ANALYSIS OF KEYWORD SPOTTING PERFORMANCE ACROSS IARPA BABEL LANGUAGES

William Hartmann[*]*, Damianos Karakos, Roger Hsiao, Le Zhang*
Tanel Alumäe[†]*, Stavros Tsakalidis, Richard Schwartz*

Raytheon BBN Technologies, Cambridge, MA, USA
{william.hartmann, damianos.karakos, le.zhang, stavros.tsakalidis, rich.schwartz}@raytheon.com

ABSTRACT

With the completion of the IARPA Babel program, it is possible to systematically analyze the performance of speech recognition systems across a wide variety of languages. We select 16 languages from the dataset and compare performance using a deep neural network-based acoustic model. The focus is on keyword spotting using the actual term-weighted value (ATWV) metric. We demonstrate that ATWV is keyword dependent, and that this must be accounted for in any cross-language analysis. Further, we show that while performance across languages does not track with any particular feature of the language, it is correlated with inter-annotator agreement.

Index Terms— ATWV, babel, cross-language analysis

1. INTRODUCTION

There is a dearth of analysis in cross-language performance for automatic speech recognition (ASR) systems. While it is common to test methods on datasets from multiple languages [1, 2], the differences between the languages can not be interpreted due to the variation in the data sets. Instead they can demonstrate how a method can work better for certain languages. Killer [3] analyzed performance using grapheme-based lexicons compared to phonetic lexicons for several languages—English performs poorly while Spanish does not suffer. Zhang et al. [4] showed that availability of web data and the gains from using web data was inconsistent across languages.

Much of the work has focused on multilingual speech recognition. Huang et al. [5] used multilingual training for a variety of languages. Adding multilingual data using their shared hidden layer structure improved all languages, but the gains varied depending on language. Knill et al. [6] showed similar results for languages in the IARPA Babel program. Recent work has shown that biasing multilingual training to similar languages—either at the corpus level, or even at the frame level—can improve performance compared to using a larger variety of data [7]. In all cases, variations across languages are seen, but the specific causes are not investigated.

One reason for the scarcity of analysis in this area is the lack of data. It is very difficult to collect equal amounts of data in the same manner for multiple languages. Prior to the IARPA Babel Program—a program designed to encourage research in resource-limited keyword spotting in a variety of languages—the Globalphone [8] database may have been the closest approximation. Data for 20 languages were collected in a similar manner to the commonly used Wall Street Journal corpus. Native speakers read articles from newspapers. This had the added benefit of eliminating the transcription requirement.

We explore possible reasons for the variance in performance across languages. The IARPA Babel data is ideal for this analysis due to the attempt to minimize confounding factors in the data collection. We show a wide range in performance across languages. While we are unable to derive a single factor explaining the variation, we show that performance is correlated with human transcription accuracy on the same data. For keyword spotting (KWS), we show that performance is highly dependent on the keyword selection. Any cross-language analysis must take this into account.

2. THE IARPA BABEL DATA AND ATWV

The IARPA Babel program recently completed its fourth and final year. The principal objective of Babel was to develop a KWS system that delivers high accuracy for any new language, in the face of very limited transcribed speech, noisy acoustic and channel conditions, and limited system build time of one week. Each year participants produced systems for an increasing number of languages. Upon completion, the IARPA Babel dataset consisted of 25 languages. We focus on the fifteen languages from the third and fourth year of the program that contained approximately 40 hours of transcribed speech for training: Amharic, Cebuano, Dholuo, Georgian, Guarani, Igbo, Javanese, Kazakh, Kurdish, Lithuanian, Mongolian, Pashto, Swahili, Telugu, and Tok Pisin [1]. We also include results from a 40 hour variant of the second year language Tamil.

The amount of training data available for the primary condition varied throughout the conditions and years. Our focus is on the full language pack (FLP) from the third and fourth year. This definition of the FLP assumes 40 hours of transcribed training data. The data consists of conversational telephone speech collected in a variety of environments. For added difficulty, approximately 15% of the conversations were actually recorded with a far-field microphone.

The principle evaluation metric is actual term-weighted value (ATWV). ATWV is computed for each keyword individually and then averaged across all keywords to produce the final result. The range of ATWV is $-\infty$ to 1; an incorrect hypothesis is worse than

[*]This research was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US 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.

[†]Now at Tallinn University of Technology, Estonia

[1]IARPA-babel{307b-v1.0b, 301b-v2.0b, 403b-v1.0b, 404b-v1.0a, 305b-v1.0a, 306b-v2.0c, 402b-v1.0b, 302b-v1.0a, 205b-v1.0a, 304b-v1.0b, 401b-v2.0b, 104b-v0.4bY, 202b-v1.0d, 303b-v1.0a, 207b-v1.0b, 204b-v1.1b}

Hours of Transcribed Audio	WER	ATWV
40 hours	47.8	0.423
80 hours	44.4	0.487

Table 1. Results demonstrating the effect of the amount of transcribed audio for Pashto.

no hypothesis. See [9], for a more detailed description of ATWV. In the case of WER, all words are treated equally. Short words are as important as long words, and common words are as important as rare words. While each keyword is treated equally for ATWV, each detection is not. The detection of a rare keyword is worth more than the detection of a common keyword for any particular instance—all false accepts are equal. This property of ATWV is important when analyzing across languages as we will see in Section 5.

3. EXPERIMENTAL SETUP

We use the Sage ASR toolkit [10] for building all systems. Sage is BBN's newly developed speech-to-text transcription (STT) platform that integrates technologies from multiple sources, each of which has a particular strength. In Sage, we combine proprietary sources, such as BBN's Byblos [11], with open source toolkits, such as Kaldi [12] and CNTK [13]. Sage also includes a cross-toolkit FST recognizer that supports models built using the various component technologies, and software supporting keyword search from Byblos [14, 15, 16].

The training recipe for the ASR system is constant across all languages. MLP bottleneck features (BN) are trained on 32-dimensional filterbanks plus pitch features. The 40-dimensional BN features are used for speaker-adaptive training with the final features being fMLLR-transformed bottleneck features. The final acoustic model is a six hidden layer DNN with 2048 nodes in each hidden layer and approximately 4500 nodes in the output layer. Lexicons are derived using simple G2P rules [17]. A trigram language model trained only the acoustic transcripts was used during decoding.

Decoding is performed on 10 hours of development data, and search is performed on approximately 2000 keywords. Both whole word and phonetic search are used [18]. We note that performance can be significantly increased through the addition of web data [4], data augmentation [19], joint decoding [20], and multilingual features [21]. However, it was beyond the scope of the study to build the best possible system for each language. Our goal was to build a competitive system that could be compared across languages.

4. RESULTS

We focused only on the FLP condition for the 15 year 3 and year 4 languages. The two previous years used between 60 and 80 hours, while the final two years used 40 hours of transcribed speech. Pashto, originally a first year language, was the only language with both an 80 hour and 40 hour FLP set defined. In Table 1 we show the performance difference between the two training sets. Doubling the amount of training data—this also implicitly increases the size of the vocabulary and strength of the language model—provides large gains in both ATWV and WER. Given the gains from additional data alone, it was important to limit each language to the same amount of training data in order to obtain a fair comparison.

In Figure 1 we plot the ATWV vs. WER for each of the 15 languages. While there is clearly a relationship between the two measures, there is still much variance. Pashto and Georgian have approximately the same WER, but ATWV performance is more than 20 points apart. The ATWV for Pashto and Telugu are nearly identical,

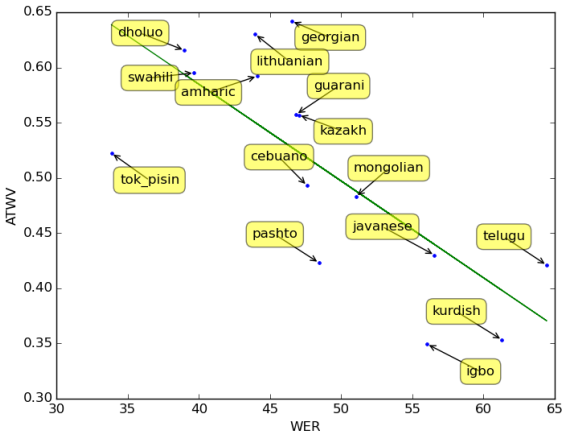

Fig. 1. ATWV vs. WER for 15 languages from year 3 and year 4. Line represents best linear fit ($r = 0.747$).

but their WER is more than 15 points apart. Tok Pisin, in particular, is an outlier with the best WER by far, but only middling ATWV. The point is that what makes a language perform well on one metric, may not apply to other metrics. Further, when making comparisons across languages, the metric used is important. We note that Gales et al. [22] found a much stronger relationship using a different subset of languages—five languages from year 2. This may be an artifact of the limited number of languages; we can see similar results when limiting the total number of languages considered.

5. KEYWORD SELECTION AND ATWV

Since ATWV is not a measure of all words, but a specific subset, it is variable based on the keyword selection. Looking at features of the keywords, we analyze how performance changes with respect to the keyword variation. Depending on the feature, the variability in performance is greater within a language than across languages.

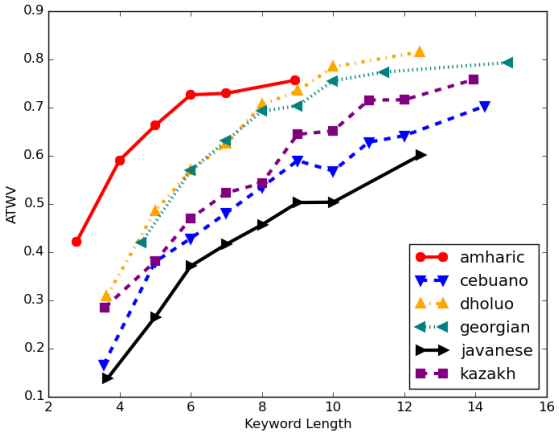

Fig. 2. ATWV vs. Keyword Length

In Figure 2 we show how ATWV changes with respect to keyword length in characters—given our lexicons use simple G2P mappings, length in characters is similar to length in phones. We only

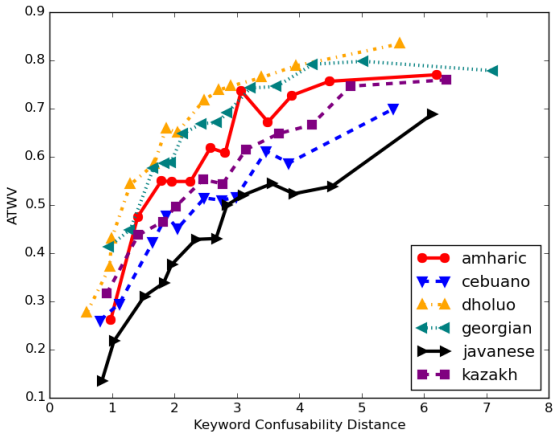

Fig. 3. ATWV vs. Confusability Distance

show results for a subset of languages as the figure becomes inscrutable with too many languages; performance on languages not shown follows the same pattern. The difference in performance for long and short keywords can be dramatic, far greater than the difference in performance between the overall ATWV between any two languages. Also, note that the gap between languages is maintained as the keyword length varies.

A similar result is shown in Figure 3—ATWV vs. keyword confusability. We define keyword confusability as the average minimum Levenshtein distance for a keyword in each utterance. It can be thought of as a weighted keyword length. Lower values indicate a more confusable keyword. Again, as the value increases, so does the ATWV. Larger, less confusable keywords are easier to detect.

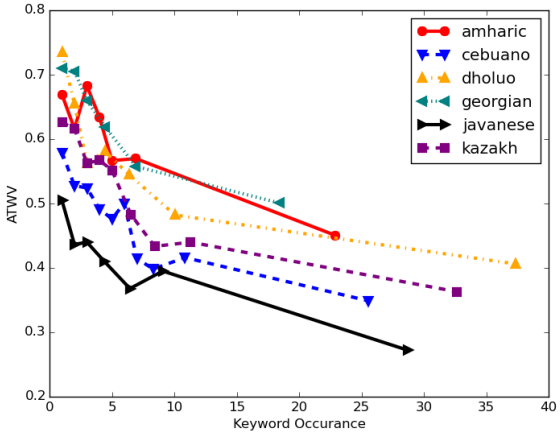

Fig. 4. ATWV vs. Keyword Occurrence

In addition to these keyword intrinsic features, other features more related to the test set are important. Due to the formulation of ATWV, the penalty for missing a keyword is inversely proportional to the frequency of the keyword, while the penalty for a false alarm is constant. It is more important to recognize a rare keyword than a keyword with many occurrences. This builds an inherent bias towards detecting rare words. Figure 4 highlights this bias. Per-

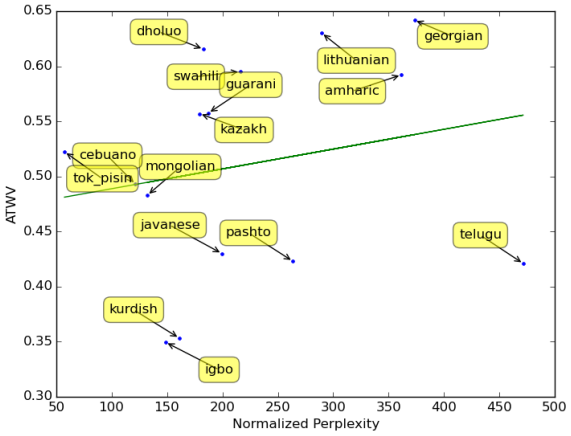

Fig. 5. ATWV vs. Normalized Perplexity. Line represents best linear fit ($r = 0.202$).

formance on rare keywords is higher than for common keywords. This bias also fits the goal of the program as more information-rich keywords are likely to be less frequent. Note that all of these measures are related. We found that all pairs of features—ATWV, length, confusability distance, and number of references—are strongly correlated across all languages. We also note these features account for the differences in IV and OOV performance.

As a collection, these figures show that ATWV is strongly related to keyword selection. Some of these features can potentially be related to the difficulty of a language for KWS. However, it is difficult to make a case that some languages inherently contain more frequent keywords, thereby making them more difficult. In addition these figures demonstrate that while the variation on ATWV based on these features is large, it does not completely account for the differences in performance across languages. Regardless of the feature chosen, Dholuo always outperforms Javanese.

6. CROSS-LANGUAGE ANALYSIS

We examined many properties of each language in an attempt to correlate them with performance—e.g. average word length, size of phonetic inventory, and out-of-vocabulary rate—but failed to find any strong relationship. As an example, we show ATWV vs normalized perplexity in Figure 5. Normalized perplexity is simply perplexity normalized by the average length of words in the language; the unnormalized plot also shows a similar lack of correlation. While all of these features are most certainly factors, there are many of them interacting together. We could find a weighted combination to fit the results even better, but this would greatly increase the probability of overfitting to the small number of languages available. While the total number of languages is large compared to other studies and collections, it is still a small number of data points for analysis.

There are also additional factors we can not measure. Obviously the data for each language was collected in different locations and countries. We do not know the relative difficulty in data collection for each country, the reliability of the cellular networks, the degree of nativeness for each speaker, the regularity of the writing system, or any other of a large number of factors that could be driving the differences in performance. It is nearly impossible to separate out the factors inherent to a language with those that are merely an artifact of the societal conditions and the country of origin. We need a single

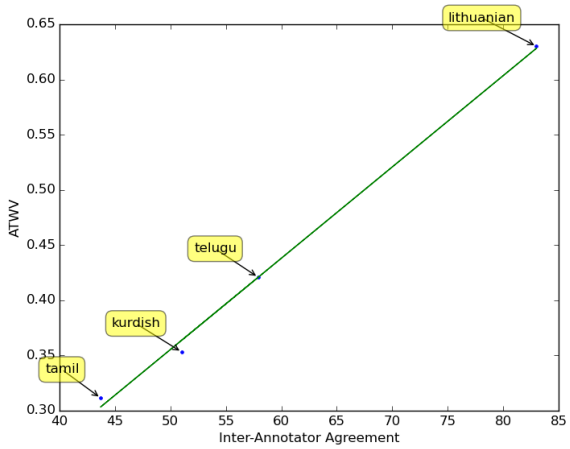

Fig. 6. ATWV vs. Inter-Annotator Agreement. Line represents best linear fit ($r = 0.998$).

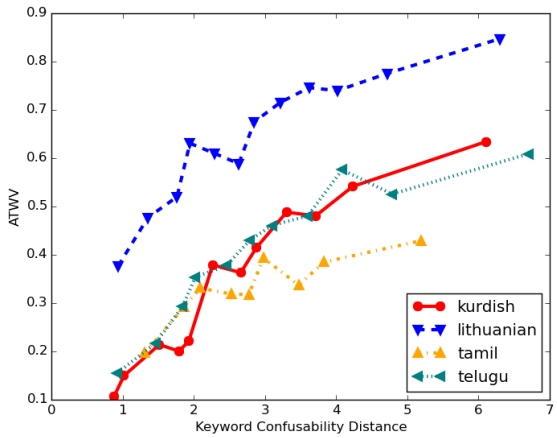

Fig. 7. ATWV vs. Keyword Confusability of the 4 language subset

measure that is a correlate of all these factors.

7. MEASURING HUMAN PERFORMANCE

One benchmark against which to compare system performance is human performance on the same task. It would be good to know, for example, if data from certain languages are more difficult to transcribe than data from other languages. If so, then we would expect to see those differences in performance reflected in system performance on the same task. If there are some fundamental differences, such as poor signal quality, poor speaking quality, lack of a well-defined writing system, etc., we will be able to measure this by an increase in the inconsistency in transcriptions by multiple transcribers.

To this end, after the initial data collection and preparation, a second set of transcribers were used to retranscribe a subset of four languages: Lithuanian, Tamil, Telugu, and Kurdish[2]. Tamil is a language from year 2 and has not been included in our previous analysis. In order to analyze Tamil with respect to these new transcripts, we trained a Tamil system using a 40 hour subset—from the original 70 hours—of transcribed audio. After the transcription, it was noticed that three of the languages contained a large number of new words. The new transcriptions were renormalized to account for the alternate spellings of words. The inter-annotator agreement of the second transcriber compared to the first ranged from as high as 83% for Lithuanian to as low as 44% for Tamil—even after normalization.

Figure 6 shows the relationship between ATWV and inter-annotator agreement. Our performance is correlated with the degree of difficulty for a native transcriber to match the transcription of another native speaker. However, as we only have four data points, this analysis needs to be repeated with a larger number of languages before strong conclusions can be drawn. The correlation with WER is not as strong ($r = -0.943$). The amount of disagreement between native speakers also highlights one of the challenges in the IARPA Babel program. If the inter-annotator agreement is below 50%, how can a system be expected to provide accurate transcription?

The final two figures demonstrate how much the inter-annotator agreement accounts for the variability in performance. The first, Figure 7, shows the relationship between ATWV and keyword confusability for the four language subset. Even given the same confusability distance, differences can be as great as 30 points. In Figure 8 we

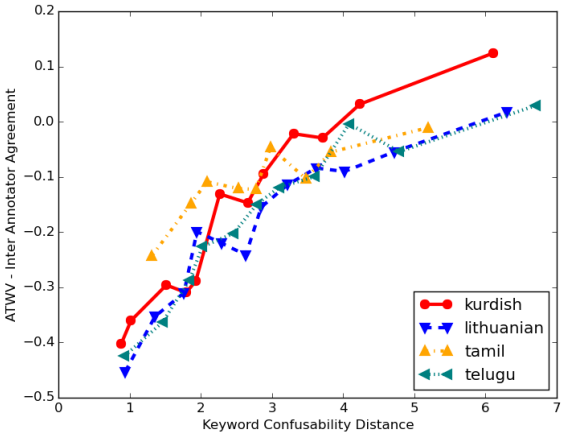

Fig. 8. Normalized ATWV vs. Keyword Confusability of the four language subset

normalize ATWV by the inter-annotator agreement by subtracting the agreement from the ATWV. The motivation is that performance can roughly only be as good as the level of inter-annotator agreement. The results are brought much closer together, and the difference between languages is no more than 10 points. Inter-annotator agreement may not explain cross-language variation, but it demonstrates a correlation between human and machine performance.

8. CONCLUSION

We built speech recognition systems for a total of 16 languages from the IARPA Babel program. The consistency of the data collection in this program allowed us to analyze performance across languages by focusing on the differences inherent to the language. We presented a detailed analysis of the relationship between keyword selection and ATWV, and how this can impact cross-lingual analysis. While we could not isolate a single language-dependent characteristic to explain the variation in performance across languages, we showed that performance is correlated with inter-annotator agreement. The factors that make it difficult for a native speaker to consistently transcribe speech also impact ASR systems.

[2]We thank the IARPA Babel T&E team for providing this data.

9. REFERENCES

[1] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Interspeech*, 2015.

[2] P. Golik, Z. Tüske, R. Schluter, and H. Ney, "Multilingual features based keyword search for very low-resource languages," in *Interspeech*, 2015, pp. 1260–1264.

[3] M. Killer, "Grapheme-based speech recognition," M.S. thesis, Carnegie Mellon University, 2003.

[4] L. Zhang, D. Karakos, W. Hartmann, R. Hsiao, R. Schwartz, and S. Tsakalidis, "Enhancing low resource keyword spotting with automatically retrieved web documents," in *Interspeech*, 2015, pp. 839–843.

[5] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proceedings of ICASSP*, 2013.

[6] K. Knill, M. Gales, A. Ragni, and S. Rath, "Language independent and unsupervised acoustic models for speech recognition and keyword spotting," in *Interspeech*, 2014, pp. 16–20.

[7] E. Chuangsuwanich, Y. Zhang, and J. Glass, "Multilingual data selection for training stacked bottleneck features," in *Proceedings of ICASSP*, 2016, pp. 5410–5414.

[8] T. Schultz, "Globalphone: a multilingual speech and text database developed at karlsruhe university," in *Proceedings of Interspeech*, 2002.

[9] S. Wegmann, A. Faria, A. Janin, K. Riedhammer, and N. Morgan, "The tao of ATWV: Probing the mysteries of keyword search performance," in *Proceedings of IEEE ASRU*, 2013, pp. 192–197.

[10] R. Hsiao, R. Meermeier, T. Ng, Z. Huang, M. Jordan, E. Kan, T. Alumäe, J. Silovsky, W Hartmann, F. Keith, O. Lang, M. Siu, and O. Kimball, "Sage: The new BBN speech processing platform," in *Interspeech*, 2016.

[11] S. Tsakalidis, R. Hsiao, D. Karakos, T. Ng, S. Ranjan, G. Saikumar, L. Zhang, L. Nyugen, R. Schwartz, and J. Makhoul, "The 2013 BBN Vietnamese telephone speech keyword spotting system," in *ICASSP*, 2014.

[12] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *ASRU*, 2011.

[13] D. Yu, A. Eversole, M. Seltzer, K. Yao, B. Guenter, O. Kuchaiev, F. Seide, H. Wang, J. Droppo, Z. Huang, Y. Zhang, G. Zweig, C. Rossbach, J. Currey, J. Gao, A. May, A. Stolcke, and M. Slaney, "An introduction to computational networks and the computational network toolkit," Tech. Rep., Microsoft Research, 2014.

[14] D. Karakos, I. Bulyko, R. Schwartz, S. Tsakalidis, L. Nguyen, and J. Makhoul, "Normalization of phonetic keyword search scores," in *Proc. of ICASSP*, Florence, Italy, 2014.

[15] D. Karakos and R. Schwartz, "Combination of search techniques for improved spotting of OOV keywords," in *Proc. of ICASSP*, 2015.

[16] D. Karakos, R. Schwartz, S. Tsakalidis, L. Zhang, S. Ranjan, T. Ng, R. Hsiao, G. Saikumar, I. Bulyko, L. Nguyen, J. Makhoul, F. Grezl, M. Hannemann, M. Karafiat, I. Szoke, K. Vesely, L. Lamel, and V.-B. Le, "Score normalization and system combination for improved keyword spotting," in *Proc. of ASRU*, Olomouc, Czech Republic, 2013.

[17] M. Davel, E. Barnard, C. van Heerden, W. Hartmann, D. Karakos, R. Schwartz, and S. Tsakalidis, "Exploring minimal pronunciation modeling for low resource languages," in *Interspeech*, 2015, pp. 538–542.

[18] D. Karakos, R. Schwartz, S. Tsakalidis, L. Zhang, S. Ranjan, T. Ng, R Hsiao, G. Saikumar, I. Bulyko, L. Nyugen, J. Makhoul, F. Grezl, M. Hannemann, M. Karafiat, I. Szoke, K. Vesely, L. Lamel, and V.-B. Le, "Score normalization and system combination for improved keyword spotting," in *Proceedings of IEEE ASRU*, 2013.

[19] W. Hartmann, T. Ng, R. Hsiao, S. Tsakalidis, and R. Schwartz, "Two-stage data augmentation for low-resourced speech recognition," in *Proceedings of Interspeech*, 2016.

[20] W. Hartmann, L. Zhang, K. Barnes, R. Hsiao, S. Tsakalidis, and R. Schwartz, "Comparison of multiple system combination techniques for keyword spotting," in *Proceedings of Interspeech*, 2016.

[21] Tanel Alumäe, Stavros Tsakalidis, and Richard Schwartz, "Improved multilingual training of stacked neural network acoustic models for low resource languages," in *Interspeech*, 2016.

[22] M. J. F. Gales, K. Knill, A. Ragni, and S. Rath, "Speech recognition and keyword spotting for low resource languages: Babel project research at CUED," in *Proceedings of SLTU*, 2014.