Inaugural lecture

Speech technology for development: working towards impact

by

Prof Etienne Barnard

Professor in the School of Information Technology
in the Faculty of Economic Sciences and Information Technology
at the Vaal Triangle Campus, North-West University

20 April 2012

Vaal Triangle Occasional Papers: Inaugural lecture 9/2012
Vanderbijlpark
2012

Printed by: The Platinum Press

North-West University

Vaal Triangle Campus

April 2012

ISBN 978-1-86822-613-9

Speech technology for development: working towards impact

Etienne Barnard

ABSTRACT: Providing people in under-served communities with information access could be of great social value. I discuss some of our attempts during the past decade to create systems based on speech technology which can be used for that purpose. The trajectory we followed during those attempts – from resource creation, through the development of basic technological building blocks, to the eventual deployment of several services aimed at information access – has been instructive from various perspectives. On the one hand, we now have a much better set of tools for the development of, especially, speech-recognition systems in under-resourced languages. On a more abstract level, these efforts underline the need to carefully choose a long-term objective which is diligently pursued – even if the need for occasional adjustment of the overall plan cannot be denied, such a long-term strategy seems like a crucial requirement for the attainment of the deep insights that drive innovation.

Contents

Speech technology for development: working towards impact

"I have always found that plans are useless but planning is indispensable."
D D Eisenhower

Information against poverty

One of the great revolutions of our time is the widespread realization that *information* is a commodity of immense value. In the same way that the industrial revolution caused a widespread change in the relative valuation of agricultural and manufacturing activities, we have more recently come to the realization that the possession of relevant and accurate information is often as valuable as substantial quantities of tangible goods. For a (slightly whimsical) demonstration of this point, consider Figure 1, which shows how often the words "gold" and "information" occurred in the wide variety of publications included in the *Google Books* collection between the years 1800 and 2000 ... Clearly, the (literate) world became a lot more interested in information from about 1940 onwards!

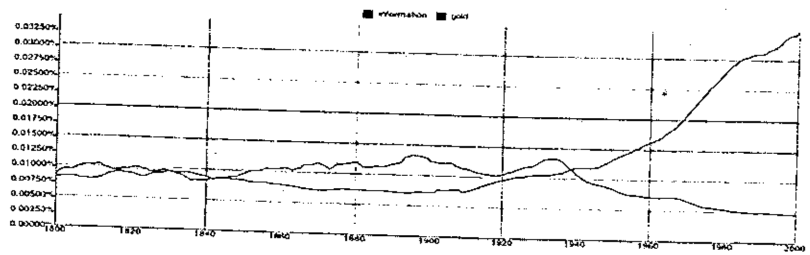

Figure 1: Frequencies of the words "information" and "gold" in a wide range of texts between 1800 and 2000

This understanding has significant implications for one of the oldest battles facing humanity, namely the struggle against poverty. It has long been understood that you can feed a man for a day by giving him a fish, but feed him for a lifetime if you teach him to use a fishing rod. However, the realization that this represents a systematic way to address poverty is much more recent -- the immediacy of tangible donations was not easily replaced by something as indirect as an instruction guide! The sea change effected by digital computers and the Internet has gone a long way to subverting that particular prejudice, and today there is a widespread understanding that one of the best ways to overcome poverty is to empower people with information.

Here in South Africa, where wealth inequality is about as bad as it gets, this understanding is both relevant and intuitively understandable. In our daily lives, we are frequently confronted with the fact that many of our fellow South Africans are unable to make the choices that would lift them out of poverty, since they do not have access to the appropriate information that would allow them to do so. "Appropriate information" can take a variety of forms, ranging from practical advice on agricultural matters and health care, through advice on training and job opportunities, to longer term investments in basic to advanced education.

Such a vision of improved information access to address the most pressing concerns of our country (and our continent) is very attractive, since it is forward-looking and supports many of the core values of our society. In addition, the massive growth in information technology during the past five decades provides a wide array of tools that can be used in support of that goal. The digital computers and computer networks referred to above have become much more powerful and affordable in this period, thus making widespread information access a reality for billions of people worldwide. Although this is far from a proven fact, I personally believe that information technology will eventually alleviate poverty to an extent that is not imaginable today.

This optimistic long-term view should not, however, blind us to the reality of millions of our fellow citizens who are currently prevented from gaining access to relevant and timely information because they are not computer users, or not connected to the Internet, or not fluent in English. It is not acceptable to tell these people – who certainly still constitute the majority of Africans and South Africans in particular – that they should await their turn. As a society, we should search for ways to accelerate the process whereby the benefits of information access are provided to the widest possible range of citizens.

Of course, this realization is not new, and there have been countless attempts to provide information technology to underprivileged communities in Africa, South America and elsewhere. By and large, these projects have been dismal failures: when people are struggling to make a living, it generally helps very little to provide them with a shiny computer system, regardless of how many Gigahertz and Gigabytes it has on display. Many are the computer laboratories in Africa, stocked with the latest in hardware and software, which sadly gather dust once the kind people from the donor agencies have made their way back to Europe or the US. Clearly, we need technologies that are more appropriate to the specific needs of the intended beneficiaries. Thus, we need to pay attention to the realities that prevail in the communities for whom we want to provide information access: most of the people in those communities are not familiar with sophisticated technologies, are often illiterate, and do not speak much English.

One of my goals is to discuss a group of technologies that are attuned to these realities. These are known as speech technologies, and their common characteristic is that they use computer algorithms to process human speech. The two most prominent examples of speech technology are *speech recognition*, which enables computer systems to understand human speech and *speech synthesis*, which generates speech based on, for example, written text. It is clear that such technologies address both the literacy concerns and issues related to technological sophistication, since they use speech to interface between people and computers. Since speech is easily transmitted (using widely-available telephone networks) these technologies hold the promise of providing information access to anyone who can speak and understand speech. The complexities of computer setup, Internet access etc. can be handled by dedicated service providers, and end users are not required to operate anything more complicated than a normal telephone.

Although this vision is very alluring, it presents one immediate problem: speech technologies are highly language dependent, and generally not available in the languages for which the social need is the highest. Thus, when my colleagues and I started pursuing the goal of speech-based information access about ten years ago, our immediate aim was to develop the nuts-and-bolts components that would lead to useable speech technologies in the official South African languages. In what follows below, I will describe the path that we have taken in this quest. However, before delving into those details, I first want to make a detour to examine the general prospects for large-scale technology interventions of the nature that I have been proposing.

What to expect when venturing into the unknown

The law of unintended consequences looms large over all human activities. We all know about those infamous rabbits that were introduced as a source of protein in Australia, and soon became a major pest with severe costs to agriculture and the environment. Or the way that prohibition in the US produced an explosive growth in organized crime. Conversely, many of the medicines the modern doctor's arsenal were discovered by happenstance – think of Alexander Fleming, who left a Petri dish uncovered in his lab. The resulting *penicillium* infestation ruined his preparation of bacteria, but lead to all the wonder drugs we now know as antibiotics. And of course, most unforeseen outcomes are not specifically good or bad, just unexpected – such as the entire suburban culture which resulted from the invention of the automobile.

Clearly, all human endeavours have to account for the slings and arrows of outrageous fortune – but this issue is particularly troubling for long-term technological planning, where both the process of development and the outcome of such development are subject to the influences of a wide range of unpredictable variables. In addition both of these classes of unknowns – those related to science and engineering of innovation, and those that describe the social impact of whatever is developed – come in a continuous range of severities. Roughly speaking, longer development times and larger differences between the affordances of the technology and those of the status quo, lead to a wider range of possible outcomes that may arise.

As we now consider the development of speech technology as a means of information access in underserved communities, we have to admit that our endeavour is fraught with fairly severe unknowns in both respects. There are approximately seven thousand living languages on earth (according to those who know about these things), and somewhere between twenty and a hundred of those have been studied in substantial detail. Hence, technology development for most of these languages is certain to be a long and wide-ranging task. Also, the shift in mind-set required of people with very limited exposure to technology, in order to utilize a disembodied and remote piece of technology in order to gain information access is large. Thus, I also want to view my narrative as a case study, relevant to the following question: *are the costs of such long-term technology projects justifiable, given all the unpredictability that we should expect during their execution?*

Anatomy of a speech-recognition system

Against that background, I now want to delve into the details of how current speech-recognition systems are constructed, and the particular challenges we had to address in order to make such systems work for the South African languages. Figure 2 below shows the main building blocks of such a system, with those components that require language-specific attention coloured in. The first module serves to capture the speech from whatever device the user is employing, such as a telephone or desktop computer. The result is a stream of numbers that describe the sampled speech signal – typically, around 16 000 samples are collected per second. These samples are then converted to an alternative representation in a process known as "feature extraction". The feature extraction process is somewhat abstract, involving several fairly sophisticated signal-processing steps. However, its purpose can easily be understood by considering the related task of computer vision. In vision, a digital camera represents an image as a table of numbers (analogous to the stream of numbers representing sampled speech), but for the purposes of recognizing objects in the image, it is much more convenient to work with representations which indicate that, for example, a patch of orange with radius 30 pixels occurs in the top left-hand corner of the image. Such a domain-specific, outcome-oriented transformation is what we mean by the term "feature extraction".

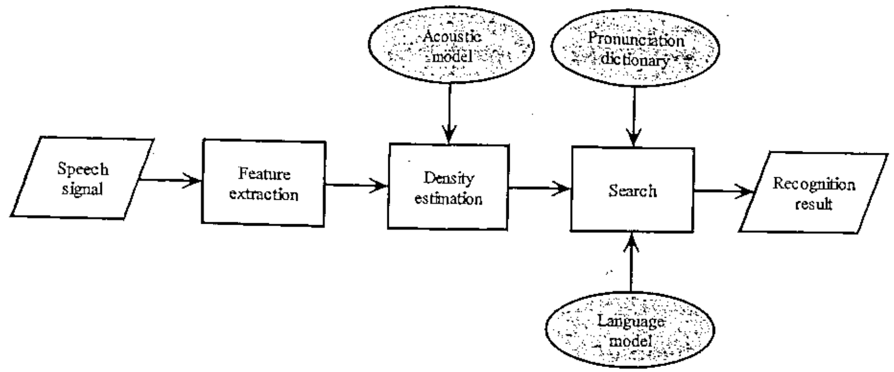

Figure 2: Block diagram of a basic speech-recognition system

The feature description produces a new stream of numbers, at a somewhat lower rate than those that occur in the sampled waveform – typically, we work with a hundred feature items per second. Whereas each sample consisted of a single number, each feature descriptor is a group of numbers (usually, 13 numbers per group), but the net result is a reduction in the amount of data that needs to be processed. More importantly, similar speech sounds will produce similar feature descriptors, so that we are able to tell – with some degree of uncertainty – which sound occurred just by looking at the feature descriptor at a given time.

To explain that more clearly, I briefly need to remind you about the building blocks of any speech sound. These basic blocks are known as *phonemes* -- they are closely analogous to letters in the written form of language: just as any written message can be constructed by stringing together a sequence of letters, all spoken utterances consist of sequences of phonemes*. Thus, the speech elements that can be deduced from the instantaneous feature descriptors are these phonemes. However, the uncertainty alluded to above means that we can only assign *probabilities* to each of the phonemes occurring – within such a short time window, for example, it is easy to confuse a "d" with a "t", for example – or an "m" with an "n". Hence, our next processing step – known as *density estimation* – produces a column of phoneme probabilities at every time step.

* This analogy is not perfect. Most importantly, whereas exactly the same set of 26 letters are shared by many languages, different languages tend to use different phoneme sets - I return to this point below.

Of course, users of a recognition system are not interested in phoneme probabilities -- we want to know which words had been spoken. Thus, we need to figure out the most likely word corresponding to a sequence of phoneme probabilities. This is a typical example of what computer scientists refer to as a *search* problem, and combines three types of information: the observed phoneme probabilities, the phonemic constituents of words (as expressed in *pronunciation dictionaries* -- see below), and the likely sequence of words that may have occurred at a particular time. These expectations play a surprisingly large role in spoken communication. In speech-recognition systems, a *language model* is used to represent such expectations. Thus, the search module produces the most likely sequence of words based on both the speech sounds that were heard and the sequences of words that are expected in the context of the utterance.

Do people use language models?

• • •

Although many aspects of our current approaches to speech recognition are not plausible psychologically, the important role of language models is also notable in human understanding. That is, we continually use cues from the context to interpret the speech that we hear. Ron Cole likes to use the example of "recognize speech", which is heard as "wreck a nice beach" if the listener is thinking of the ocean!

How to develop speech recognition for a "new" language

Based on this description, we are able to discuss the steps that must be undertaken in order to produce a speech-recognition system for a particular language. Importantly, we need to create a phonemic description of each word that we hope to recognize -- that is, we need to produce a pronunciation dictionary, which is a tabulation of the sequence of phonemes that corresponds to each of the words in our dictionary. However, before this can be done, we need to agree on the inventory of phonemes that occur in our target language, which is a surprisingly subtle task: whereas, say, the 26 letters of the Roman alphabet suffice for all of the European languages and many of the African languages, it turns out that virtually every language uses a distinct set of phonemes, and the exact details of these phoneme sets are subject to significant debate, even in relatively well-studied languages such as Sesotho. Hence, we have found that one needs to start from a standard linguistic reference, or if that is not available, discussions with at least a few independent linguists with detailed knowledge of the target language. From such sources, an initial phone† set is chosen; our goal is that all the words in the target

† Being an engineer, I am allowed to obfuscate on the distinction between phonemes and phones! The distinction does not play an important role in speech technology, so for now I will pretend that the terms are synonymous.

language must be describable in terms of these phones, but that there should be no redundancy in the phone set.

For the actual dictionary creation process, we need to obtain the sequence of phones corresponding to each of the words we wish to recognize. In the languages that we study, this invariably means that a human has to enter these phonetic descriptions, and we have found that a very systematic process is required in order to ensure that accurate and consistent pronunciations are generated. My colleague, Marelie Davel, has lead the development of a highly structured and efficient software package, known as DictionaryMaker, to assist in this process. The two main inputs to DictionaryMaker are the phone set, and a list of words for which pronunciations will be generated. The easiest way to create such a list is to extract the words from a large collection of text in the target language. Such a text *corpus* is useful for other reasons as well, as we will see below, so we spend much of our early effort when working on a new target language on gathering suitable corpora, from sources such as the World Wide Web, collections available at universities and publishing houses, and wherever else we can find large collections of text.

DictionaryMaker presents each of the words in our word list to a human transcriber in turn, and asks the transcriber to enter the phone string that corresponds to the "standard" pronunciation of the word. As the pronunciation of each word is entered in this way, the software employs a machine-learning algorithm to hypothesize a set of rules that are able to predict the pronunciation from the written form of the word. These rules are also applied to each new word that the transcriber has to process, giving the transcriber the option to either enter a new pronunciation or simply correct the automatic prediction. For languages with regular writing systems, such as those of the indigenous South African languages, the appropriate rules are learnt very quickly, and after a few hundred words the transcriber is generally only making small corrections to the predicted pronunciations (except for foreign or loan words, which inevitably require more manual intervention). Once this process has been completed, we have a list of words with pronunciations that have been verified by a human, but also a set of rules that can be used to predict the pronunciations of words that do not appear in the list.

So, we now have models for the phonetic decomposition of each word in our language; the next step is to develop models that represent the sounds corresponding to each of the phones -- that is, we need to compute the probability density functions of Fig. 1. One of the great success stories of speech recognition in the past 30 years is the development of increasingly capable automatic algorithms for computing those functions. Besides the components described above, the most important input to this process is an *orthographically transcribed speech corpus* -- that is, a collection of spoken utterances in the target language, along with the corresponding text representing the words that were spoken. Two characteristics of the speech corpus play a crucial role in the accuracy of the resulting density functions:

- The contents of the corpus must be *diverse*, containing the speech of a wide range of speakers, balanced for age, gender, and dialect, and also containing a wide range of spoken content.
- The corpus must also be *representative* of the target applications, as far as possible – thus, if we wish to recognize people speaking commands to an information-access application, the corpus should preferably contain numerous recordings of such commands, rather than of, say, spontaneous conversations or readings of newspaper articles.

These requirements suggest a laborious process of speaker recruitment, recording and transcription, and it certainly is true that the process of speech corpus development is the largest obstacle in the process of recognizer development. In wealthy countries, it is not unusual for tens of millions of rands to be invested in the development of useful speech corpora for a single language. In the developing world – our focus – we do not have the luxury of such large investments, in light of both our financial constraints and the diversity of languages spoken here. Consider that the single-language costs must be multiplied by eleven to cover all the official languages of South Africa!

The early work in our group did indeed suffer under the complexities of corpus development. We made a small breakthrough in discovering that a smaller group of contributing speakers can provide most of the benefit of speaker diversity, thus reducing the potential cost of corpus development to some extent. However, the costs of recruitment, recording and transcription still limited the size of corpora that could be developed economically, and we had decided on a compromise: accept that smaller corpora are the best that we will have available (at least initially), and therefore limit our speech applications to tasks that require smaller recognition vocabularies, since such vocabularies can be recognized with acceptable accuracy even if your training corpus is small.

This was the main trajectory of our work until mid-2009, when our entire mindset was changed by a number of conversations with Johan Schalkwyk and Pedro Moreno (who are speech scientists with Google in New York). Along with their colleagues, Johan and Pedro had developed an approach to data collection with so-called smartphones that promised to revolutionize speech data collection. The core of the smartphone-based approach is that the guidance to a speaker during data collection can be displayed on the screen of the telephone, *while the data collection is taking place*. Thus, all the complexities of speaker recruitment, server set-up, recording and transcription are reduced considerably, since a field worker can take a batch of pre-loaded smartphones to a location where many first-language speakers of the target language are available, and rapidly collect speech from several speakers in parallel. Since the speakers can be prompted to say particular words, sentences or phrases, transcription is replaced by verification – a considerably simpler task which can, in fact, be automated to a great extent.

Nic de Vries of the Meraka Institute, then a Master's student under Marelie's guidance, adapted this idea to the practicalities of the developing world, and added a number of additional refinements of his own. The resulting open-software source package, *Woefzela*‡, comprehensively changed our data-collection landscape. In short order we had access to staggering amounts of data, by our earlier standards – up to 100 hours of speech per language.

To process such large amounts of data is a challenge in its own right; fortunately, Cambrdge University Engineering Department has long supported an extremely powerful open-source toolkit called HTK for just that goal, and Charl van Heerden has done stellar work to develop a suite of scripts around HTK, so that we can now train highly accurate speech-recognition systems in relatively short amounts of time. We have done so for all of our official languages, and I will describe some of the characteristics of those systems when I delve into applications below.

The remaining resource in the language-specific triad of Fig. 1 is the *language model*. Since this model represents the sequences of words that users are expected to say when using a speech-recognition system, it is highly application specific. That is, whereas a general-purpose pronunciation dictionary or acoustic model may be a reasonable (if sub-optimal) entity to use in a speech application, a general-purpose language model usually makes no sense. Two broad classes of models are employed in practice: hand-crafted models, which require a human to express (in a suitable format) the expected utterances in a particular situation, and statistical models, which represent the expected likelihood of any word given some surrounding context. As can be expected, these two classes are at opposite ends of the difficulty spectrum for under-resourced languages: whereas it is relatively straightforward to list the things a speaker may reasonably say when asked about, say, her date of birth, it is extremely complex to specify all the probabilities of words and phrases that may be spoken in general. For the latter case, systems in well-resourced languages utilize truly huge speech corpora to estimate the relevant statistics: hundreds of millions or even billions of words of text are commonly used. For our languages, with the exception of English, such corpora are simply not available. Fortunately, the applications that formed our initial focus could all be constructed around hand-crafted models; we therefore have so far not developed statistical language models for most of our languages. I will return to this issue below.

Applying speech technology for information access: first steps

After this tortuous detour through technical details, it is perhaps useful to remind the reader of our long-term goal with all this development: recall that we wish to create systems that allow a

‡ The name is a pun on the canine theme of *DataHound*- an early name of the Google data-collection software – and *Vuvuzela*, the noisemaking instrument, probably of Brazilian or American origin, used to such destructive effect during the Soccer World Cup. The connection is that "Woef" is an Afrikaans term of endearment for dogs, and acoustically similar to "Vuv-".

much larger cross section of citizens to gain access to information. I will now describe three projects in which we have investigated how such impact could be achieved. Although these represent only a small portion of the worldwide attempts to deliver information access with speech technology, I believe they are a fair representation of some of the most important trends.

OpenPhone

The OpenPhone service is a general health information line designed for caregivers of HIV positive children in Botswana; these caregivers are individuals who provide personal care for one or more HIV positive children, and include parents, family members and other community members. The vast majority of caregivers are females between the ages of 18 and 65, generally with low levels of literacy. Most caregivers tend to have low-income jobs and many are frequently unemployed. The Botswana-Baylor Children's Clinical Centre of Excellence in Gaborone, Botswana, is a HIV paediatrics hospital where children receive free treatment; it also provides the caregivers with free lectures on various aspects of living with HIV and caregiving. Each caregiver on average attends two lecture sessions. It was observed that caregivers often forget the material covered in the lectures. Reinforcement through written material is not viable as many caregivers are semi-to-low literate. The Baylor lectures and all interactions with caregivers are in Setswana (the local language of Botswana) since most caregivers are not proficient in English. Baylor staff members therefore explain complex health information in accessible terms in the local language.

It was also observed that caregivers often travel large distances (average 28 km and as far as 500 km) with high costs and time spent during a working day. In addition, they often have general health information queries (e.g. nutritional needs, hygiene, etc.) beyond the material covered in the lectures. Although caregivers are encouraged to call Baylor with any questions they may have, and although mobile phone usage and ownership is commonplace (up to 90% of the population has easy access to a mobile telephone), most caregivers are reluctant (and unable) to make person-to-person calls due to the high costs of mobile phone calls. These challenges and issues formed the basis of the design for OpenPhone, an IVR health information service in Setswana and accessible at any time through a simple telephone call. After extensive consultation with experts on HIV caregiving and Baylor staff members, we developed a simple list of menus that callers can access either by selecting choices on the telephone keypad or by speaking keywords when they are prompted. An example of a typical keypad-driven interaction is given in Figure 3; Figure 4 shows the equivalent interaction when speech recognition is used.

User: [Dials number...]

System (Introduction): *Hello and Welcome to the Health Helpline, I am Nurse Lerato and I know you probably have many questions about caring for someone with HIV.*

System (Overview): *I can tell you about Hygiene & Cleanliness, Nutrition, Common Sicknesses, ARV Medication, and Facts about HIV. If at any time during your call you want to start our conversation again, you can press 0.*

System (Main Menu): *For Hygiene & Cleanliness, please press 1, for Nutrition, press 2, for Common Sicknesses, press 3, for ARV medication, press 4 or for Facts about HIV, please press 5.*

User: [Presses 2.]

System: *Eating a balanced diet of different foods helps people with HIV stay healthy and strong. A healthy diet does not have to be costly and contains food from all the different food groups. Healthy food is always prepared and stored in a clean environment...*

Figure 3: keypress-based interaction with OpenPhone

User: [Dials number...]

System (Introduction): *Hello and Welcome to the Health Helpline, I am Nurse Lerato and I know you probably have many questions about caring for someone with HIV.*

System (Overview): *I can tell you about Hygiene & Cleanliness, Nutrition, Common Sicknesses, ARV Medication, and Facts about HIV. If at any time during your call you want to start our conversation again, you can say "Start Over", or press 0.*

System (Main Menu): *Please say the name of the topic you want to hear about... you can say "Hygiene & Cleanliness", "Nutrition", "Common Sicknesses", "ARV medication" or "Facts about HIV".*

User: *Nutrition*

System: *Eating a balanced diet of different foods helps people with HIV stay healthy and strong. A healthy diet does not have to be costly and contains food from all the different food groups. Healthy food is always prepared and stored in a clean environment...*

Figure 4: interaction OpenPhone using speech recognition

In order to determine whether this service would be useful to our target population, and to investigate how users experience the speech-based interaction mode, we piloted OpenPhone for a week at Baylor. The service was tested with 33 caregivers. We asked these caregivers to use both the keypress-based and speech-based systems, and then asked them several questions related to their personal circumstances, the potential role of a service such as OpenPhone in their lives, and their experiences with both types of systems. As expected (in light of the practicalities summarized above), the majority of the caregivers were quite enthusiastic about the option to obtain information in this way, and also about the type of content that was made available to them. They even suggested novel motivations for such a service: for example, one person mentioned that she would like to use the service at home in order to convince her family members that they really had to act according to the guidelines provided by the Baylor personnel.

However, we were surprised to learn that these users – who had little prior exposure to IVR systems – generally preferred to use the keypress-based system rather than the version using speech recognition which we had developed with such care! Users understood that the speech-based system was not really faster than the version using keypresses, and they quickly learned how to use both. Crucially, the subject matter of the service is culturally sensitive, and members of trial group said that it was less awkward to request the information by pressing keys on the telephone than to speak the keywords out loud.

Voice search

When designing OpenPhone, one of our design criteria was that the speech-recognition component should only have to recognize small sets of words, as reflected in the prompt of Figure 4. That was the only feasible option at the time, since our limited corpora severely restricted the accuracy that could be achieved on larger vocabularies. During the time that OpenPhone was being deployed, however, Johan Schalkwyk visited our research group, and initiated the conversations that changed our thinking about corpus size (as described above). We now realized that large-vocabulary recognition was achievable even for under-resourced languages, and thus initiated the process to collect much larger speech corpora. At the same time, Google was interested to expand the number of languages that they cover with their speech services, and Pedro Moreno and Johan arranged for sponsorship from Google for the development of large-vocabulary recognizers in South African English, isiZulu and Afrikaans.

For this development, the folks at Google Research had an extremely interesting application in mind, namely the use of speech to search the World Wide Web. For well-resourced languages – American English in particular – Google had found that spoken searches of the Web were extremely popular with a certain segment of their users, and they wanted to see how well we could do with the three languages mentioned above. While the data collection was proceeding at Meraka Institute (under the guidance of Alta de Waal and her team), Charl van Heerden

went to the Google offices in New York to develop the components required for such Voice Search applications in these three languages. It turned out that some of the tools we had developed for other purposes, such as Marelie Davel's DictionaryMaker as well as several resources that had been developed with the guidance and support of the national Department of Arts and Culture, were extremely useful in this process. As for the remaining obstacle I emphasized above, namely the need for large amounts of text to develop language models, Google had the perfect answer: the queries that users enter when typing into Google constitute a massive reserve of textual material, perfectly suited to this task. When using Google, one can select the language of the interface, and the starting point for language modelling was that this "interface language" was also the language of the query. Of course, this assumption is sometimes not valid: a user may prefer using, say, a Sesotho interface, but then enter English queries into the search. By counting the number of English words in a sample of queries, we found that this was indeed the case for isiZulu: more than 90% of the queries made by users of the isiZulu interface contained either proper names or English words. For Afrikaans, on the other hand, around 50% of the requests were valid Afrikaans words, and another 40% were proper names. Thus, Google could build powerful Afrikaans and English language models based on the text queries alone; for isiZulu, we had to add text from several other isiZulu language sources (collected by Martin Puttkammer and his group at CTexT in Potchefstroom) in order to have a reasonable language model.

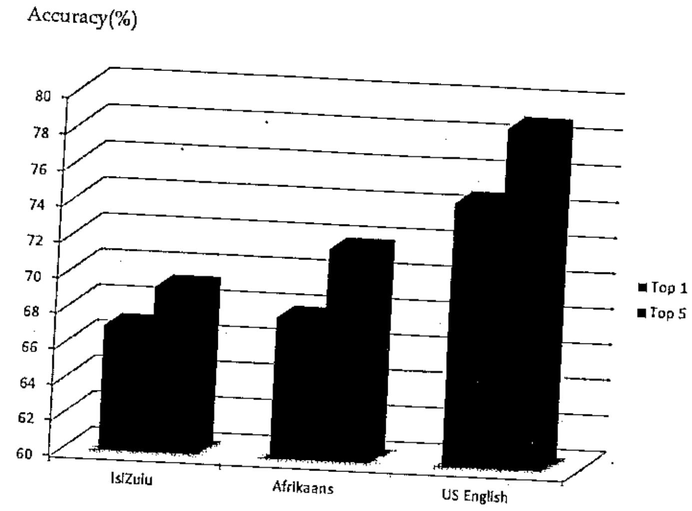

Figure 5: Comparative accuracies of three voice-search systems. "Top 1" and "top 5" refer to the percentage of cases with the correct answer as the first answer returned by the system, or within the top 5 returned by the system, respectively.

Under the guidance of Pedro and Martin Jansche, Charl was then able to complete the language-specific components of Google Voice Search in less than three months. Figure 5 gives an indication of the level of success that was achieved by this process: it shows the accuracies achieved on a set of held-out queries for the Afrikaans and isiZulu voice-search systems, with the accuracy of the Google US English system as a reference value[§]. These results are quite impressive, and Google was happy to release Voice Search in all three languages as part of their commercial offering. The usage statistics of these services are unfortunately not made public by Google, but from talking to smartphone users in South Africa, I have the impression that especially the Afrikaans service is used fairly extensively by smartphone users. Usage of the South African English service is also picking up, but for isiZulu the problem is that there are so few resources in isiZulu on the Web – hence, users do not even attempt to use their home language to search for information.

§ For technical reasons, we did not compute the equivalent value of the South African English system; however, our anecdotal experience is that its accuracy is about mid-way between that of the Afrikaans and US English systems.

Lifelines

Our third and final exploration of information access through speech technology takes us to India, where we have recently completed a very informative project along with Stephane Boyera and Aman Grewal of the Web Foundation. This project, which was sponsored by the Rockefeller Foundation, investigated user acceptance of speech technology in an agricultural helpline called Lifelines. This helpline is operated by a non-governmental organization called OneWorld South Asia. It is aimed at small farmers in three Indian states, and has been running for more than four years, providing over 200,000 farmers with tailored, specific advice to assist them in their farming practices.

The current operation of Lifelines is schematized in Figure 6 (which was provided to me by Shakeel Ahmad, who was then at Lifelines). The interaction starts with a farmer making a phone call to an automated system (managed by OneWorld), which simply records the farmer's query, and provides him with an identification number that can be used to retrieve the answer to his query at a later stage. A group of knowledge workers at OneWorld retrieve these queries, and then use various means to answer them as well as possible. They first consult a database of queries that had been answered previously, then search text books or the Web, and finally contact one of a set of contracted agricultural experts. The resulting answer is added back into the database of "Frequently asked questions" (FAQs), and recorded to the automatic system. Thus, when the farmer calls back after 24 hours or more to retrieve the answer (using the ID number assigned to his query), he has access to expert-curated information that is pertinent to his specific concerns and circumstances.

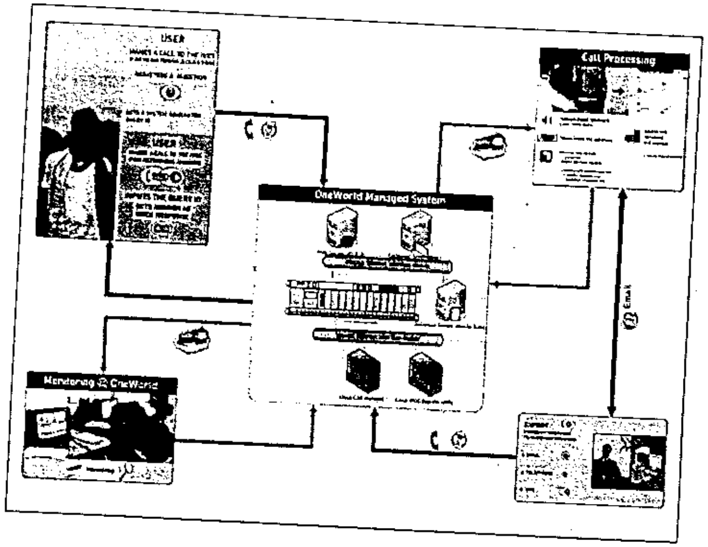

Figure 6: Flow diagram of the Lifelines service

This system has been evaluated in several ways, and it has been a remarkable success in terms of the number of people reached and the impact on their farming practices. Despite this great track record, however, it is still not entirely self-sustaining: the extensive human involvement at all stages of the process is rather costly, and the targeted users are, of course extremely price sensitive. Thus, the nominal payment received from those users does not entirely cover the costs of the service, and increasing the scale of the service would cause a proportionate increase in the external funding required. This creates an ideal opportunity for speech technology, since we can aim to multiply the efficiency of the paid workers through the use of such technology.

As a first step in that direction, we obtained funding from the Rockefeller Foundation to investigate whether the user population would find such technology acceptable (from a usability perspective), and trustworthy. Our investigation was purposefully limited: given all the complications experienced with voice-based services such as OpenPhone, we wanted to modify the Lifelines agricultural service as little as possible, while still getting good insights on the potential of speech technology in this context. We therefore extended the menu of choices provided to callers, by giving them the option to listen to one of three frequently-asked questions (FAQs) rather than leaving a query. The FAQs were carefully selected to be issues that had been relevant to callers from those regions in preceding years, but the lack of a speech-recognition system forced us to use only three FAQs per trial location. (Using caller

identification on the telephone system, we could readily determine the location from which each call originated, and the caller could select one of these questions by keypress, or select to leave a new query. If an FAQ were selected, the corresponding answer would be played back immediately, whereas new queries required the normal 24-hour waiting time.) For the purposes of the experiment, two systems were implemented: in one case the answers to the FAQs were read by a human, in the other by a TTS system.

Our investigation of the FAQ-enhanced Lifelines system consisted of two phases. In the first, we employed field workers to accompany our experimental team; the farmers whom the field workers encounter in their normal routine were asked to try the original system and the FAQ-enhanced version (with either recorded or TTS-produced responses being selected at random for different trials). The experimenters then asked the farmers a series of questions to assess how they perceived the various systems in comparison to one another. In the second phase, we simply replaced the normal Lifelines service with the FAQ-enhanced version for a period of 5 weeks for selected callers, and monitored their usage of the FAQ portion of the service.

The results from both phases were highly encouraging. During the interviews, the farmers suggested that they found the new information source to be equally credible to the old one, but substantially more convenient. As the comments in the sidebar indicate, farmers everywhere would like more handouts ... but these farmers in India also understood the value of the new information source, and even suggested novel uses of the information that we provided to them (for example, making printed versions as a general way to support farmers with relevant information). These subjective responses were confirmed by the usage patterns during Phase 2: callers spontaneously listened to the FAQ content, and their usage of that information was in fact somewhat higher than the frequencies we had predicted based on previous years' requests to the system.

User comments during Lifelines field trial

The users that we interviewed during our Lifelines field trial gave a number of responses that were encouraging, insightful, surprising, ... Here is a sample (summarized and translated from Hindi):

If pesticide is freely available, will experiment and then trust.

Response to fresh query should be faster

If we get answer same time, it is better

What would be better than getting suggestion sitting at home.

Difficult to dial 1, 2 or 3.

Whatever information computer is providing; should also be given in a farmer's meet for awareness.

With this feasibility study behind us, it is clear that there is great potential for speech recognition in this service – several users commented on the fact that only three FAQs are too limited for their needs, and the large database of existing FAQs at Lifelines should enable us to answer a substantial fraction of the questions automatically – the Lifelines staff estimate that about 80% of their queries are currently being answered from that database. An automated system will, of course, not match human ingenuity in retrieving all relevant answers, but if we could answer 50% of the calls automatically, the service would probably be self-sustaining. We are currently looking for funding to implement such a solution.

Conclusion: looking back, looking forwards

Looking back at our work of the past 10 years or so, I am struck by the unpredictability of it all. Rather than a bundle of lines leading from insights to impact, I see a patchwork of stirring successes mingled with severe disappointments. In the latter category, the most significant item is probably the fact that our "build it and they will come" approach to information access has not taken off yet. This, despite the fact that some of our most basic insights have repeatedly been confirmed in the ten years we have been working on this problem. The importance of information is even more widely appreciated now, and speech-enabled access to such information has been a core focus for many of the largest technology companies in the world (including Apple, Google, Microsoft and IBM). Well-considered analyses of developing-world requirements, such as a multinational study lead by the Web Foundation, have also confirmed the substantial potential of speech technology in the developing world. Nevertheless, a development-oriented voice-based service that can sustain itself can still not been found anywhere. I am still not clear on what we are doing wrong: availability of content has certainly been a much tougher problem than we had imagined, and the high cost of telecommunications in South Africa has also stymied some of our own efforts. I am hopeful that our partnerships with Google, the Web Foundation and OneWorld will enable us to overcome these particular hurdles, but I cannot predict whether other unforeseen barriers will hold us back down the line.

On the positive side, our technology-development efforts have been most gratifying. We now have high-quality speech-recognition systems in all the official languages of South Africa, and a good understanding of the process of developing additional languages or achieving better accuracies in particular applications. Our technology for text-to-speech systems is not quite as sophisticated, but progress in that domain has also been remarkable given the relatively limited resources at our disposal and the multiplicity of languages in our focus. Even more satisfying is the crop of excellent young scientists and technologists who have taken up the

challenge to develop these technologies. I have mentioned the names of several of my colleagues and students who have been responsible for many of our basic developments, and there are at least two dozen others who have become highly proficient in the general scientific method, but also in technical areas encompassing speech technology, software development, signal processing, African linguistics, and other domains.

On balance, of course, I remain optimistic. Under the spell of the Internet bubble, I had certainly hoped that our activities would lead straight from resource and technology development to impact, but I now realize that such a short path is the exception in the world of technology innovation. A much more likely trajectory is a steady progression that leads from technology to user acceptance, sustainability, scalability ... and from that solid platform, impact is more likely to arise. Viewed from such a perspective, it is important for us to continue building our foundations while striking up partnerships to ensure that we have the building blocks in place, all the way up to impact. Along the way, we must keep listening to our end users – they have repeatedly demonstrated innovative insights to us, and perhaps our biggest failure to date has been that we have not interacted with those users at sufficient length.

Finally, what lessons regarding the process of technology planning can be distilled from the events I have described? On the one hand, the doubts about long-term plans reflected in my epigraph are certainly supported by the surprises (both positive and negative) we have encountered along the way. But Eisenhower's respect for the planning process is also duly supported by our experiences. Our wholehearted – and sustained – attempts to carry out those same plans were critical to building the in-depth knowledge and skills that allowed us to capitalize on the opportunities that came our way. As is often the case, Voltaire seems to have known the truth all along:

A state of doubt is unpleasant, but a state of certainty is ridiculous.

Acknowledgments

Every step I have described above was taken in collaboration with my sometimes manager, sometimes graduate student, and constant collaborator, Marelie Davel. I am extremely grateful to her, to the excellent graduate students who have contributed so much to this work, and to my various colleagues and collaborators at North-West University, Meraka Institute, Google, the Web Foundation and elsewhere. Funding from the Department of Arts and Culture, Google and other generous supporters has played a crucial role in getting us this far. Finally, heartfelt thanks to my family members and friends, who have supported both the planned and the unexpected aspects of my journey with such kindness.