RICH SYSTEM COMBINATION FOR KEYWORD SPOTTING IN NOISY AND ACOUSTICALLY HETEROGENEOUS AUDIO STREAMS

Murat Akbacak[1], Lukas Burget[2], Wen Wang[3], Julien van Hout[3]

[1]Microsoft, Sunnyvale, CA, U.S.A.
[2] Brno University of Technology, Czech Republic
[3]SRI International, Menlo Park, CA, U.S.A.

ABSTRACT

We address the problem of retrieving spoken information from noisy and heterogeneous audio archives using system combination with a rich and diverse set of noise-robust modules. Audio search applications so far have focused on constrained domains or genres and not-so-noisy and heterogeneous acoustic or channel conditions. In this paper, our focus is to improve the accuracy of a keyword spotting system in highly degraded and diverse channel conditions by employing multiple recognition systems in parallel with different robust frontends and modeling choices, as well as different representations during audio indexing and search (words vs. subword units). After aligning keyword hits from different systems, we employ system combination at the score level using a logistic-regression-based classifier. Side information such as the output of an acoustic condition identification module is used to guide system combination system that is trained on a held-out dataset. Lattice-based indexing and search is used in all keyword spotting systems. We present improvements in probability-miss at a fixed probability-false-alarm by employing our proposed rich system combination approach on DARPA Robust Automatic Transcription of Speech (RATS) Phase-I evaluation data that contains highly degraded channel recordings (signal-to-noise ratio levels as low as 0 dB) and different channel characteristics.[1]

Index Terms— Keyword spotting, spoken term detection, channel degradation, fusion, acoustic noise, robust audio search.

1. INTRODUCTION

Information search in audio recordings is expanding at an increasing rate as more audio data (e.g., audio broadcasts, archives from digital libraries, audio/video content on the Internet, meeting recordings) becomes available. Different audio search applications have been studied in the past, such as keyword spotting, spoken term detection [1], and spoken document retrieval [2]. These studies have mostly focused on constrained and somewhat acoustically homogeneous domains or genres and not-so-noisy acoustic conditions. When the searchable audio content is drawn from diverse and acoustically degraded sources, it is challenging to build robust and up-to-date audio search systems. System tuning to reduce acoustic mismatch could help to maintain retrieval performance at desired levels, although this can be a costly (time, labor, money) solution. Therefore, finding automatic ways to maintain audio search performance at desired levels across different acoustic conditions becomes a practical concern.

[1]Approved for Public Release, Distribution Unlimited.

The effect of acoustic condition mismatch and variation on spoken document retrieval performance has been heavily observed in audio search applications for digital archive projects such as [3, 4, 5], where there is a variety of different acoustic conditions, recording media, speakers, emotions, accents, and dialects. In these studies, the quality of automatic speech recognition (ASR) transcripts is improved via robust speech recognition methods (e.g., robust feature extraction, model adaptation, speech enhancement) to minimize the impact of acoustic mismatch or variation on retrieval performance. The strategy is to pick the best system configuration for all conditions without analyzing which frontend or modeling choice works best for what kind of acoustic condition. In [6], the authors cluster the acoustic conditions via an Environmental Sniffing module [7], taking a first step in this direction. Based on this side information, they decide the system combination and back-off weights during a parallel and hybrid search, respectively, for a spoken document retrieval task where they employ a single word-based and phonetic system. Although this approach uses side information to guide a system combination of word and phonetic systems, the way that the system combination is done is somewhat *ad-hoc*, and it does not investigate using several recognition systems with different features or modeling approaches in parallel. The approach cannot be extended to use other side information with soft decisions. On the other hand, score-level system combination has been heavily used for speaker [8], dialect [9], and language identification [10] systems. In [8], side information is used to guide system combination. In this study, we apply similar techniques to the keyword spotting task.

The DARPA RATS program deals with clean speech that has been degraded by transmission through eight different radio channels [11]. The resulting speech varies widely in quality and intelligibility, with various distortions, dropouts, frequency shifts, push-to-talk noise, and so on. The speech varies from somewhat intelligible to barely intelligible. The original speech was taken from the Levantine Fisher Corpus [12] which was produced at LDC by having different native speakers of Levantine Arabic speak on the telephone about different topics. To mitigate the problem of noisy and heterogeneous acoustic conditions in this scenario, we employ rich and diverse set of recognition (with several noise-robust features, as well as advanced modeling techniques) and keyword spotting systems (with different units) in parallel, and employ system combination at the end by using the output of an acoustic condition identification module as side information to allow adaptive and robust combination.

In Section 2, system components, namely speech recognition, indexing, metadata extraction, and system combination are presented. Evaluation of the proposed system combination methods is presented in Section 3. Discussion and future work are presented in Section 4. This is followed by conclusions in Section 5.

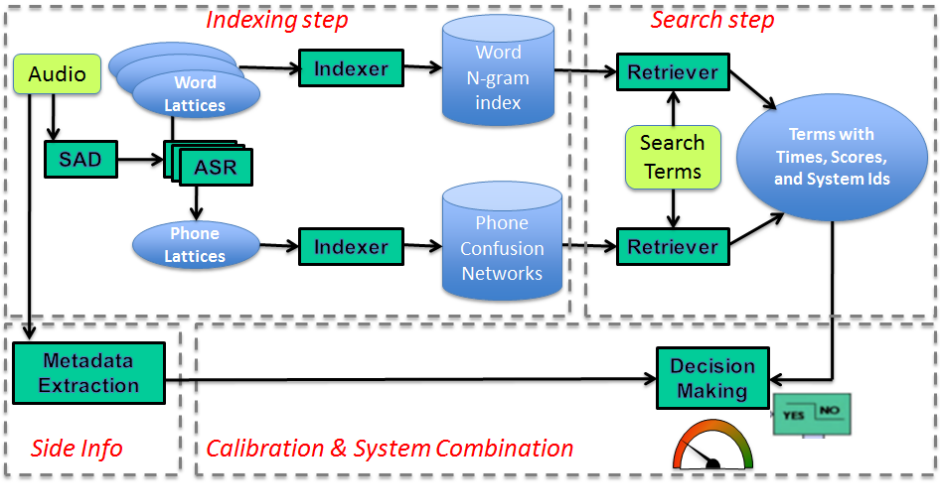

Fig. 1. Overview of proposed rich system combination approach fusing robust and diverse set of keyword spotting systems.

2. SYSTEM COMPONENTS

In our system, as shown in Figure 1, we employ multiple recognition systems with different robust front-ends, and modeling choices, as well as different representations during the audio indexing and search step. In the following sections, we first present the choices made for these different components of the keyword spotting systems. Then, we present the metadata or side information extraction module. Finally, we present the system combination module that is used to combine outputs from these rich and diverse keyword spotting systems, with or without side information.

2.1. Automatic Speech Recognition

The training data for both acoustic and language models comes from the RATS program. For acoustic modeling, we have around 50 hours of transcribed audio data in each of eight different channels. From this data, we choose the portion that has SNR bigger than 15 dB. After feature normalization, we trained maximum likelihood estimated (MLE) cross-word HMM-based acoustic models with speaker clustering and speaker adaptive training. The lexicon contains all non-singleton words from the training data. Grapheme-based pronunciations are used in the lexicon. Language models (LMs) are trained with modified Kneser-Ney smoothing [13] and include bigrams (for lattice generation) and trigrams (for lattice rescoring).

For the Decipher recognition systems, the features were normalized using standard cepstral mean and variance normalization, vocal tract length normalization (VTLN), and heteroscedastic linear discriminant analysis (HLDA). The cross-word MLE models were trained with decision-tree clustered states. When decoding the testsets, the cross-word MLE model was first adapted through maximum-likelihood linear regression (MLLR) using a phone-loop model as reference and then used for 1-best decoding. The cross-word MLE model was adapted again through MLLR on the 1-best decoding output and the adapted model and the bigram LM were used for generating HTK lattices. Speaker-clustered regression class trees were used to improve robustness of MLLR adaptation [14]. The bigram HTK lattices were then rescored with a trigram LM and the resulting trigram lattices were used for lattice indexing and search.

On the front-end side, in addition to conventional front-end features, such as MFCC and PLP, we employ noise-robust features such as NMCC, PNCC, and CSAWH. Further details of these front-end features and how they are extracted can be found in [15].

On the acoustic modeling side, in addition to standard modeling approaches, for example Gaussian Mixture Models (GMMs) as Hidden Markov Model (HMM) state density functions, where there is no parameter sharing between Gaussians and it is hard to adapt to new acoustic conditions with few training samples, a subspace GMM (SGMM) approach [16] where Gaussian parameters are projected into pre-trained low-rank subspaces is used for acoustic modeling. This allows fast acoustic adaptation to unseen data as SGMM provides very compact representation of complex distributions, which can be robustly trained with a limited amount of training data. We use the KALDI speech recognition toolkit [17] for SGMMs. For the GMM-based system, we use SRI's Decipher engine. We explore only multi-style training where data from all channels are pooled to train one model in each recognition system.

2.2. Lattice Indexing and Search

Since the lattice structure provides additional information about the correct hypothesis, compared to 1-best recognition output, to avoid misses (which are more likely to occur in noisy recordings such as RATS data), we employ word and phone lattices to generate the searchable index. During indexing, audio input is run through each of the recognition systems to produce word or phone recognition hypotheses and lattices. Each set of word lattices is converted into a candidate term index, one for each system, with times and detection scores (posteriors) as shown in Figure 1. During the retrieval step, first the search terms are extracted from the system-specific candidate term lists, and then detection scores from each system are combined into one detection score via system combiner, as will be explained later in this section. We used the *lattice-tool* in SRILM toolkit to extract the list of all word N-grams during lattice indexing (up to $N = 3$ for word-only systems as this is the maximum length of keywords in our termlist). The term *posterior* for each N-gram is computed as the forward-backward combined score (acoustic, language, and prosodic scores were used) through all the lattice paths that share the N-gram nodes. We used a time tolerance of 0.5 seconds to merge the same N-grams with different times. Further details for indexing and search can be found in [18]. For the phonetic system, we employ the UTD Phone Confusion Network (PCN)-based keyword spotting system [19] after converting phone recognition lattices to phone confusion networks via SRILM toolkit.

2.3. Metadata extraction

To provide auxiliary information to the keyword spotting system, a channel identification system was developed specifically for the RATS channels. The objective of this system is to produce relevant information for an audio excerpt that reflects the property of the channel in what it was transmitted. We use i-vectors as features for Linear Gaussian classifier, which is trained to recognize one of the predefined acoustic conditions or channels. This system is based on the work done in [8]. In the context of RATS, the system extracts a vector of eight values, each corresponding to the likelihood of the audio file belonging to the respective channel. In this way, eight channels are used as bases to characterize a channel condition. The system was trained on data from the LDC corpus using standard MFCC features.

2.4. System Combination

When we merge different hits coming from different systems, for a specific target reference keyword location, as long as one of the systems finds it correctly, it helps to reduce the P(Miss), but the extra hits contribute to P(False-alarm). As mentioned earlier, system combination has been applied to identification tasks extensively, and good improvements are obtained. For keyword spotting, which is a detection task, the first thing that needs to be done is to align the detections provided by all the individual systems. This process is demonstrated in Figure 2. We consider 10 sec floating window, which corresponds to the tolerance required in the RATS project. For a speech recording and a particular position of the floating window, we consider only the best detection (with highest posterior score) from each sub-system. Such set of detections forms a candidate for the final combined detection. For each recording, all unique detection candidates corresponding to different floating window positions are collected. However, some of the window position are clearly suboptimal as they contain only subset of detections from different nearby position. Such window positions are not considered for forming detection candidates.

After this alignment step, for each recording, we have a collection of detection candidates, each corresponding to detections from one or more sub-systems. To obtain the final combined detections, each detection candidate has to get assigned time (e.g. average time from the participating detections) and posterior score. We explore two approaches for assigning the final scores: (1) *max-posterior filtering (MaxPost)*, and (2) *linear logistic-regression (LLR) score combination* with and without side information. In the first approach, detection candidate gets assigned score from the highest scoring sub-system (i.e., the highest score in the window). Although this approach reduces number of false alarms, it assumes that the posterior scores from the different sub-systems are comparable (i.e they are reasonably calibrated). This might be a problem when diverse sub-systems output keyword hits in different posterior score ranges as we will demonstrate with experimental results in the next section.

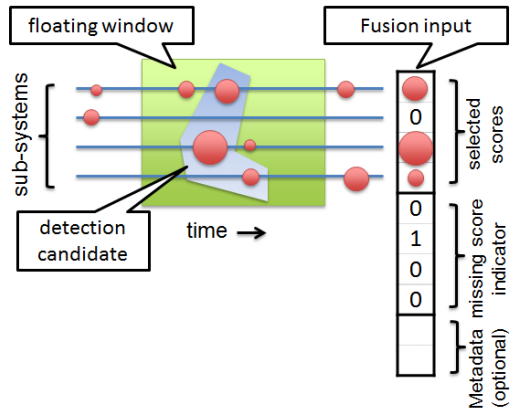

Fig. 2. System combination of different keyword spotting systems, including alignment and filtering step, as well as creation of feature vector for logistic-regression classifier.

In the second approach, for each detection candidate, we create a vector of the scores from all the participating detections as illustrated in figure 2. We convert posterior scores into log-likelihood ratio domain using logit function to make them better suitable for the following logistic regression based fuser [2]. We use zero vaue for missing scores (i.e. sub-systems with no detection in the corresponding window). The vector of scores is further augmented with vector of binary indicators of missing scores (values 0/1; one per each subsystem). Optionally, we augment the vector with metadata describing the detection (e.g., SNR at that time), acoustic characterization (e.g., channel ID) or keyword (e.g., number of phones), though we evaluated only the channel ID metadata in this work. The resulting vector $\mathbf{x}$ is used as an input to linear fusion, where the fused (log-likelihood ratio) score is calculated as $s = \mathbf{w}^T\mathbf{x}+b$. The fusion weights $\mathbf{w}$ and bias b are learned as a binary logistic regression classifier trained on positive and negative examples (i.e. correct hits and false alarms) from development data. It was shown in [20] that logistic regression optimizes performance of the combined system for a wide range of operating points (i.e. any point on the DET curve).

3. EXPERIMENTAL RESULTS

We evaluated the proposed approaches on the keyword spotting portion of RATS Phase-I evaluation data which is in Levantine Arabic. The test keyword set contains 200 keywords with at least three syllables. The system combination is trained on a development data with a much larger keyword set to generalize keyword models better. We ran all keyword spotting systems with 2000 keywords, and then used resulting detections to train system combination parameters.

Figure 3 shows the Detection Error Tradeoff (DET) curves for the 5 word-based systems and the phonetic system as well as *MaxPost* and *LLR* combined systems. As you can see in Figure 3, logistic-regression based combiner, listed as *LLR fusion*, achieves 5% relative reduction in P(Miss) at 4% P(False-Alarm) compared to max-posterior combiner which is listed as *MaxPost fusion*.

Since word-based systems have different posterior ranges than the phonetic system, this causes *MaxPost* approach combining uncalibrated scores. When we introduce a manual calibration step for max-posterior combiner by placing posterior scores from word-based systems into a different range (e.g., [1.0-2.0]) from that of phonetic system ([0-1.0]), we obtain a very similar performance compared to the logistic-regression combiner. This system is listed as *MaxPost fusion with manual calibration* in Figure 3. Yet the latter approach is a more principled way and it provides a framework where side information can be used. The LLR combiner does calibration internally, where at the same time MaxPost combiner requires an extra step of calibration. For word-based systems, calibration on this noisy test set was not very critical as the posterior distributions are similar. However, when we combine the phonetic system with the word-based systems, calibration becomes more critical since the posterior distributions are very different.

Next, we explore using channel identification as side information in LLR combiner. The channel identification system was run at the conversation level on both the training and the test data. Its output is an 8-dimensional feature vector that models the log-likelihood

[2]Here, "fusion" and "combiner" terms are used interchangeably.

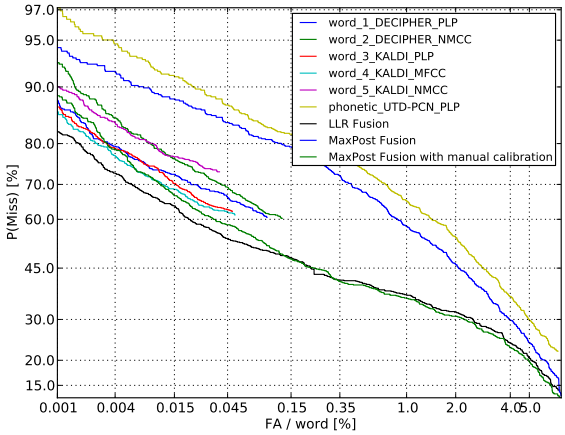

Fig. 3. DET curves for the five word-based systems (yielding more than 60% lowest-P(Miss)), and UTD phonetic system as well as the combination of these six systems with the proposed fusion schemes: linear logistic regression (LLR) fusion, max-posterior filtering (MaxPost) fusion, and MaxPost with manual calibration.

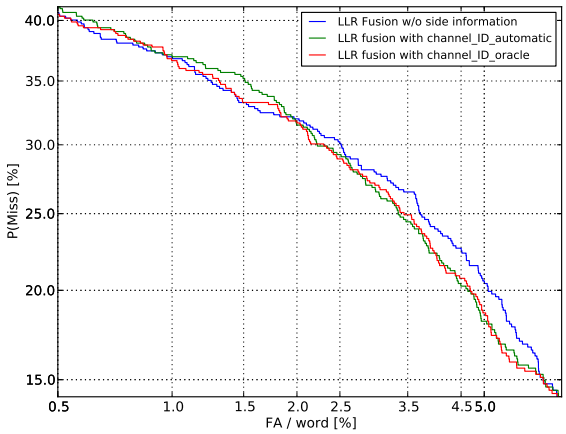

Fig. 4. DET curves for the fusion of the six systems using various fusion techniques: the proposed LLR fusion without side-information, LLR fusion using estimated as well as oracle channel information.

of the given conversation to originate from each of the eight RATS channels. As described in Section 2.4, for each detection, this 8-dimensional feature vector is appended to the 12-dimensional vector of detection scores and missing scores indicators. This 20 dimensional vector is used to train a logistic regression model. This system is listed as *LLR fusion with automatic channel ID* in Figure 4. A second LLR model is trained using an 8-dimensional feature vector representing oracle information about the true channel. This vector is set to have the value 1 along the dimension of the true channel and zero value along other dimensions. This system is listed as *LLR fusion with oracle channel ID*.

The results presented in Figure 4 show that using channel information as side info is very beneficial to the fusion of our six keyword spotting systems, especially below 30% P(Miss). In this range of the DET curve, LLR combination using *automatic* or *oracle* channel information, brings about 2% to 3% improvements in P(Miss) for a False-Alarm rate per word in the range of 3% to 6%. It is very encouraging that keyword spotting fusion with *automatic* channel side information performs as well as fusion using the *oracle* channel labels, since the latter is typically not available in practice.

4. DISCUSSION AND FUTURE WORK

In the current system, during recognition decoding we do not try to boost target keywords except for boosting their portion of the training data during language model training. A separate system that boosts keywords during decoding, similar to [21], can be added to the set of systems we use in parallel. We use channel identifiers provided by LDC as channel labels during channel ID training/testing. In the future, we would like to explore capturing acoustic conditions, not necessarily tied to channel labels, but a bigger set of conditions, similar to Environmental Sniffing work [6] where acoustic conditions are extracted in an unsupervised way. We plan to extend the acoustic side information by including features like signal-to-noise ratio (SNR). In addition to acoustic side information, we would like to extract other types of non-acoustic side information, such as exploring topic models, which will help to reduce potential mismatches on the language modeling side. To diversify system outputs, we also would like to employ acoustic condition specific, in the RATS scenario channel-specific, models as well during the recognition step in addition to multi-style-trained models.

5. CONCLUSION

We addressed the problem of retrieving spoken information from noisy and heterogeneous audio archives using rich system combination with a diverse set of robust modules and audio characterization. Our focus is to improve the accuracy of a keyword spotting system in a highly degraded and diverse set of channel conditions by employing multiple recognition systems with different robust frontends and modeling choices, as well as different lattice-based representations during audio indexing and search (words vs. subword units). At the end, we employ logistic-regression based system combination at the score level, after aligning keyword hits among different systems, and if available use side information such as the output of an acoustic condition identification module to guide the system combination module. We obtained significant improvements in P(Miss) at a fixed P(False-Alarm) by employing our proposed rich system combination approach on a dataset containing highly degraded and diverse channel characteristics.

6. ACKNOWLEDGMENTS

We thank Dimitra Vergyri, Arindam Mandal, and Jing Zheng of SRI International for providing the ASR system outputs developed under the DARPA RATS project. We thank Vikramjit Mitra for providing different frontend modules used in these ASR systems. We thank Abhijeet Sangwan for helping us to run the UT Dallas PCN-based keyword spotting system. We thank Aaron Lawson for providing the channel identification system that we use during side information extraction. We thank Luciana Ferrer for early discussions on system combination. This material is based on work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. D10PC20024. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of the DARPA or its Contracting Agent, the U.S. Department of the Interior, National Business Center, Acquisition & Property Management Division, Southwest Branch. The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

7. REFERENCES

[1] NIST, "The Spoken Term Detection STD 2006 Evaluation Plan", http://www.nist.gov/speech/tests/std/docs/std06-evalplan-v10.pdf, 2006.

[2] J. Garofolo, G. Auzanne, and E. Voorhees, "The TREC Spoken Document Retrieval Track: A Success Story," *Proc. of the Recherche d'Informations Assiste par Ordinateur: Content Based Multimedia Information Access Conference*, 2000.

[3] J.H.L. Hansen, B. Zhou, M. Akbacak, R. Sarikaya, B. Pellom, "Audio Stream Phrase Recognition for a National Gallery of the Spoken Word: One Small Step, *Proc. of ICSLP Conference*, 2000.

[4] J.H.L. Hansen, R. Huang, B. Zhou, M. Seadle, J.R. Deller, A.R. Gurijala, M. Kurimo, and P. Angkititrakul, "Speechfind: Advances in Spoken Document Retrieval for a National Gallery of the Spoken Word", *IEEE Transactions on Speech and Audio Processing*, Vol. 13(5), 2005.

[5] M. Franz, B. Ramabhadran, and M. Picheny, "Information Access in Large Spoken Archives", *Proc. of the ISCA Multilingual Spoken Document Retrieval Workshop*, 2003.

[6] M. Akbacak, J.H.L. Hansen, "Robust Spoken Document Retrieval in Acoustically Heterogeneous Historical Audio Archives", *submitted to Special Issue in Journal of Computer, Speech, and Language (CSL)*, August 2012.

[7] M. Akbacak, J.H.L. Hansen, "Environmental Sniffing: Noise Knowledge Estimation for Robust Speech Systems,"*IEEE Transactions on Speech & Audio Processing*, February 2007.

[8] L. Ferrer, L. Burget, O. Plchot, N. Scheffer, "A Unified Approach for Audio Characterization and its Application to Speaker Recognition", *Proc. of Odyssey Workshop*, 2012.

[9] M. Akbacak , D. Vergyri, A. Stolcke, N. Scheffer, and A. Mandal, "Effective Arabic Dialect Classification Using Diverse Phonotactic Models", *Proc. of Interspeech Conference*, 2011.

[10] A. Stolcke, M. Akbacak, L. Ferrer, S. Kajarekar, C. Richey, N. Scheffer, and E. Shriberg, "Improving language recognition with multilingual phone recognition and speaker adaptation transforms," *Proc. Odyssey Speaker and Language Recognition Workshop*, 2010.

[11] K.Walker, S. Strassel,"The rats radio traffic collection system," *Proc. of ISCA Odyssey Speaker and Language Recognition Workshop*, 2012.

[12] M. Maamouri, et al., "LDC2006S29, Arabic CTS Levantine QT training data set 5", *Linguistic Data Consortium*, 2006.

[13] S. F. Chen, J. T. Goodman, "An Empirical Study of Smoothing Techniques for Language Modeling", *Technical Report, Harvard University*, 1998.

[14] A. Mandal, M. Ostendorf, A. Stolcke, "Improving robustness of MLLR adaptation with speaker-clustered regression class trees", *Computer Speech and Language*, 23, 176 199, 2009.

[15] V. Mitra, H. Franco, M. Graciarena, A. Mandal, "Normalized Amplitude Modulation Features for Large Vocabulary Noise-Robust Speech Recognition", *Proc. of IEEE ICASSP Conference*, 2012.

[16] D. Povey, L. Burget et al., "The subspace Gaussian mixture model–A structured model for speech recognition," *Computer Speech, and Language*, 2011.

[17] D. Povey, A. Ghoshal, et al., "The Kaldi Speech Recognition Toolkit," *Proc. of IEEE ASRU Workshop*, 2011.

[18] D. Vergyri, I. Shafran, A. Stolcke, R. R. Gadde, M. Akbacak, B. Roark, and W. Wang, "The SRI/OGI 2006 Spoken Term Detection system," *Proc. of Interspeech Conference*, 2007.

[19] A. Sangwan, J. H. L. Hansen, "Keyword Recognition with Phone Confusion Networks and Phonological Features based Keyword Threshold Detection," *Proc. of ASILOMAR Conference*, 2010.

[20] N. Brummer, J. A. du Preez, "Application-independent evaluation of speaker detection", *Computer Speech and Language*, 20(2-3), 230-275, 2006.

[21] B. Zhang, R. Schwartz, S. Tsakalidis, L. Nguyen, S. Matsoukas, "White Listing and Score Normalization for Keyword Spotting of Noisy Speech", *Proc. of Interspeech Conference*, 2012.