Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 81 (2016) 128 – 135

Procedia Computer Science

5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016, 9-12 May 2016, Yogyakarta, Indonesia

Code-switched English Pronunciation Modeling for Swahili Spoken Term Detection

Neil Kleynhans[a,b,*], William Hartman[c], Daniel van Niekerk[a,b], Charl van Heerden[a,b], Rich Schwartz[c], Stavros Tsakalidis[c], Marelie Davel[a,b]

[a] *Multilingual Speech Technologies, North-West University,Vanderbijlpark, South Africa.*
[b] *CAIR, CSIR Meraka, South Africa*
[c] *Raytheon BBN Technologies, Cambridge, MA 02138, USA.*

Abstract

We investigate modeling strategies for English code-switched words as found in a Swahili spoken term detection system. Code switching, where speakers switch language in a conversation, occurs frequently in multilingual environments, and typically deteriorates STD performance. Analysis is performed in the context of the IARPA Babel program which focuses on rapid STD system development for under-resourced languages. Our results show that approaches that specifically target the modeling of code-switched words, significantly improve the detection performance of these words.

Peer-review under responsibility of the Organizing Committee of SLTU 2016

Keywords: Spoken term detection; code switching; Swahili, pronunciation modeling

1. Introduction

Multilingual environments provide many challenges to current speech recognition systems. One aspect that impacts the performances of such systems is code switching. Code switching occurs when speakers switch to a different language during a conversation, borrowing words, phrases or sentences. In terms of code switching, it is useful to define the primary language as the matrix language and the language used to source borrowed terms, as the embedded language. An important part of current speech recognition acoustic modeling, is the pronunciation dictionary. This resource defines the manner in which speakers generally pronounce words in a language. An implication of code switching is that additional languages are effectively introduced into the pronunciation dictionary. The goal is then to develop an approach to model the code-switched words accurately, given the constraints of the matrix language and available resources.

* Neil Kleynhans. Tel.: +27-28-312-1907 ; fax: +27-16-910-3116.
E-mail address: ntkleynhans@gmail.com

Peer-review under responsibility of the Organizing Committee of SLTU 2016
doi:10.1016/j.procs.2016.04.040

Our research focuses on pronunciation modeling of English (embedded language) words within a Swahili (matrix language) domain, and the implications on Swahili spoken term detection (STD). The Swahili orthography lends itself to graphemic pronunciation modeling as the written and spoken forms are quite regular. English, however, performs significantly poorer with such modeling given its irregular orthography. Thus, the challenge is to produce robust English pronunciations given the under-resourced Swahili context, and specifically the absence of any pronunciation lexicons.

Following a brief overview of the Babel program and code-switched speech recognition (Section 2), we describe the experimental setup and data used (Section 3). A description of our approach to various aspects of the tasks and the results obtained are interspersed in Section 4: we describe the baseline we wish to optimize, the process of identifying code-switched words, and modeling options with and without using additional English resources.

2. Background

The *Babel* program[1] is an international collaborative effort sponsored by the US Intelligence Advanced Research Projects Activity (IARPA) and aims to develop techniques for the rapid development of STD systems. On completion in 2016, this four-year project will have investigated 26 different languages from across the globe. Audio data consists mostly of spontaneous speech recorded over telephone lines, with a variety of recording conditions and noise levels included.

The program is run as a challenge, with initial analysis on so-called "development languages" culminating in a yearly "surprise language" evaluation. Once the surprise language has been announced and data made available, participating teams have a limited amount of time to process data and develop fully-fledged STD systems (one week in 2016). Various training and test conditions are applicable, with our focus here the *Very Limited Language Pack (VLLP)* condition, which makes only three hours of transcribed audio available for training. Apart from allowing the inclusion of multilingual data from earlier Babel languages, no external resources are allowed for this condition, and no pronunciation lexicon provided.

The 2015 surprise language was Swahili, as spoken in Kenya. Swahili is a Bantu language spoken by more than 120 million people across especially East Africa, even though less than 2 million are considered native speakers[2] (quoting Wald[3]). It is therefore not unexpected that, in conversational speech, Swahili is often mixed with other languages, most notably English and Arabic[2].

Code switching is frequently observed in many of the Babel languages (including, for example, Igbo, Javanese and Zulu). This reflects a known tendency when languages co-occur, with high frequencies of code-switching observed for a variety of languages[4]. This phenomenon has been studied in the context of both STD and speech-to-text (STT) systems, with various approaches to model adaptation and combination demonstrated[5,6,7]. These studies typically assume language-specific pronunciation lexicons exist. When pronunciation lexicons exist, a simple strategy for modeling code-switched speech, or otherwise generating multilingual acoustic models, is to use some form of mapping from the embedded to the matrix language. Approaches that have been applied in practice, include knowledge-based mappings such as IPA, and data-driven approaches such as confusion matrices[5], or hierarchical clustering based on phone mappings[7]. This phenomenon is not as well studied for English / African language pairs, but some results are available[8,9].

3. Experimental Setup

The Swahili analysis was performed using the *IARPA-babel 202b-v1.0d* release. Data consisted of audio collected over telephonic channels (land-lines, cellular, in-car) and split into conversational and scripted speech. Evaluation data contained only conversational speech. Each training utterance had an accompanying transcription that contained time-marked word-level orthography. In addition, untranscribed speech was made available for unsupervised acoustic model improvement. No pronunciation lexicon was made available, but a document describing the language, referred to at the language-specific peculiarities (LSP) document[2] was provided. For the VLLP condition analyzed here, three hours of transcribed training data, three hours of tuning data and 10 hours of evaluation data were provided.

State-of-the-art techniques were used to develop STT and STD systems. Briefly, the system utilizes multilingual stacked bottleneck features for feature extraction[10] and discriminatively trained GMMs for acoustic modeling. To

improve keyword spotting, a modified search algorithm was implemented, as described by Tsakalidis *et al.*[11]. When combining whole word systems with sub-word systems, the techniques described in Karakos and Schwartz[12] were used. Given the limited amount of transcribed training data, the text used for language model training and the pronunciation lexicon were augmented with web data[13].

4. Approach

The starting point for our investigation was to perform an acoustic analysis of the Swahili audio, specifically listening to randomly selected samples of English word pronunciations. From the analysis it was concluded that in most cases English words were pronounced using standard English letter-to-sound rules with limited nativization. The pronunciation variation was quite large and ranged from standard to accented English. All spelled words (e.g. *SMS*) were pronounced in an English manner.

Based on this analysis, different approaches to modeling the English words were investigated. We first apply an existing approach to lexicon development, as used in earlier work[14] (Section 4.1). This provides the baseline we wish to optimize. We then consider the task of identifying possible code-switched words, using techniques from text-based language identification (Section 4.2). The 2015 VLLP condition specified that no external resources (such as an English pronunciation lexicon) could be used. We therefore first consider approaches to modeling English words without having access to English resources (Section 4.3), before considering the improvements possible if English resources were allowed (Section 4.4).

4.1. Baseline process

The baseline process is described in more detail in Davel *et al.*[14] and relies heavily on the LSP document. This document contains a host of language-specific information such as an overview, written orthography, potential dialect differences, and importantly, basic phonemic translation of letters into IPA and X-SAMPA representations. The initial step was to create a letter-to-phone (L2P) mapping based on the LSP document. Once a basic mapping has been generated, rare phonemes are removed. This process is iterative, where phones are merged based on phonemic distance. A decision to remove a phone is based on frequency of occurrence and a threshold of 30 was chosen. Careful consideration is given to resultant homophone creation when merging phones.

Two maps are created: one for general words and another for spelled letters – the LSP document contains different definitions for these two cases. The spelled letter map transliterates spelled letters to phonemic representation. These are predominantly defined for English pronunciations of spelled letters. When merging phones, both maps contribute to the phone occurrence counts.

Syllables were generated automatically for sub-word modeling purposes using a simple, language-independent algorithm also described in Davel *et al.*[14]. The algorithm is applied to pre-processed text, and uses known word-beginning and word-ending consonant clusters to discover the syllable structure of the language.

4.2. Language identification

The English word identification process was constrained as only a pre-defined list of resources could be used to develop the system. Our sources were: BABEL-wide shared web data, Omniglot[15] and common English knowledge. The initial English word classifier was a direct look-up. The English words were extracted from the web data and 100k words were selected based on frequency. After a manual inspection of the identified words, it was found that the look-up method generated many false positives. This is due to noisy English word list extracted from the web data. Therefore, two follow-up processes were introduced to prune non-English words. The processes were rule-based inclusion and exclusion, and, word removal based on length.

The defined rules were:

- English sub-word pattern detection with exclusion e.g. x, q, c but not ch
- Letter pair detection e.g. ss, ll
- Swahili sub-word pattern exclusion e.g. hawa, sai

The first version of the rules contained Swahili sub-word pattern detection and vowel pair detection rules. These rules were excluded, however, as they generated many false positives. Lastly, all words that were three letters or less in length were excluded from the English word detection. This threshold was chosen based on visual inspection and balancing false positives and misses.

Table 1 shows number of unique Swahili and English words detected in the *Full Language Pack* (FullLP) transcription set. As can be seen, there is a large proportion of unique English words (types): 13% of all words. For comparison, a joint-sequence-model (JSM), developed using Sequitur[16,17] and trained on the CMUDict0.7b[18] American English dictionary (over 134k words), was used to detect English words in the FullLP set. The JSM system detected 5,467 words at a 90% confidence threshold. This indicates that there may be many more English words present, that are not detected via the closed resource English word identifier.

Table 1. English/Swahili words in FullLP training data.

# unique words	26,202
# Swahili words	22,726 (87%)
# English words	3,476 (13%)

Table 4.2 shows the manner in which English words were detected for both the FullLP and VLLP data-sets. The results show that for the vast majority of cases the look-up methods detects the most number of English words. The English sub-word patterns detect around 20% for FullLP and less than 10% for VLLP. The exclusion rules remove roughly 10% of the English detected words. These results show the importance of a large clean English word set.

Table 2. English word detected based on method of classification

Method	*# English words*	
	FullLP	*VLLP*
lookup	3,026	1,542
rules detection	693	124
rules exclusion	170	34
length exclusion	70	52
Total	3,479 (13%)	1,580 (30%)

Given the reliance on English word look-up, as shown in Table 4.2, a comparative evaluation was run on the VLLP set using an external source of 30k most frequent English words derived from the 100M word British National Corpus (BNC)[19]. The results showed a 72% precision and 94% recall.

4.3. No additional English resources

Building on the baseline process, a series of increasingly complex L2P maps were created. The first version (*gra*) was a close to graphemic system, with each letter mapped to a single phone. The next L2P map, (*swa-lsp*), implemented the baseline approach, using information from the LSP to expand the graphemic system by adding rules for Swahili letter sequences (e.g. $nj \rightarrow /n\ dZ/$, $ng \rightarrow /N\ g/$). The following map (*mixed*) augmented the *swa-lsp* Swahili mappings with English-specific letter sequence mappings. These included double letters (e.g. $ll \rightarrow /l/$) and other unique English sequences (e.g. $qu \rightarrow /k\ w/$, $chr \rightarrow /k\ r/$. Where a sequence is realized differently in Swahili or English, only the Swahili sequence was applied.

For the next two sets, text-based language identification was used to identify English words, and an English-specific L2P map applied. For one of the sets (*eng-tagged*) vowels were kept separate and tagged with predicted language origin. For the other (*eng-translit*) all English phones were transliterated to Swahili graphemes, as also proposed in Basson and Davel[20]. This map contained an extended set of English letter sequences which included diphthongs. Diphthongs were modeled using the closest Swahili vowel or vowel combination. In both cases these English L2P predictions were added to a dictionary as variants to *swa-lsp* Swahili predictions – a single English-specific prediction was generated for each identified word. Both variants were added given (a) the possibility that the word may exist

in both English and Swahili, and (b) for true English, the possibility that a specific Swahili speaker may produce a Swahili pronunciation for the English word.

Table 3 shows an example of the *swa-lsp*, *eng-tagged* and *eng-translit* L2P predictions.

Table 3. Swahili L2P mappings for English word "anyway"

Mapping	*Result*
swa-lsp	A n j w A j
eng-tagged	A n j w A j
eng-translit	A n j w A j

A_EN n j w A_EN j
A n i w E j

Results are summarized in Table 4. Standard Word Error Rate (WER) and Actual Term-Weighted Value (ATWV)[21] are used to report on STT and STD results, respectively. The Term-Weighted Value (TWV) indicates the STD's performance as a trade-off between false accepts and misses – a perfect system would score 1. TWV is calculated by subtracting, from 1, the average loss per term. The ATWV indicates the TWV at the system's set detection threshold. STD results are shown separately for in vocabulary (IV) and out-of-vocabulary (OOV) keywords. (OOV keywords were not included in the language model during term detection.)

Table 4. STD and STT results obtained with different mappings, when no additional English resources are used

	WER	*ATWV*		
Mapping		*IV*	*OOV*	*All*
gra	54.95	0.4756	0.4533	0.4716
swa-lsp	55.19	0.4693	0.4715	0.4697
mixed	54.62	0.4821	0.4629	0.4786
eng-tagged	54.82	0.4773	0.4551	0.4733
eng-translit	54.81	0.4803	0.4621	0.4771

The results in table4 show the *mixed* map produces the best system across measures. Surprisingly, the graphemic (*gra*) system performs marginally better compared to the initial LSP map (*swa-lsp*) system – except for the OOV ATWVs. In terms of the English mappings, the *eng-translit* map produced a slightly better result compared to the *eng-tagged* map, but both these maps do not out-perform the *mixed* map.

4.4. Additional English resources

Given the restrictions determined by the Babel evaluation process, we modeled code-switched English words by performing simple LID followed by "common knowledge-based" L2P mappings specific to English. However, another realistic scenario involves adapting and applying freely available and extensive pronunciation resources to this task (such resources certainly exist for English and many other major world-languages). Investigating this, we analyzed the STD results obtained by the *eng-tagged* system on English queries as a function of overlap/correspondence with an existing reference English pronunciation dictionary.

As the reference dictionary, we used a South African English (*SAE*) pronunciation dictionary[22] and mapped these to the Swahili phoneset. Unseen words were generated using Default&Refine-extracted G2P rules[23], obtained from the same dictionary. The *SAE* dictionary was selected because it contains simplified versions of British pronunciations, and had previously been applied in the context of English code switching. The experiment was performed on an English-only subset of the VLLP development (*dev*) set (a total of 234 phrases). A histogram plot of ATWV vs dictionary overlap score (*dictscore*) and phrase length is shown in Figure 1. The dictionary overlap score is the phone accuracy averaged over all the variants in the reference dictionary (mapped *SAE*), where for each variant the closest mapping in the target dictionary (*eng-translit*) was selected. Higher values of *dictscore* indicate greater overlap between *eng-tagged* and the mapped *SAE* English pronunciations. Notwithstanding sparseness in sections of the result space, the consistent gradient for the medium to longer phrases as *dictscore* increases, suggests the potential utility of the mapped SAE English dictionary in the STD task.

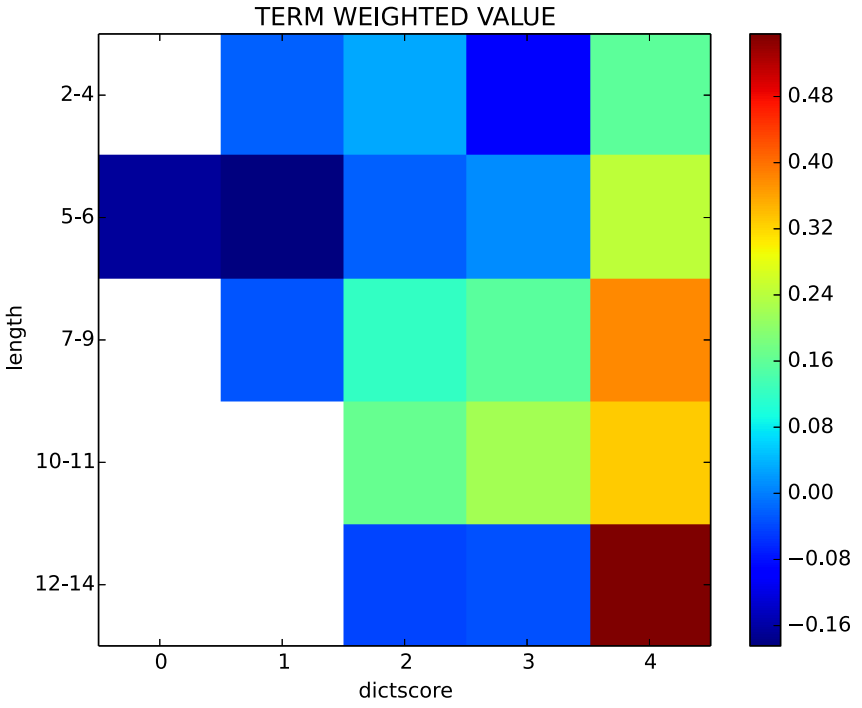

Fig. 1. Histogram of the ATWV as a function of dictionary overlap (*SAE* and *eng-tagged*) and phrase length.

Following up the result in Figure 1, further STD and ASR experiments with the mapped *SAE* dictionary were performed. Tables 5 and 6 contain results on the VLLP *dev* and *dev+tune* sets respectively. Table 5 compares the ASR and STD results when using the *Web data* and BNC English word lists for LID combined with the mapped SAE approach. The WERs and ATWVs show that the LID systems produce comparable results. The BNC LID system was trained on "cleaner" English data but this did not translate into a performance gain. This may allude to a minimal overlap between the Swahili and reference English words.

Table 5. ASR and STD performance on the VLLP *dev* set using different LID resources.

LID system	*WER*	*ATWV*		
		IV	*OOV*	*All*
Web data	54.67	0.4898	0.4749	0.4871
BNC	54.30	0.4929	0.4783	0.4903

In Table 6 STD results are presented using comparable English word detection. Results are partitioned for English-only, Swahili-only and Mixed phrases, indicating potential gains for both English-only and Mixed classes at a slight cost to the Swahili-only accuracy. Focusing on the English-only result we present the phrase-length-dependent results in Figure 2 which also supports the application of the mapped SAE dictionary.

5. Conclusion

The work presented here describes an investigation into pronunciation modeling approaches for code-switched English in a Swahili spoken term detection system.

Table 6. STD results for all of the systems tested on the VLLP *dev+tune* set.

-	*gra*	*swa-lsp*	*mixed*	*eng-translit*	*SAE*	*# kwds*
English	0.2193	0.2066	0.2111	0.2174	0.2732	415
Mixed	0.4906	0.4865	0.5107	0.5179	0.5203	519
Swahili	0.5251	0.5297	0.5310	0.5295	0.5266	2,111
All	0.4776	0.4783	0.4840	0.4850	0.4910	3,045

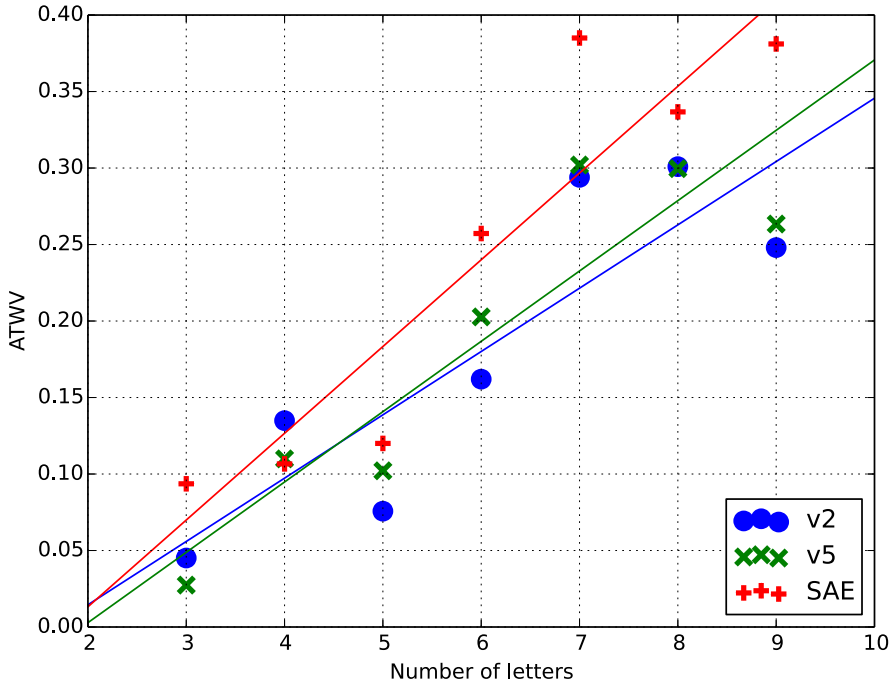

Fig. 2. STD results (*dev+tune* keywords on *dev* set) for English using the *mixed*, *eng-translit* and *SAE* mapped dictionaries, with linear interpolation used to indicate trends.

As expected, English words perform much more poorly than Swahili or mixed phrases, across all systems. As there is a difference in keyword length, these results are not directly comparable, but the same trends are seen when analyzing results per keyword length (results are not presented here).

English-only results are comparable across systems, however, and there was a small gain in English STD performance from *mixed* to *eng-translit*. The largest gain in English STD performance occurs when using an English pronunciation dictionary for English words (*SAE*), where a more significant gain (0.217 to 0.273) is observed. Furthermore, while there is a benefit from both basic Swahili L2P over plain graphemic (*gra* vs *swa-lsp*) and mixed L2P over plain graphemic (*gra* vs *mixed*), a slight degradation is observed when modeling English words more aggressively (*eng-translit* and *SAE*).

A surprising result is the fact that the graphemic English average ATWV is comparable with and slightly higher than all of the other maps apart from the *SAE* map. This is due to a lower number of false accepts (FAs) than the other maps. (The *eng-translit* had 1,084 correctly identified English keywords compared to 1,038 using the *gra* map; however *eng-translit* had 1,876 FAs while the system using the *gra* mapping had only 1,686 FAs.)

This work confirmed that without explicit modeling, code-switched speech degrades both ASR and STD performance, and showed that improved performance can be achieved using fairly straightforward approaches. It also highlighted the need for trustworthy English resources during this process.

6. Acknowledgment

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory contract number W911NF-12-C-0013. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either express or implied, of IARPA, DoD/ARL, or the U.S. Government.

References

1. Harper, M.P.. Babel. `http://www.iarpa.gov/index.php/research-programs/babel`; 2016. Accessed: 2016-01-20.
2. Language-specific peculiarities document for Swahili as spoken in Kenya. `https://mig.nist.gov/BabelOP3/LSPs/LSP_202_final.pdf`; 2016. Accessed: 2016-01-20.
3. Wald, B.. Swahili and the Bantu languages. In: Comrie, B., editor. *The major languages of South Asia, the Middle East, and Africa*. London: Routledge; 1990.
4. Nilep, C.. Code switching in sociocultural linguistics. *Colorado Research in Linguistics* 2006;**19**:1–22.
5. White, C., Khudanpur, S., Baker, J.. An investigation of acoustic models for multilingual code switching. In: *Proc. INTERSPEECH*. 2008, p. 2691–2694.
6. Imseng, D., Bourlard, H., Magimai-Doss, M., Dines, J.. Language dependent universal phoneme posterior estimation for mixed language speech recognition. In: *Porc. ICASSP*. IEEE; 2011, p. 5012–5015.
7. Yeh, C.F., Lin, Y.C., Lee, L.S.. Minimum Phone Error model training on merged acoustic units for transcribing bilingual code-switched speech. In: *Proc. ISCSLP*. IEEE; 2012, p. 320–324.
8. Modipa, T.I., Davel, M.H., De Wet, F.. Implications of Sepedi/English code switching for ASR systems. In: *Proc. PRASA*. 2013, p. 64–69.
9. Modipa, T.I., Davel, M.H.. Predicting vowel substitution in code-switched speech. In: *Proc. PRASA-RobMech*. IEEE; 2015, p. 154–159.
10. Grézl, F., Karafiát, M., Vesely, K.. Adaptation of multilingual stacked bottle-neck neural network structure for new language. In: *Proc. ICASSP*. 2014, p. 7654–7658.
11. Tsakalidis, S., Hsiao, R., Karakos, D., Ng, T., Ranjan, S., Saikumar, G., et al. The 2013 BBN Vietnamese telephone speech keyword spotting system. In: *Proc. ICASSP*. Florence, Italy: IEEE; 2014, p. 7829–7833.
12. Karakos, D., Schwartz, R.. Subword and phonetic search for detecting out-of-vocabulary keywords. In: *Proc. INTERSPEECH*. Singapore; 2014, p. 2469–2473.
13. Zhang, L., Karakos, D., Hartmann, W., Hsiao, R., Schwartz, R., Tsakalidis, S.. Enhancing Low Resource Keyword Spotting with Automatically Retrieved Web Documents. In: *Proc. INTERSPEECH*. Dresden, Germany; 2015, p. 839–843.
14. Davel, M., Karakos, D., Barnard, E., van Heerden, C., Schwartz, R., Tsakalidis, S., et al. Exploring minimal pronunciation modeling for low resource languages. In: *Proc. INTERSPEECH*. Dresden, Germany; 2015, p. 538–542.
15. Ager, S.. Omniglot: English. `http://www.omniglot.com/writing/english.htm`; 2016. Accessed: 2016-01-21.
16. Bisani, M., Ney, H.. Joint-Sequence Models for Grapheme-to-Phoneme Conversion. *Speech Communication* 2008;**50**(5):434–451.
17. Giwa, O., Davel, M.H.. Language identification of individual words with joint sequence models. In: *Proc. INTERSPEECH*. 2014, p. 1400–1404.
18. University, C.M.. The CMU pronouncing dictionary. `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`; 2016. Accessed: 2016-01-21.
19. Kilgarriff, A.. BNC database and word frequency lists. `http://www.kilgarriff.co.uk/bnc-readme.html`; 2016. Accessed: 2016-01-21.
20. Basson, W., Davel, M.H.. Category-based phoneme-to-grapheme transliteration. In: *Proc. INTERSPEECH*. Lyon, France; 2013, p. 1956–1960.
21. Fiscus, J.G., Ajot, J., Garofolo, J.S., Doddingtion, G.. Results of the 2006 spoken term detection evaluation. In: *Proc. of ACM SIGIR Workshop on Searching Spontaneous Conversational*. Amsterdam, The Netherlands; 2007, p. 51–55.
22. Loots, L., Davel, M., Barnard, E.. Comparing manually-developed and data-driven rules for P2P learning. In: *Proc. PRASA*. Stellenbosch, South Africa; 2009, p. 35–39.
23. Davel, M., Barnard, E.. Pronunciation prediction with Default&Refine. *Computer Speech & Language* 2008;**22**(4):374–393.