CONSTRUCTING SUB-WORD UNITS FOR SPOKEN TERM DETECTION

Charl van Heerden[1], Damianos Karakos[2], Karthik Narasimhan[3], Marelie Davel[1] and Richard Schwartz[2]

[1]Multilingual Speech Technologies, North-West University, South Africa.
[2]Raytheon BBN Technologies, Cambridge, MA 02138, USA
[3]CSAIL, MIT, Cambridge, MA 02138, USA

ABSTRACT

Spoken term detection, especially of out-of-vocabulary (OOV) keywords, benefits from the use of sub-word systems. We experiment with different language-independent approaches to sub-word unit generation, generating both syllable-like and morpheme-like units, and demonstrate how the performance of syllable-like units can be improved by artificially increasing the number of unique units. The effect of unit choice is empirically evaluated using the eight languages from the 2016 IARPA BABEL evaluation.

Index Terms— Spoken term detection, BABEL, sub-words, syllables, morphemes.

1. INTRODUCTION

The detection of out-of-vocabulary (OOV) keywords[1] is a significant area of research in keyword spotting, mainly because OOV keywords tend to correspond to named entities or other content-bearing words of interest.

As has been documented in a number of papers [1, 2, 3, 4, 5, 6, 7], the best techniques for OOV detection involve decoding with a variety of units, such as syllables, morphemes and 1-n phone units. The sets of hits (postings lists) generated from these decodings are subsequently combined together (system fusion) using techniques such as the one described in [8]. The reason for the gains is diversity: each decoding unit (e.g., syllable) has a different degree of acoustic confusability (e.g., shorter units are more confusable with each other) and expressiveness (e.g., shorter units are more flexible in that they can represent a larger set of OOV keywords). So, all these units offer complementary strengths.

Important questions that arise when detecting OOV keywords with sub-word units relate to the level of granularity; the approach one should follow when generating such units; and whether one should use *query expansion* during search. By query expansion we mean the process of searching for closely related sequences of units, or allowing phone insertions, deletions and substitutions with some cost. For example, these alternatives could be close to the keyword of interest in some feature space, or close in terms of weighted phonetic edit distance.

A number of papers that utilize sub-word units for decoding and search [3, 4] just search for the exact sequence of units. Other papers [6, 9, 10, 1, 11] use a confusion model to come up with "proxy" keywords which are phonetically close (or, alternatively, allow fuzzy match). As we have found in our experiments, it is of paramount importance to utilize query expansion, no matter what sub-word unit is used. One notable exception is [12], where it is shown that the performance of the proxy keyword search (which employs query expansion) is almost as effective as searching sub-word (syllable) lattices with an exact match. This is different from our finding, but there are several differences between systems.

In this paper we present a variety of methods for generating diverse sub-word units and we show that they offer significant gains in the detection of OOV keywords. Our techniques are tunable, so that a particular level of granularity can be achieved, offering a trade-off between confusability and OOV keyword detectability.

The paper is organized as follows: Section 2 offers background on the task of keyword spotting, Section 3 describes different approaches to generating and utilizing sub-word units, while the experimental setup and results are presented in Sections 4 and 5, respectively.

2. BACKGROUND

In this paper we focus on the problem of keyword spotting from speech in an off-line mode. That is, we assume that a speech recognition system processes the incoming audio and saves lattices and/or an index to disk for later processing. Searching for queries (keywords) is then done by matching a representation of these keywords (e.g., in terms of sequences of phones) with sequences of units in the lattice or the index. Although, in principle, one could perform the search in real time using the 1-best answer or even a lattice, the results presented in this paper are done using a combination of techniques and decoding schemes that can be very challenging to do in an on-line mode.

Out-of-vocabulary (OOV) keywords pose a challenge, as they are absent from word-based lattices. The usual strategies for detecting such words involve (i) doing an approximate match in word-based lattices (the recognizer will replace the OOV words with in-vocabulary words that are usually acoustically similar) [6, 13, 14]; (ii) generating lattices that contain units of a finer granularity (syllables, morphemes) and then searching for sequences of those, but also allowing for inexact matches [1, 15, 2, 3, 16, 17, 18]. The latter is the approach that we follow in this paper.

One reasonable choice of sub-word for tackling the problem of OOV keywords is linguistically-motivated morphemes of the language. Morphemes represent the basic meaning-bearing units of a language and we know that segmenting words in terms of a good morphology can decrease the OOV rate significantly. A recent study [15] used a high-quality supervised morphology system to provide alternative segmentations for Turkish speech. As they demonstrate, even though many of the OOV keywords can be represented in terms of such morphemes, there were still a significant number of keywords (which consisted of "new" morphemes) that could not be seg-

[1]An OOV keyword is defined as a keyword which contains at least one OOV word.

mented. This is, of course, a consequence of sparsity. If, instead, the morphemes were broken down into smaller pieces (without necessarily adhering to linguistic rules) the generalization would have been better. This conclusion motivates us to consider alternatives (similar to those of [6]) which are able to represent virtually any possible OOV word, without necessarily going to the extreme of using a phonetic recognizer.

Performance of keyword spotting is measured using the Average Term Weighted Value (ATWV) [19], which is defined as

$$\text{ATWV} = 1 - \frac{1}{K}\sum_{w=1}^{K}\left(\frac{\#miss(w)}{\#ref(w)} + \beta\,\frac{\#fa(w)}{T - \#ref(w)}\right) \quad (1)$$

where K is the total number of keywords, $\#miss(w)$ is the number of true tokens of keyword w that are not detected, $\#fa(w)$ is the number of false detections of w, $\#ref(w)$ is the number of reference tokens of w, T is the total number of trials (for this challenge, approximated by the duration of the test audio in seconds), and β is a constant, set at 999.9.

3. APPROACH

Two different approaches to unit construction are experimented with: the first (Section 3.1) attempts to approximate syllables, using the phonemic structure of the word to guide decisions. The second (Section 3.2) attempts to approximate morphemes, and is therefore guided more strongly by the orthographic structure of the word. In both cases, language-independent algorithms are developed to extract sub-word units from a given set of words in an automated fashion.

For each set of units an individual speech recognition / spoken term detection (STD) system is developed, following the system description in [1, 11]. While we report on the performance of individual systems, our real interest lies in the effect when combining different sub-word systems; our approach to system combination is described in Section 3.3.

3.1. Syllable-based units

During the first two years of the IARPA BABEL project, pronunciation lexicons with syllable-marked pronunciations for all words were provided as part of each language pack. Letter-to-phoneme (L2P) models were trained to predict not only the pronunciation of new words, but also syllable boundaries. Starting with year 3 of the program, pronunciation lexicons were omitted from the language packs. To compensate, simplified L2P maps were derived from the language-specific peculiarities (LSP) document to create basic pronunciation lexicons [20, 21]. An automatic syllabification algorithm was also developed to generate syllables for sub-word modeling [20].

3.1.1. Generating longer syllables

For all year 4 languages, syllables were initially letter-based as opposed to phoneme-based. While the phoneme-based syllables provided comparable gains for OOV words on most languages, there was a significant drop in ATWV for Pashto, which has unwritten vowels. The main difference between the letter- and phoneme-based syllables for Pashto was thus syllable length and hence the number of unique syllables; the letter-based syllables were generally longer (in terms of phones), and there were more unique syllables. This led us to investigate the use of longer and more unique syllables for all languages.

Lang	Number of unique syllables				
	v1	v2	v3	v4	v5
Amharic	467	11,304	38,326	72,340	76,475
Guarani	8,442	26,707	46,483	59,607	67,639
Igbo	16,582	32,034	50,256	59,920	63,333
Javanese	10,246	36,141	51,598	60,198	63,336
Dholuo	15,463	33,055	45,719	58,390	64,201
Mongolian	10,094	31,510	39,672	47,407	55,072
Pashto	46,643	57,671	60,290	62,494	63,116

Table 1. The number of unique syllables using the original approach ($v1$), and after reclassifying one additional vowel per iteration ($v2$ to $v5$).

There are many approaches to generating longer and more unique syllables; our main goal was to develop a consistent, language-independent approach to generating syllables without having to involve a language expert. Pashto's unwritten vowels led us to experiment with changing the vowel/consonant classification of some phonemes to artificially generate longer and more unique syllables. An iterative approach was followed to determine which vowel contributes to the biggest increase in unique syllables if re-classified as a consonant (an approach that consistently leads to increasing the number of unique syllables – see Table 1). In the extreme case of relabeling all vowels as consonants, no syllabification would be possible, and all the words would be "syllables".

Our syllabification algorithm generally creates open syllables. As a consequence, many word-initial syllables are single-phoneme vowel-only syllables. In a further experiment, in order to reduce the number of short syllables, all word-initial vowel-only syllables were merged with the following syllable, as discussed further in Section 5.

3.2. Morpheme-based units

We use a publicly available implementation of MorphoChain [22] to generate morpheme-based units. MorphoChain considers the process of word formation as a sequence of morphological changes applied to a root word. For example, the word *beautifully* is formed from the word *beauty* in steps of *beauty* → *beautiful* → *beautifully*. This chain can be broken down into pairs of words – for instance, (*beautiful*, *beauty*), where *beauty* is referred to as the *parent* of *beautiful*. The probability distribution over word pairs is represented using a log-linear model. A word's segmentation can be obtained from its chain in a straightforward manner. We use MorphoChain on the *phonemic* representations of words to derive segmentations, which are used directly in our experiments.

3.3. Combining different sub-word approaches

The combination approach we used for coming up with a single hit list is the one described in [8]. Briefly stated, it proceeds in an incremental fashion, combining two system outputs at a time. For each pair of systems, it finds hits which are overlapping in time (up to a minimum proportion) and then merges them into a single hit that has the extremal times of the two hits as its times. The score of the new hit is a linear combination of the scores of the combined hits. The weights of the linear combination are trained using Powell's method [23], with the objective of maximizing the ATWV on a development set.

Lang	ATWV (IV / OOV / ALL) v2	v3	v4	v5
Amharic	0.513 / 0.704 / 0.534	0.640 / 0.731 / 0.650	0.659 / 0.716 / 0.665	0.665 / 0.719 / 0.670
Guarani	0.565 / 0.741 / 0.585	0.588 / 0.721 / 0.603	0.602 / 0.736 / 0.618	0.606 / 0.716 / 0.619
Igbo	0.370 / 0.477 / 0.381	0.392 / 0.495 / 0.402	0.399 / 0.496 / 0.409	0.404 / 0.503 / 0.414
Javanese	0.480 / 0.559 / 0.485	0.500 / 0.534 / 0.503	0.501 / 0.554 / 0.504	0.507 / 0.545 / 0.510
Dholuo	0.631 / 0.713 / 0.638	0.647 / 0.726 / 0.654	0.664 / 0.724 / 0.669	0.664 / 0.719 / 0.668
Mongolian	0.514 / 0.533 / 0.516	0.526 / 0.568 / 0.529	0.536 / 0.5720 / 0.539	0.545 / 0.556 / 0.546
Pashto	0.476 / 0.479 / 0.476	0.481 / 0.468 / 0.480	0.482 / 0.479 / 0.481	0.482 / 0.481 / 0.482

Table 2. ATWV for all IARPA BABEL year 4 languages as one (v2) to four (v5) different vowels are reclassified as consonants, for the purposes of creating syllable-like units.

4. EXPERIMENTAL SETUP

The corpora that we used in our experiments are the "Full language packs" that were distributed in the fourth year of the IARPA project BABEL, each containing approximately 40 hours of training data. The languages under investigation consisted of seven so-called development languages – Amharic (amh), Dholuo (luo), Guarani (grn), Igbo (ibo), Javanese (jav), Mongolian (mon) and Pashto (pus) – and the final evaluation language, Georgian (geo)[2].

The main recognition system was a variant of the Kaldi toolkit [24]. The search pipeline involved various modes of search, as described in [1, 11]. Syllables were trained on a vocabulary which included all words from the orthographic transcriptions, extended with web data. For the development languages we extend the vocabulary with 50k unique words obtained from web data; for the "surprise" evaluation language (Georgian), the amount of web data is increased significantly. All results are reported on the official development sets, using the official development keywords generated by IBM and BBN.

5. RESULTS

We first consider syllable-based sub-word unit performance for the development languages, when using an 50K vocabulary from web data. (See Section 4.) The effect of generating longer, more unique syllables during sub-word construction is shown in Table 2. The set empirically determined to provide the best performance for OOVs was $v4$, which is the set generated by reclassifying three vowels as consonants. From Table 2 it can be seen that all languages benefited from this approach.

These syllables could be further improved for some of the languages by merging vowel-only initial syllables with the following syllable, as shown in Table 3. All languages showed a small gain, except for Dholuo (small decrease) and Mongolian (larger decrease).

Table 4 shows the best comparable results for the different types of sub-word units, whole-word systems, as well as a combination of all three approaches. As expected, for IV keywords the whole-word systems' performance is better than either syllables or morphemes, while the syllable- and morpheme-based systems outperform whole words for OOV keywords[3]. By combining system results, improvements for both IV and OOV keywords are observed (OOV keywords have higher ATWV's than IV keywords as they are generally longer words which are easier to detect correctly than shorter words[25]).

[2]The data releases used: IARPA-babel307b-v1.0b (amh), IARPA-babel403b-v1.0b (luo), IARPA-babel305b-v1.0b (grn), IARPA-babel306b-v2.0c (ibo), IARPA-babel402b-v2.0b (jav), IARPA-babel401b-v2.0b (mon), IARPA-babel104b-v0.4bY(pus), IARPA-babel404b-v1.0a (geo)

[3]The whole-word sytem handles OOVs by doing approximate matches on whole-word lattices: see Section 2.

Lang	ATWV (IV / OOV / ALL) v3 (without)	v3 (with)
Amharic	0.640 / 0.731 / 0.650	0.652 / 0.741 / 0.662
Guarani	0.588 / 0.721 / 0.603	0.591 / 0.728 / 0.606
Igbo	0.392 / 0.495 / 0.402	0.399 / 0.501 / 0.409
Javanese	0.500 / 0.534 / 0.503	0.507 / 0.561 / 0.511
Dholuo	0.647 / 0.726 / 0.654	0.651 / 0.717 / 0.656
Mongolian	0.526 / 0.568 / 0.529	0.532 / 0.544 / 0.532
Pashto	0.481 / 0.468 / 0.480	0.481 / 0.476 / 0.481

Table 3. ATWV for all IARPA BABEL year 4 development languages with and without vowel-initial syllable merging.

Interestingly, when comparing OOV results, whether to use syllables or morphemes becomes a language-specific choice. For Amharic, Dholuo and Pashto, the morpheme-based OOV results are best, while for the rest of the languages, syllable-based OOV results are slightly better (Guarani, Igbo, Javanese) to significantly better (Mongolian).

Results for the development languages are summarized in Figures 1 and 2. From Fig. 1 it is clear that performance increases as the syllables become longer and more unique – eventually approximating performance of the whole word models, as expected. The one outlier is Amharic: when the standard algorithm is applied to Amharic, the number of syllables generated is substantially less than for the other languages, resulting in poor performance.

For the IARPA BABEL evaluation, the best system to use was selected using a development set. For the final evaluation language (Georgian) the syllable-based system performed best: in Table 5 we display the results for the morpheme-based and two syllable-based systems ($v3$ and $v4$, as above), using a system that is comparable to the one used for the other seven languages. This time the web data vocabulary was increased well beyond the initial 50k.

The Georgian systems are described in more detail in [26]. The final single-best system submitted for the BABEL evaluation incorporated the syllable-based units described in this paper. When incorporating additional techniques, ATWVs of 0.738 (IV) and 0.827 (OOV) were attained [26].

6. CONCLUSION

STD systems benefit significantly from the use of sub-word systems. We show that, while these systems are particularly useful for OOV keywords, they also offer gains for IV keywords. Different language-independent sub-word systems were presented: syllable-like systems and morpheme-based systems. Apart from reporting new results on the 2016 BABEL corpora, an additional contribution relates to the automated construction of syllable-like units: we demonstrate

Lang	ATWV (IV / OOV) Whole-word	Syllables (v4)	Morphemes	Combination
Amharic	0.677 / 0.675	0.659 / 0.716	0.656 / 0.728	0.680 / 0.744
Guarani	0.615 / 0.689	0.602 / 0.736	0.560 / 0.733	0.615 / 0.746
Igbo	0.412 / 0.486	0.399 / 0.496	0.362 / 0.491	0.414 / 0.507
Javanese	0.515 / 0.553	0.501 / 0.554	0.427 / 0.550	0.519 / 0.589
Dholuo	0.670 / 0.722	0.664 / 0.724	0.611 / 0.736	0.673 / 0.756
Mongolian	0.585 / 0.528	0.536 / 0.572	0.511 / 0.536	0.587 / 0.596
Pashto	0.504 / 0.500	0.482 / 0.479	0.438 / 0.498	0.504 / 0.514

Table 4. ATWV for all IARPA BABEL year 4 development languages for whole-word, syllable and morpheme based sub-word modeling approaches respectively, as well as a combination of all three approaches.

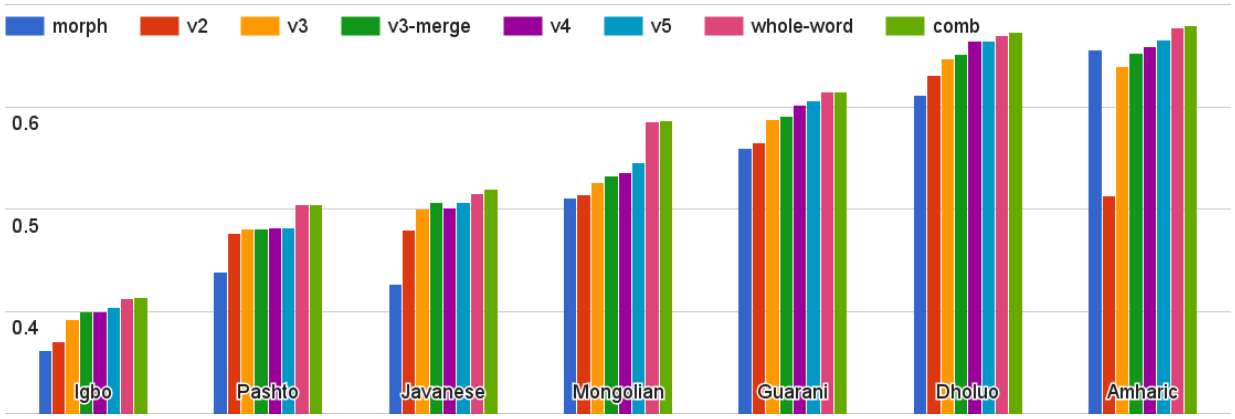

Fig. 1. *Comparing ATWVs across languages and approaches for in vocabulary (IV) keywords.*

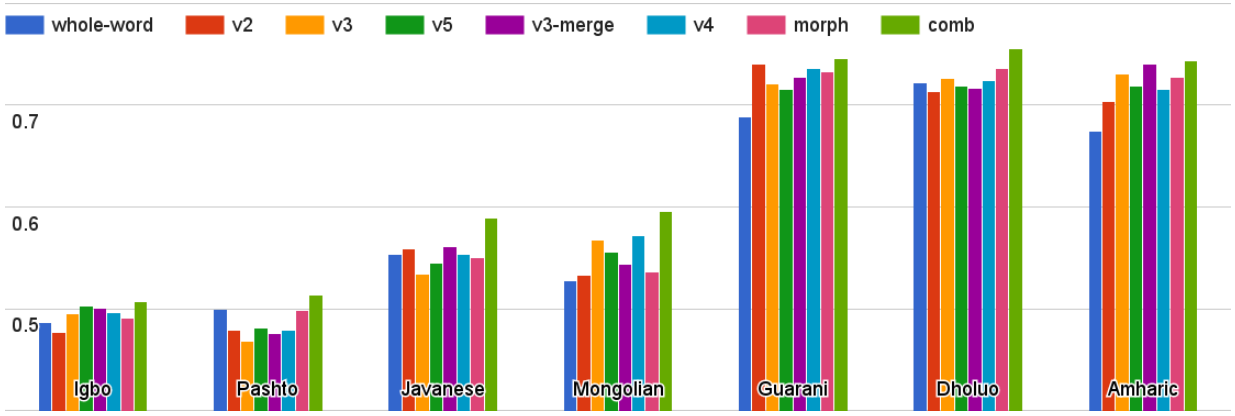

Fig. 2. *Comparing ATWVs across languages and approaches for out of vocabulary (OOV) keywords.*

Vocabulary	ATWV (IV / OOV) Morphemes	Syllables (v3)	Syllables (v4)
50k	0.681 / 0.802	0.709 / 0.808	0.718 / 0.785
100k	0.681 / 0.806	0.708 / 0.804	0.721 / 0.798
150k	0.680 / 0.805	0.708 / 0.806	0.722 / 0.807
1mill	-	-	0.731 / 0.840

Table 5. ATWV for Georgian when comparing morpheme-based, $v3$ and $v4$ syllable-based sub-word modeling approaches when increasing the web data vocabulary from 50k to 1 million additional words.

that by artificially increasing the number of unique syllables, significant improvements in STD performance can be observed.

7. ACKNOWLEDGMENTS

We would like to thank all members of the Babelon team at Raytheon BBN Technologies, and especially Tanel Alumae, William Hartmann and Stavros Tsakalidis. 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 expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

8. REFERENCES

[1] Damianos Karakos and Richard Schwartz, "Subword and phonetic search for detecting out-of-vocabulary keywords," in *Proc. Interspeech*, 2014.

[2] William Hartmann, Lori Lamel, and Jean-Luc Gauvain, "Cross-word subword units for low-resource keyword spotting," in *Proc. SLTU*, 2014.

[3] Yanzhang He, Peter Baumann, Hao Fang, Brian Hutchinson, Aaron Jaech, Mari Ostendorf, Eric Fosler-Lussier, and Janet Pierrehumbert, "Using pronunciation-based morphological subword units to improve OOV handling in keyword search," *IEEE/ACM Transactions on Audio Speech and Signal Processing*, vol. 24, no. 1, pp. 79–92, 2016.

[4] William Hartmann, Viet-Bac Le, Abdel Messaoudi, Lori Lamel, and Jean-Luc Gauvain, "Comparing decoding strategies for subword-based keyword spotting in low-resourced languages," in *Proc. Interspeech*, 2014.

[5] Jan Trmal, Guoguo Chen, Dan Povey, Sanjeev Khudanpur, Pegah Ghahremani, Xiaohui Zhang, Vimal Manohar, Chunxi Liu Aren Jansen, Dietrich Klakow, David Yarowsky, and Florian Metze, "A keyword search system using open source software," in *Proc. SLT Workshop*, 2014.

[6] Ivan Bulyko, José Herrero, Chris Mihelich, and Owen Kimball, "Subword speech recognition for detection of unseen words," in *Proc. Interspeech*, 2012.

[7] Mike Schuster and Kaisuke Nakajima, "Japanese and Korean voice search," in *Proc. ICASSP*, Kyoto, Japan, March 2012, pp. 5149–5152.

[8] Damianos Karakos, Richard Schwartz, Stavros Tsakalidis, Le Zhang, Shivesh Ranjan, Tim Ng, Roger Hsiao, Guruprasad Saikumar, Ivan Bulyko, Long Nguyen, et al., "Score normalization and system combination for improved keyword spotting," in *Proc. ASRU*. IEEE, 2013.

[9] Guoguo Chen, Ozgur Yilmaz, Jan Trmal, Daniel Povey, and Sanjeev Khudanpur, "Using proxies for oov keywords in the keyword search task," in *Proc. ASRU*. IEEE, 2013.

[10] Murat Saraclar, Abhinav Sethy, Bhuvana Ramabhadran, Lidia Mangu, Jia Cui, Xiaodong Cui, Brian Kingsbury, and Jonathan Mamou, "An empirical study of confusion modeling in keyword search for low resource languages," in *Proc. ASRU*. IEEE, 2013.

[11] Damianos Karakos and Richard M Schwartz, "Combination of search techniques for improved spotting of OOV keywords," in *Proc. ICASSP*. IEEE, 2015.

[12] Hang Su, Van Tung Pham, Yanzhang He, and James Hieronymus, "Improvements on transducing syllable lattice to word lattice for keyword search," in *Proc. ICASSP*, 2015.

[13] D. R.H. Miller, M. Kleber, C.-L. Kao, O. Kimball, T. Colthurst, S. A. Lowe, R. M. Schwartz, and H. Gish, "Rapid and accurate spoken term detection," in *Proc. Interspeech*, 2007.

[14] J. Mamou, B. Ramabhadran, and O. Siohan, "Vocabulary independent spoken term detection," in *Proc. SIGIR*, Amsterdam, The Netherlands, July 2007.

[15] Karthik Narasimhan, Damianos Karakos, Richard Schwartz, Stavros Tsakalidis, and Regina Barzilay, "Morphological segmentation for keyword spotting," in *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, October 2014, pp. 880–885, Association for Computational Linguistics.

[16] F. Seide, P. Yu, C. Ma, and E. Chang, "Vocabulary-independent search in spontaneous speech," in *Proc. ICASSP*, 2004.

[17] A. Rastrow, A. Sethy, and B. Ramabhadran, "A new method for OOV detection using hybrid word/fragment system," in *Proc. ICASSP*, 2009.

[18] A. Rastrow, A. Sethy, B. Ramabhadran, and F. Jelinek, "Towards using hybrid word and fragment units for vocabulary independent LVCSR systems," in *Proc. Interspeech*, 2009.

[19] NIST, "OpenKWS13 keyword search evaluation plan," http://www.nist.gov/itl/iad/mig/upload/OpenKWS13-EvalPlan.pdf, 2013.

[20] Marelie Davel, Damianos Karakos, Etienne Barnard, Charl van Heerden, Richard Schwartz, Stavros Tsakalidis, and William Hartmann, "Exploring minimal pronunciation modeling for low resource languages," in *Proc. Interspeech*, Dresden, Germany, September 2015, pp. 538–542.

[21] Neil Kleynhans, William Hartman, Daniel van Niekerk, Charl van Heerden, Rich Schwartz, Stavros Tsakalidis, and Marelie Davel, "Code-switched English pronunciation modeling for Swahili spoken term detection," in *Proc. Workshop on Spoken Language Technologies for Under-resourced languages (SLTU)*, Yogyakarta, Indonesia, May 2016, pp. 128–135.

[22] Karthik Narasimhan, Regina Barzilay, and Tommi Jaakkola, "An unsupervised method for uncovering morphological chains," *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 157–167, 2015.

[23] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes: The art of Scientific Computing*, Cambridge University Press, 2007.

[24] Dan Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlcek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The Kaldi speech recognition toolkit," in *Proc. ICASSP*, 2011.

[25] William Hartmann et al., "Analysis of Keyword Spotting Performance across IARPA BABEL Languages," in *Proc. ICASSP (accepted for publication)*, New Orleans, USA, March 2017.

[26] Tanel Alumae et al., "The 2016 BBN Georgian telephone speech keyword spotting system," in *Proc. ICASSP (accepted for publication)*, New Orleans, USA, March 2017.