife, e.g. rhinos, are expensive to attach and replace. For this reason, it will be desirable to replace the sensors as seldom as possible. The sensors must therefore be energy efficient, limiting the transmission power and reducing the connection range. A sensor network implementing relay drones can be utilised to perform the power draining operations and extend the connection range. This paper contains the work in progress on the design of the routing protocol controlling the data transfer in this drone network.**

**Index Terms— animal tracking, drone network, energy efficiency, routing protocol**

## I. INTRODUCTION

Rhino poaching has become a crisis in South Africa. From 2008, numbers of poached rhinos have increased excessively. Statistics according to StopRhinoPoaching [1] are displayed in the table below.

**Table I-1: Rhino poaching statistics in South Africa**

| Year | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014* |
|---|---|---|---|---|---|---|---|
| Quantity | 83 | 122 | 333 | 448 | 668 | 1004 | 558 |

**As at 9 July 2014*

To combat rhino poaching, this research will focus on the tracking of rhinos using electronic equipment.

YRless International [2] uses strap-on sensors to track rhinos. The circumference of the neck of a rhino is similar to the circumference of its head; for this reason, the sensor can't be strapped around the neck of a rhino. As a result of the growing horn, if the sensor is placed into the horn, the horn may break off or the sensor is grown out in a few years. The sensor is consequently strapped onto the leg of the rhino, reducing the sensor's communication distance.

It is costly to find and dart a rhino to attach the sensor, and therefore has to be done as few times as possible. Energy efficient sensors with reduced connection ranges are used to increase the battery life, extending the time before the sensors have to be replaced.

YRless experienced that the placement and energy constraints of the sensors reduce the maximum communication distance to 800m. A system that can provide a greater coverage (covering an area of a game reserve) is required.

The project's research group proposed a system that YRless can implement, as follows:

- Make use of relays that will enable the sensors to connect to a base station over a sufficient distance.
- Implement it as a mobile relay network. In a stationary approach, a relay must be installed every 800 metres resulting in an unappealing sight. A solar rechargeable drone[1] used as a relay will also be able to achieve a higher altitude, giving it a greater connection area.
- Drone landing spots are determined such that neighbouring drones can communicate when hovering above landing spots.
- The positioned drones will track the rhinos and give real-time distress signals to the base station.
- The drone network is reconfigured every night and drones are repositioned to form the network topology.
- To save energy, not all the drones will be airborne when transferring data. This is a significant constraint on the network that causes intermittent connectivity between nodes.

## II. PRELIMINARY MODEL

The execution of the analysis of the proposed drone coverage system is currently in progress to determine which parameters in the operation mode of the system will impact the design and selection of routing protocols.

By considering the constraints and characteristics of the above mentioned proposed system, one scenario is:

- Once the drones are positioned on their landing spots for the first time, they will ascend to be able to connect.
- Once all of them are airborne, the network will be configured based on a connection oriented approach. The drones will identify all the possible paths to the base station and store the possible 'next hop nodes' in their routing tables.
- After the network is configured the drones will return to the ground.
- If a drone has data to transmit, it ascends.
- Once airborne, the drone will try to connect to the next drone ('next hop node') according to the first entry in its routing table. If a connection can be established, data will be transferred to the 'next hop node'.
- Once a link breakage occurs the drone attempts to connect to the next drone, according to the second entry in its routing table. This process will continue until the data reaches the base station.

## III. PRIOR ART

Existing protocols were classified to identify what their constraints are when applied to the scenario described above. Figure III-1 depicts this classification; with protocol constraints indicated with X's.

[1]The complex problem of deploying drones/UAV's in a wildlife scenario is not considered part of the scope of this study.

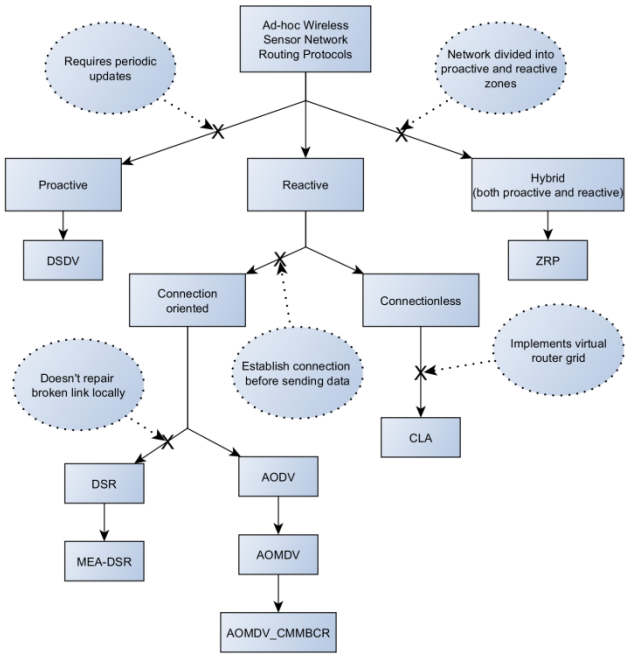



**Figure III-1: Well-known ad-hoc wireless sensor network routing protocols**

## IV. Problem Statement

Existing protocols cannot be implemented in the proposed drone coverage system without modifications. Wireless sensor network protocols, such as AODV [3], DSR [4], DSDV [5] and ZRP [6] require an established connection between the source and destination to be able to transmit data. It is therefore part of the study to determine how the system can be modified and how protocols can be combined and/or modified to overcome the problem.

## V. Research Objectives

### A. *Primary objective*

Design a routing protocol that will operate in the proposed drone coverage system. During the design one or more existing protocols will be modified and/or combined.

### B. *Secondary objectives*

- Simulation model
  Models will be constructed that incorporate the modified and/or combined protocols into a simulated animal tracking environment. These simulation models will allow one to determine the operating parameters and ideal performance of the modified and/or combined protocols in the proposed drone coverage system.

- Model comparison
  Simulation models will be evaluated against each other using parameters such as bandwidth, throughput, latency (delay) and jitter.

- Recommended protocol
  A protocol will be recommended that will be suitable for the proposed drone coverage system according to its performance in the simulation and comparisons.

## VI. Conclusion

From the scenario stated in section II it can be deduced that the following parameters must be included in the new protocol:
The physical layer must be connectionless because while data is transferred in one part of the network, nodes will not be connected in another part of the network even though they are on the same logical path. A connection oriented approach can be implemented into the network layer because a logical path is created before data is sent. In conclusion, logical paths will be established during configuration, but when the data is transmitted, it will try to follow the logical paths one hop at a time.

From Figure III-1 and the discussion above we see that an appropriate protocol has the characteristics of a reactive protocol. Both the connectionless and connection oriented approaches will be considered as part of the new protocol. In consequence of the fact DSR doesn't repair a broken link locally; AODV is a more appropriate starting point. We can attempt to integrate some parameters of connectionless protocols into AODV to produce a preliminary model.

## VII. Acknowledgment

The authors gratefully acknowledge the financial support of the Technology and Human Resource for Industry Programme (THRIP) and the Telkom Centre of Excellence (CoE) at the North-West University.

## VIII. References

[1] StopRhinoPoaching. (2014, April) StopRhinoPoaching. [Online]. http://www.stoprhinopoaching.com/statistics.aspx

[2] YRless International. (2014, April) YRless International: Animal Tracking. [Online]. http://www.yrless.co.za/animals/animalsGI.html

[3] M.K. Marina and S.R. Das, "On-demand Multipath Distance Vector Routing in Ad Hoc Networks," in *IEEE International Conference on Network Protocols*, Cincinnati, 2001, pp. 14-23.

[4] D.B. Johnson, D.A. Maltz, and Y.C. Hu. (2003, April) The Dynamic Source Routing Protocol for Mobile Ad Hoc Networks. Internet-Draft.

[5] C.E. Perkins and P. Bhagwat, "Highly Dynamic Destination Sequence-Vector Routing (DSDV) for Mobile Computers," in *SIGCOMM '94 Proceedings of the conference on Communications architectures, protocols and applications*, New York, October 1994, pp. 234-244.

[6] Z.J. Haas, M.R. Pearlman, and P. Samar. (2002, July) The Zone Routing Protocol (ZRP) for Ad Hoc Networks. draft-ietf-manet-zone-zrp-04.txt.

This work is based on the research supported wholly by the National Research Foundation of South Africa (Grant specific unique reference number 90087). The Grantholder acknowledges that opinions, findings and conclusions or recommendations expressed in any publication generated by the NRF supported research are that of the author(s), and that the NRF accepts no liability whatsoever in this regard.

**CJ van der Spuy** received his undergraduate degree in 2013 from the North-West University and is presently studying towards his Master of Engineering degree at the same institution. He is a member of the TeleNet Communications Research Group at NWU, and his research interest is wireless sensor network routing protocols.

# A Comparison of Wireless Sensor Network Routing Protocols on a Low Cost Prototype Experimental Test Bed

John G. Moutzouris and Ling Cheng
Department of Electrical and Information Engineering
University of the Witwatersrand, 1 Jan Smuts Avenue, Braamfontein 2000
Tel: +27 73 0090536, Fax: +27 86 5857443
Email: john.moutzouris@students.wits.ac.za, ling.cheng@wits.ac.za

**Abstract-An investigation in determining characteristics that enable certain wireless sensor network routing protocols to perform better and worse than others in specific applications and network situations is presented. Wireless sensor routing protocols will be implemented on an experimental test bed. The design and implementation of a low cost wireless sensor node is briefly discussed. A brief background of wireless sensor networks and WSN routing protocols is provided.**

**Index Terms—Wireless Sensor Network, Routing Protocols, Low Cost Test bed, WSN**

## I. INTRODUCTION

Advances in technology have allowed microprocessor and wireless devices to not only improve in terms of performance and size, but have also allowed them to become more energy efficient and low cost. All these factors have allowed Wireless Sensor Networks (WSNs) to become more feasible for use in a variety of different applications. From detecting fires in the forests of Cyprus[1], to the idea of a smart city where sensors everywhere report on almost anything; from the quality of air in various locations to the detection of any cracks or physical damage of buildings and city infrastructure.[2]

WSNs consist of a number of interconnected nodes, each node sensing and reporting on the physical environment it is in. Nodes coordinate with their neighboring nodes in order to get their data through the WSN to the data sink, where the data can be aggregated and computed on. Routing protocols are designed and implemented for WSNs to coordinate communications in the network. Many routing protocols have been designed, each having distinct strengths and weaknesses in different situations and network structures.

While development platforms are readily available (in the form of kits) from a number of vendors, they are typically very expensive. Development kits usually only include 6-8 nodes, making it infeasible to procure a large number of nodes, resulting in a rather small and limited WSN test bed. The design and development of a low cost, modular WSN node would allow researchers to put together WSNs consisting of many more nodes, allowing for larger networks to be experimented on. This will allow for more accurate and reliable results to be obtained for the better understanding of the physical constraints, limitations and behaviors of WSNs.

## II. WIRELESS SENSOR NETWORKS

A WSN is a network of autonomous sensor units, cooperatively sensing and gathering information of the conditions or events in the physical environment they are sensing. These sensor units send the relevant data to a data sink in the network, where data is collected from the entire network of sensor nodes and is used in the generation of a digital virtual model of the physical world. WSNs could consist of only a few tens of nodes or possibly even many thousands of nodes spanning over very large distances.

WSNs usually do not consist of any network infrastructure, such a routers or switches, nodes connect to and coordinate communications with one another in order to transfer data across the network. Nodes are also usually placed at points of measurement or convenience. This could often result in the physical location of the WSN nodes to be random and unstructured. WSN nodes usually connect up to one another in a mesh network topology, provided the nodes are in wireless range of one another.

Wireless sensor nodes consist of a microprocessor, which is usually low power and can only perform limited signal processing and computation on the output of connected sensors. As the name suggests, wireless sensor nodes connect to one another wirelessly, the microprocessor is connected to and controls a wireless transceiver. Wireless sensor nodes are also able to store a limited amount of data, usually on an additional memory module is connected to the microprocessor. Wireless sensor nodes are often deployed in remote or hard to access locations and are often powered by limited power sources such as batteries.

WSN nodes have limited resources; primarily limited power supply, computation power and storage. It is very important that sensor nodes conserve there energy as best as possible. For this reason, sensor nodes cannot be transmitting and receiving continuously. Additionally, nodes cannot spend too much time performing large computations.

Wireless communication is very expensive for a node as it consumes a considerably large amount of energy compared to that of the microprocessor which consumes an order of magnitude or two less energy. To aid energy conservation, but maintain network coordination, wireless sensor nodes rely on network routing protocols that are usually developed for efficient communications and low energy consumption.

## III. ROUTING PROTOCOLS

Routing protocols in WSNs define the how WSN nodes communicate with one another. Routing protocols are designed to maximize the use of the nodes' limited resources.

Due to factors, such as the limited power source, nodes could occasionally 'die' or drop out of the network. Routing protocols therefore have to accommodate for this and be able to cater for changing network environments, allowing the network and nodes to reconfigure on the occurrence of these events. Routing protocols need to allow the network to be self-organizing, allowing the network to adapt to dynamic changes over time.

WSN Routing protocols are categorized into four main categories [3, 4]; flat, hierarchical, location-based and data-centric routing protocols. Flat routing protocols make use of a flat addressing scheme, therefore, all nodes in the network play an equal role. Hierarchical protocols cluster groups of nodes, allowing cluster heads to perform some data aggregation, reducing the amount of data needd to be sent across the network, saving energy and improving scalability. Nodes in hierarchical routing protocols therefore have different roles in the network. Location-based protocols send data to particular regions in the network, based on the location of the nodes. Data-centric protocols save energy by allowing intermediate nodes to aggregate data as it is sent through the network to the data sink, thereby reducing the number of data transmissions necessary.

## IV. PROPOSED RESEARCH AND METHODOLOGY

This study will involve the design and development of a WSN node. This WSN node is being designed for the purpose of protocol design and implementation. Nodes will be able have extensive debugging and batch firmware updating capabilities. The firmware of the node will provide it with capabilities such as software defined network structures and multi-channel communications allowing for the defining of a particular network structure in software, making it easier to evaluate protocols in more extensive testing environments. The most important aspect of the nodes is that they are going to be low cost, which will be achieved by using low cost components such as an 8-bit microcontroller and 2.4GHz wireless transceiver. The prototypes will be evaluated by students in the WSN research group at the University of the Witwatersrand, in order to verify and provide feedback on the design of the WSN development nodes.

Using the proposed low cost experimental test bed, a number of existing WSN routing protocols will be implemented on specifically designed network structures. The performance of these different routing protocols will then be evaluated in various network structures in order to identify their strength and weaknesses.

This study will focus on the evaluation of hierarchical and data-centric routing protocols such as SPIN (Sensor Protocols for Information via Negotiation), Rumor Routing, LEACH (Low Energy Adaptive Clustering Hierarchy), PEGASIS (Power-Efficient Gathering in Sensor Information Systems), etc. While these protocols may be more complicated, they allow for highly scalable WSNs while attempting to minimize energy usage in order to prolong the lifetime of the nodes. During the evaluation of the protocols, performance indicators such as energy usage, network scalability, etc will be obtained. These indicators will highlight the strengths and weaknesses of these protocols. The study will then attempt to correlate these strengths and weaknesses to the design features of the protocols in question. This will allow for the identification of certain characteristics in the design of a protocol that allow it to perform better or worse in particular situations.

Work completed so far includes the development of an initial prototype. This prototype served as a proof of concept and aided in the discovery of a few issues and changes in requirements of nodes. The hardware design of the second, tailored, prototype is almost complete and preparations are being made for construction of a limited number of these prototypes. Research into a number of different routing protocols is being conducted in order to implement the protocols on the prototype test bed.

## V. CONCLUSION

By designing and developing an extensible, modular and low cost development platform it becomes feasible for many research institutions and individuals, even with very limited budgets, to obtain a suitable number of nodes to create a WSN. By highlighting the design features that make a protocol better or worse than others in certain situations can aid future researchers in the design and developments of WSN routing protocols, helping to better understanding the physical limitations of WSNs and ultimately aiding to further advance the development of WSN technology.

## REFERENCES


[1] Andreou, Panayiotis G., et al. "FireWatch: gis-assisted wireless sensor networks for forest fires." Scientific and Statistical Database Management. Springer Berlin Heidelberg, 2012.

[2] Libelium. Libelium World. http://www.libelium.com/es/smart_cities/. Last accessed: 02/08/2014

[3] Hong, Xiaoyan, Kaixin Xu, and Mario Gerla. "Scalable routing protocols for mobile ad hoc networks." Network, IEEE 16.4 (2002): 11-21.

[4] Singh, Shio Kumar, M. P. Singh, and D. K. Singh. "Routing protocols in wireless sensor networks–A survey." International Journal of Computer Science & Engineering Survey (IJCSES) Vol 1 (2010): 63-83.



**John G. Moutzouris** has a BSc in Electrical Engineering (Information Engineering option) from the University of the Witwatersrand. He is currently pursuing an MSc in Electrical Engineering. His research interests include telecommunications and network and information security.

# WORK IN PROGRESS:

# MANAGEMENT

# Application-based Network Policy Management for Machine-to-Machine Devices

Nyasha A. Mukudu and Neco Ventura
Department of Electrical Engineering, University of Cape Town
Rondebosch, Cape Town, South Africa
Tel: +27 21 6502813, Fax: +27 21 6503465
{nyasha, neco}@crg.ee.uct.ac.za

**Abstract- Machine-to-machine communication refers to communications between two or more entities that occurs without the need of human intervention. Advances in automation have led to many diverse applications being developed that take advantage of M2M services. These applications and their associated devices present new challenges to operators in terms of effective use of network resources and supplying the required services. This work investigates the use of Virtualization and Software-Defined Networking to enable efficient policy and resource management between different applications sharing the same operator network.**

**Index Terms—Machine Type Communication, Machine-to-Machine, Software-Defined Networking, Virtualization**

## I. INTRODUCTION

Machine-to-Machine Communication (M2M) is used to enable a diverse range of applications across different domains, for example Smart Metering, Asset Tracking, Remote patient Monitoring and Fleet Management. These applications require various network and device functions in order to deliver service. Unfortunately for the application developers, modern networks are optimized for Human-to-Human communication, and may require some modification to meet the specific application's needs.

The number of M2M connections are expected to grow rapidly over the next five years [1]. This growth will present an opportunity for Mobile Network Operators (MNO) to increase their profits. To fully take advantage of this opportunity will require MNO to invest in infrastructure and platforms to provide Value-added Services to M2M applications. This will lead to an increase in capital expenditures (CAPEX) and operational expenditures (OPEX). The use of virtualisation technology would potentially reduce the expenditures and provides flexibility in delivering these services [3].

A lot of work has been carried out to define common network services that are required by M2M applications.

The rest of this paper is organised as follows: Section II focuses on related work, Section III presents the aims of this research and Section IV looks at the proposed architecture. Section V concludes the paper.

## II. RELATED WORK

### A. M2M Standards

The European Telecommunications Standards Institute (ETSI) defines a horizontal framework to manage the complexity of multiple M2M Applications. It gives a description of service capabilities that a network may present to applications, without placing restrictions on how they should be implemented [2]. These service capabilities define what is required of the network. ETSI also has a list of general requirements that are needed by an M2M System.

### B. Software-defined Networking

Using software-defined networks allows enterprises to create their own private, virtual networks. This is especially of interest as it fosters innovation, and enables the creation of numerous secure slices that can accommodate individual data traffic related to a particular M2M application. These slices can be used to accommodate the individual needs of different data streams. The OpenFlow protocol enables third-party control of network routing by providing access to the forwarding plane of a network switch or route [3] .This would potentially allow M2M applications to reconfigure network devices to meet the specific requirements of the application and its associated devices.

### C. Network Function Virtualisation

Network functions virtualisation (NFV) provides a means of leveraging virtualisation technology to consolidate, many network equipment types onto high volume servers, switches and storage. Using NFV techniques potentially allows for targeted service introduction based on customer sets and the services can be rapidly scaled up/down as required [4]. It would also make it possible to support multi-tenancy thereby allowing network operators to provide tailored services and connectivity for multiple applications co-existing on the same hardware with secure separation of administrative domains as suggested by [5]. Applications will be able to tailor the network to effectively meet its requirements, without interrupting other applications.

## III. RESEARCH GOALS

The aim of this work is to enable M2M Applications to customize the Network and gateway services to meet the needs of their devices by:

A. Providing a means for part of the network to be optimized to meet an Application's requirements with limited impact on other applications using the same network.
B. Providing enhanced security and access control for Applications and devices by investigating possible security vulnerabilities and providing solutions
C. Providing a mechanism to monitor the use of the network and devices by Applications

## IV. Proposed Architecture

The proposed high-level architecture is shown in Figure1. This architecture is based on the ETSI High level architecture for M2M [6]. The only addition is an operator managed platform in between the M2M applications and the M2M service Capabilities. The role of this layer would be to implement services associated with receiving the resource requests from the applications and determining the required services. This layer will also handle approving or denying request for services.
Some of the required management functions will be implemented through the use of the management plane and service capabilities in the network or gateway domains. Virtualisation and Software-defined networking will be used to create logically separated networks, to allow the applications to reconfigure an instance of the network domain to meet its needs.
An ongoing literature review is being carried out to ensure the design of a functional architecture that will ensure efficient resource allocation.

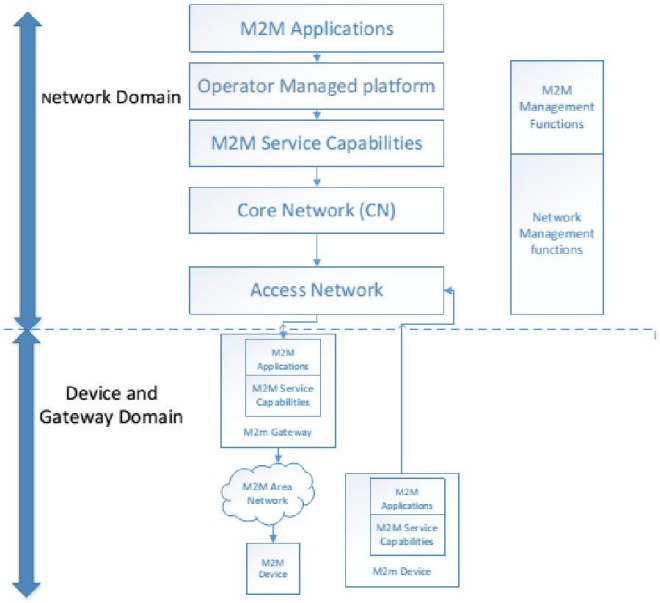



**Figure 1: Proposed High level architecture**

The proposed solution will be implemented using the OpenMTC platform, which is compatible with ETSI M2M standards [7] .

## V. Conclusion

M2M provides a massive opportunity for Network Operators to increase their revenues. Simply providing a connection from the Devices to a M2M server is no longer enough. It is now imperative to provide additional service capabilities to the applications.
We propose to design a means for Applications to adjust their networks to meet both the needs of the Applications and gateway devices.


## Acknowledgements

This research is supported by Telkom South Africa, Jasco / TeleSciences, and the Department of Trade and Industry / National Research Foundation / Technology and Human Resources Programme (DTI/NRF/THRIP).


## References


[1] Cisco. (February 5, 2014). "*Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2013–2018 "[white paper]*. Retrieved 31 July 2014 from http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.pdf
[2] ETSI Technical Specification RTS/M2M-00001ed211, *"Machine-to- Machine communications (M2M); M2M service requirements"*, July 2013
[3] S. Sezer, S. Scott-Hayward, P.K. Chouhan, B. Fraser, D. Lake, J.Finnegan, N. Viljoen, M. Miller and N. Rao, *"Are we ready for SDN? Implementation challenges for software-defined networks."* IEEE Commun. Mag., vol. 51, no. 7, Jul. 2013.
[4] ETSI, *"Network Functions Virtualisation – Introductory White paper"* Retrieved 31 July 2014 from http://portal.etsi.org/NFV/NFV_White_Paper.pdf
[5] Shafiq, M.Z.; Lusheng Ji; Liu, A.X.; Pang, J.; Jia Wang, *"Large-Scale Measurement and Characterization of Cellular Machine-to-Machine Traffic," Networking, IEEE/ACM Transactions on* , vol.21, no.6, pp.1960,1973, Dec. 2013
[6] ETSI Technical Specification RTS/M2M-00002ed211, *"Machine-to- Machine communications (M2M); Functional architecture"*, July 2013
[7] OpenMTC platform." [Online]. Available: http://www.open-mtc.org/index.html.



**Nyasha Mukudu** received his B.Tech (Hons) degree in Electronic Engineering from the Harare Institute of Technology, Zimbabwe in 2012. He is currently in his first year of study towards an MSc (Eng) degree in the Department of Electrical Engineering at the University of Cape Town. His research interests include Machine Type Communication, Internet of Things, Software Defined Networking and Network Function Virtualization.

# WORK IN PROGRESS:

# STANDARDS, REGULATORY & ENVIROMENTALS

# Analysis and Design of Wireless Mobile Charging System for mobile phones using Communication Networks

Frederick G. Kumi, Mqhele E. Dlodlo
Department of Electrical Engineering
University of Cape Town, P. O. Box 7701, Cape Town, Rondebosch
Tel: +27 21 650 4801, Fax: +27 21 650 2795
Email: kmxfre001@myuct.ac.za, Mqhele.dlodlo@uct.ac.za

**Abstract-Wireless Power Transmission (WPT), the technology to transfer power using electromagnetic (EM) waves, has been a topic of major interest in wireless communications for the past few decades**[1][2]**. Since communication is invaluable for human life, efficient charging system to power our communication equipment for constant communication availability is vital. For example; mobile phone charging in outdoors, covert or remote operations, medical implants, the internet of things technology (IoT)** [2] **are not easy to implement with cables. Many inductive and directive EM radiation technologies in literature do not address distance or mobility, or both. In this paper, the third type of WPT, non-directive far-field WPT, is the focus of this research, where receiver position varies in time. We propose a WPT system, through mathematical modelling of the mobile environment, rectenna design, and power management, to provide a reliably ubiquitous wireless charging system for mobile phones at arbitrary locations using the Communications Network. Our experiment is verified in MATLAB, FEKO, and PSim simulation environment.**

**Index Terms—wireless power transmission, Rectenna, FEKO, MATLAB, PSim, Inductive, and Far-field**

## I. INTRODUCTION

Wireless power transmission (WPT) is the technology of transmitting power from a source to a sink using EM waves. It begun in the works of Henrich Hertz and Tesla in 1899 [1] and can be implemented from Very Low Frequency (VLF) to Microwaves [3], [4]. Studies, however, show higher efficiency at 2.45GHz [1], [3]. WPT involves three steps: 1) DC power converted to RF power, 2) RF power transmitted with the help of antennas, and 3) RF power collected and converted back to DC [1], [3]. But why should this research merit attention?

Communication is a basic necessity for life. Therefore it is crucial that access to communication at any time and place – at home, on the street, at the beach, on the farm, or in the forest, just to name a few – are equally fundamentally important. In addition, it is a frequent occurrence that our cell battery runs out, and we turn to the nearest wall socket. It becomes quite embarrassing if this occurs in a location without a wall socket, or charge from our cars or computers. Moreover, today's society is developing toward creating smart environments in which a multitude of sensors interact to deliver information for Machine-to-Machine and Machine-to-Human Communications [2]. Also, the design of energy efficient systems that aim toward a low-carbon-emission society is essential to the implementation of the Internet of Things (IoT) technology. Within this context, wireless power transfer is an apt alternative to providing these devices with self-sustained operation. Related works of WPT system is given in section II, followed by a proposed work in section III. The conclusion follows on in section IV.

## II. RELATED WORK

[5]–[8] show some related works of inductive coupling, also called near-field WPT. Zhu et al in 2007 proposed the "Switched Beam Smart Antenna Technique" for far-field WPT to charge the cell phone. With electronically steerable beam-forming antenna technology, Zhu and colleagues address indoor concerns to a maximum distance of 5m [9]. Huiqing Zhai et al, in 2010, based their research on "Ultra-Wideband Retro-Reflective Beam-forming". They used multiple antennas distributed in space. Pilot signals are sent from the target device (mobile phone) to a transmitter. When the transmitter receives this signal the antennas in the vicinity jointly construct a focused EM beam onto the device (beam-forming) [10].

Finally, Md M. Biswas et al, in 2012, used high gain, large transmitting dish antenna (12m) and patch antenna arrays for power summation. Their research is based on Friis free space transmission equation $\frac{P_r}{P_t} = \frac{A_{er}A_{et}}{R^2\lambda^2}$. Given a transmit power of 20W, they achieved a received power of 2W, in a direct Line-of-Sight distance of 56m [11]. However, Friis equation is best used for free space line of sight propagation only, not multipath. [12]

It is clear, then, that none of the literature discussed addresses the issues of wireless fading and multipath channels, which is elicited from reflections, scattering, and diffractions. These results, therefore, can hardly be applied in real time.

## III. PROPOSED RESEARCH WORK

Using the equation;

$$PL(d)[dB] = \overline{PL}(d) + X_\sigma = \overline{PL}(d_0) + 10n\log\left(\frac{d}{d_0}\right) + X_\sigma \dots\dots\dots\dots (1)$$ and

$$P_r(d)[dBm] = P_t[dBm] - PL(d)[dB] \dots\dots\dots (2)$$

Where

$$\overline{PL}(d_0) = -10\log\left[\frac{G_tG_r\lambda^2}{(4\pi)^2d^2}\right] \dots\dots\dots (3)$$

Where

$$d \geq d_f \text{ And } d_f = \frac{2D^2}{\lambda} \dots\dots (4)$$

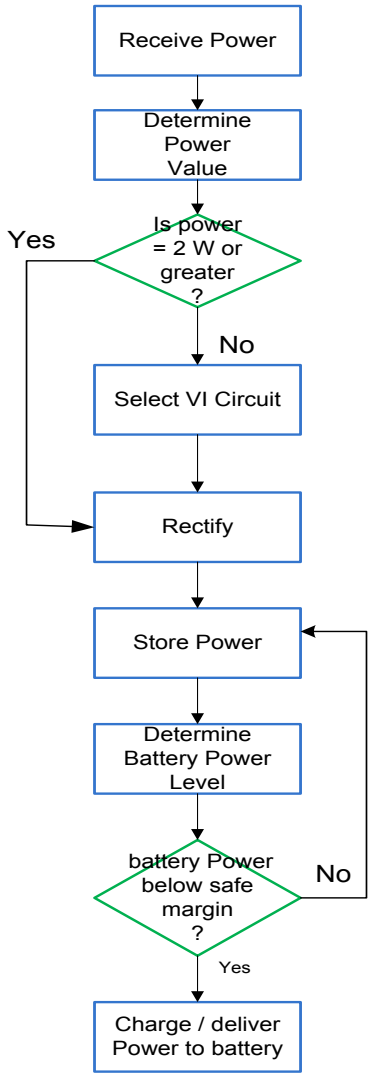
$d_f \gg D$ and $d_f \gg \lambda$ [12]

Where $G_t$, $G_r$, $\lambda$, $X_\sigma$, $d_f$, $D$, $d_0$, $PL(d)[dB]$, n are, respectively; transmit antenna gain, Receive antenna gain, wavelength, Zero-mean Gaussian distributed random variable with standard deviation, antenna far-field distance, antenna diameter, reference distance, Path loss in dB, and

the path loss exponent. The environmental impact (path loss) and other required parameters for experimental setup are determined.

### A. *System algorithm Design*

The proposed power "buffering" (storage) combined with Decision Logic system algorithm to provide ubiquitous wireless charging system, which addresses multipath and fading, is shown in figure 1 in the form of a flow chart.




## IV. CONCLUSION

This paper has shown some of the compelling reasons for pursuing research in WPT. It has also highlighted the challenges of WPT. The novel proposed algorithm, to tackle successfully both distance and mobility, makes it a promising solution. Thus far, the research project is under continuous steady progress with the aim of completion by the end of the year. It is our strong hope that, upon successful completion and implementation, this project will improve upon wireless communication and standard of living.

## REFERENCE


[1] William C. Brown, "The History of Wireless Power Transmission by Radio Waves," in *IEEE Transactions on Microwave Theory and Techniques*, Vol. 32, No. 9, pp. 1230 – 1242. September 1984, 1984.

[2] A. Boaventura, A. Collado, N. B. Carvalho, G. Alirio, "Optimum Behavior: Wireless Power Transmission System Design through behavioral models and eficient synthesis techniques," *IEEE Microwave Magazine*, Vol. 14, no. 2, pp. 26–35, March-April 2013.

[3] W. C. Brown, "The technology and application of free-space power transmission by microwave beam," *Proceedings of the IEEE*, vol. 62, no. 1, pp. 11–25, Jan. 1974.

[4] O. E. Maynard, W. C. Brown, A. Edwards, J. T. Haley, G. Meltz, J. M. Howel, "*Microwave Power Transmission System Studies,*" National Aeronotics and Space Administration, Tech. Rep. NASA-CR-134886, Vol. IV, pp. 1 – 236, 1975.

[5] X. Zhang, S. L. Ho, and W. N. Fu, "Quantitative Design and Analysis of Relay Resonators in Wireless Power Transfer System," *IEEE Transaction Magazine*, vol. 48, no. 11, pp. 4026–4029, Nov. 2012.

[6] W. P. Choi, W. C. Ho, X. Liu, and S. Y. R. Hui, "Bidirectional communication techniques for wireless battery charging systems & portable consumer electronics," Twenty-Fifth Annual IEEE Applied Power Electronics Conference and Exposition (APEC), pp. 2251–2257, Feb. 2010.

[7] E. Waffenschmidt and T. Staring, "Limitation of inductive power transfer for consumer applications," 13th European Conference on Power Electronics and Applications (EPE), pp. 1–10, 8-10 Sept. 2009, Barcelona.

[8] P. Manivannan, S. Bharathiraja, "Qi Open Wireless Charging Standard – A Wireless Technology for the Future," International Jounal of Engineering And Computer Science, vol. 2, no. 3, pp. 573–579, 2013.

[9] W. S. Jhong and W. J. Liao, "Serially-fed beam scanning antenna for wireless charging application," IEEE International Conference on Microwave Technology & Computational Electromagnetics (ICMTCE), pp. 297–300, 22-25 May 2011, Beijin.

[10] U. Huiquing Zhai, Helen K. Pan, and Mingyu Lu, "A practical wireless Charging System based on Ultra-Wideband Retro-Reflective Beamforming," in *IEEE* Antennas and Propagation Society International Symposium (APSURSI), no. 1, pp. 2–5. 11-7 July 2010,Toronto, CA.

[11] M. Biswas, U. Zobayer, and J. Hossain, "Design a Prototype of Wireless Power Transmission System Using RF / Microwave and Performance Analysis of Implementation," IACSIT International Journal of Engineering and Technology, vol. 4, no. 1, pp. 61–66, Feb. 2012.

[12] T. S. Rappaport, Wireless communications, Principles and practice, Ed, 2nd ed. Prentice-Hall, Inc., 2010.



**Frederick Gyampoh Kumi** received his undergraduate degree in 2011 from the All Nations University College (KNUST Kumasi, Ghana) and is presently studying for his Master of Science degree at University of Cape Town. His research interests include Wireless channels, electromagnetics, antennas, and circuits.