Strategies for High Accuracy Keyword Detection in Noisy Channels

Arindam Mandal, Julien van Hout, Yik-Cheung Tam, Vikramjit Mitra, Yun Lei, Jing Zheng, Dimitra Vergyri, Luciana Ferrer, Martin Graciarena, Andreas Kathol, Horacio Franco

Speech Technology and Research Laboratory, SRI International, Menlo Park, California, USA

{arindam,julien,wilson,vmitra,yunlei,zj,dverg,lferrer,martin,kathol,hef}@speech.sri.com

Abstract

We present design strategies for a keyword spotting (KWS) system that operates in highly degraded channel conditions with very low signal-to-noise ratio levels. We employ a system combination approach by combining the outputs of multiple large vocabulary automatic speech recognition (LVCSR) systems, each of which employs a different system design approach targeting three different levels of information: front-end signal processing features (standard cepstra-based, noise-robust modulation and multi layer perceptron features), statistical acoustic models (gaussian mixtures models (GMM) and subspace GMMs) and keyword search strategies (word-based and phone-based). We also use keyword-aware capabilities in the system at two levels: in the LVCSR language models by assigning higher weights to n-grams with keywords in them and in LVCSR search by using a relaxed pruning threshold for keywords. The LVCSR system outputs are represented as lattice-based unigram indices whose scores are fused by a logistic-regression based classifier to produce the final system combination output. We present the performance of our system in the phase II evaluations of DARPA's Robust Automatic Transcription of Speech (RATS) program for both Levantine Arabic and Farsi conversational speech corpora.

Index Terms: noise-robust keyword detection, automatic speech recognition, system combination, noise robustness

1. Introduction

Modern KWS systems typically employ sequence modeling approaches that use hidden Markov models (HMM) to model keywords and all other words (the *garbage model*). HMM based approaches can be grouped into four categories: whole-word or acoustic KWS that model entire keywords and other words (*garbage words*) as HMMs[1]; phonetic KWS use HMMs to model phone-level (or triphone-level) representations of keywords and ergodic HMMs for garbage words[2], ASR-based KWS use standard HMM-based ASR to produce word-level lattices that are represented as indices for keywords search[3]; and hybrid KWS that combine phonetic and ASR-based KWS approaches to produce sub-word lattices for generating keyword search indices. A detailed survey of existing KWS techniques can be found in [4, 5].

In this work we focus on KWS in conversational speech that is distorted by the transmission channel with the resulting signal-to-noise ratios (SNR) ranging from 20 db to as low as 0 db collected under DARPA's RATS program. To achieve low false alarm rates at such SNR levels, we employ a system combination approach based on combining a diverse set of outputs of multiple ASR systems. Fig. 1 shows the architecture of our KWS system. Each large vocabulary ASR system applies a different system design strategy, targeting three different levels of information: signal processing features, where we use standard cepstra-based features such as mel frequency cepstra coefficients (MFCC) and perceptual linear prediction (PLP), and noise-robust features based on normalized modulation of cepstra and Gabor/Tandem posteriors; statistical acoustic modeling, where we use standard Gaussian mixture models (GMMs) and subspace GMMs; and KWS search space representations, where we use both word lattices and phone lattices produced by LVCSR systems. Each ASR system also uses keyword-aware capabilities to improve KWS performance by first assigning higher weight to n-grams containing keywords in the ASR decoding language models (LM) and second by using relaxed pruning thresholds for keywords during ASR search. The ASR systems produce word-lattices that are in turn converted to phone confusion networks (PCN). Keyword search is carried out for each system on word-lattices and PCNs and the final KWS output is obtained by applying a logistic-regression based fusion to a subset of these system's outputs. In this study, we present analysis of the KWS performance of the above mentioned system design alternatives that led to our final system configuration.

2. Corpora and Task

The speech corpora used in this study was originally collected under DARPA's RATS program by the Linguistic Data Consortium (LDC), which was focused on speech in noisy or heavily distorted channels in two languages: Levantine Arabic and Farsi. The type of noisy channels targeted by the RATS program is similar to the distortion characteristics of air traffic controller radio communication channels such as sideband mistuning, tonal interference and multi path interference. Pre-existing conversational speech corpora in these two languages were retransmitted through eight such channels [6]. For Levantine Arabic acoustic model (AM) training we use approximately 250 hours of retransmitted conversational speech (LDC2011E111 and LDC2011E93) and for LM training we use various sources: 1.3M words from LDC's EARS data collection (LDC2006S29, LDC2006T07); 437K words from Levantine Fisher (LDC2011E111 and LDC2011E93); 53K words from RATS data collection(LDC2011E111); 342K words from GALE Levantine broadcast shows (LDC2012E79) and 942K words from web data in dialectal Arabic (LDC2010E17). We set aside a LM tuning set selected from the Fisher data collection of about 46K words. To evaluate ASR and KWS performance for Levantine Arabic, we use two different test sets, each consisting of 10 hours of held-out conversational speech. These test sets are referred to as **alv dev-1** and **alv dev-2** in the rest of this paper. For Farsi AM training, we use approximately 339 hours of retransmitted conversational speech from three corpora (LDC2001E111, LDC2012E132 and LDC2013E03) and

for LM training we use 147K words and for LM tuning we use 5K words selected from these three corpora. To evaluate ASR and KWS performance for Farsi, we use a set of 10 hours of held-out conversational speech referred to as **fas dev** in the rest of this paper. A set of 200 keywords are pre-specified for each language and test set, where each keyword is composed of up to three words and at least three syllables long and appearing at least three times on average in the test set.

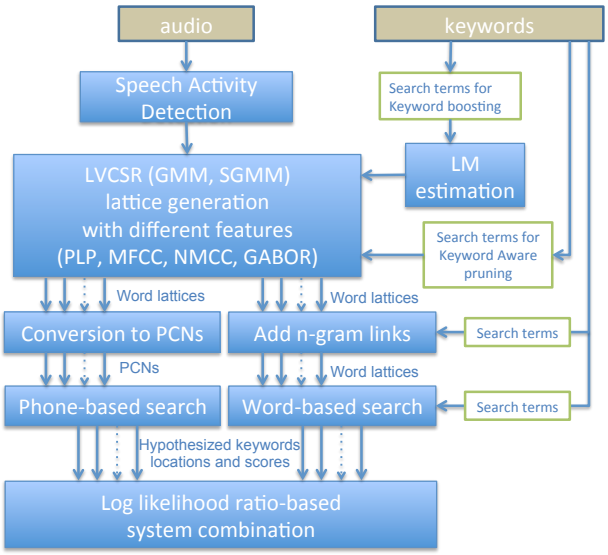


Figure 1: KWS system architecture

3. Automatic Speech Recognition

3.1. Speech Activity Detection

We developed a speech activity detector (SAD) to segment the speech signal and perform ASR on the detected speech. SAD was composed of two HMMs, one for speech one for non-speech, each configured as three state tied HMMs using 1024 dimension GMM. The GMM feature was a 51 dimensional composition of four features: PLP plus deltas and double deltas, a Gabor spectro-temporal feature followed by MLP postprocessing[7], a voicing and spectral flux composition feature[8] and a voicing feature from subband correlogram processing. A Viterbi pass was performed to estimate the speech segments.

3.2. Noise Robust Features

Apart from using standard ASR front-end feature representations: MFCC and PLP, we also explored ways to compensate for the severe channel-degradation in the speech signals used in this study. We developed three different noise-robust signal processing features: normalized modulation cepstral coefficients (NMCC), medium duration modulation cepstral (MDMC) and Gabor/Tandem posteriors. NMCC[9] is obtained from tracking in time domain the amplitude modulations of subband speech signals by using a Hamming window of 25.6 ms with a frame rate of 10 ms to generate 13 cepstral coefficients and uses up to triple delta coefficients to yield a 52-dimensional feature vector. MDMC is similar to the NMCC feature but has several advanced signal processing steps and uses a medium duration analysis window of 51.2 ms. The Gabor/Tandem posteriors [7] feature uses a mel-spectrogram convolved with spectro-temporal Gabor filters at different frequency channels. This feature uses a MLP to predict the monophone class posteriors of each frame given the Gabor values as of the present frame and the surrounding frames, which are appended standard 39-dimensional MFCC features.

3.3. Acoustic Modeling

We pool training data from all the eight noisy channels to train multi-channel acoustic models of two types, both of which use three-state left-to-right HMMs to model crossword triphones. The two types of models differ in the way they model the HMM state output probability, one uses standard GMMs and the other one uses subspace GMMs. The training corpus is clustered into pseudo speaker clusters using unsupervised agglomerative clustering. The front-end feature vector is normalized using standard cepstral mean and variance normalization and vocal tract length normalization (VTLN) over the pseudo speaker clusters, and the features are also transformed using heteroscedastic linear discriminant analysis (HLDA). For both GMM and SGMM AMs we train speaker-adaptive maximum likelihood (ML) models. The GMM models are speaker-adapted using maximum likelihood linear regression (MLLR) and the SGMM models are adapted using feature-space MLLR and speaker subspace adaptation. We perform cross-adaptation of systems by exchanging the MLLR reference hypothesis between SGMM and GMM systems each of which use a different front-end feature. For the GMM system we use SRI International's DECIPHER™ASR engine[10] and the the KALDI speech recognition toolkit[11] for training SGMM systems.

3.4. Language modeling

The LM vocabulary was selected using the approach described in [12]. Using our held-out tuning set for each language, this approach selected a vocabulary of 47K words for Levantine Arabic and 42K for Farsi, which resulted in an out of vocabulary (OOV) rate of 4.3% on alv dev-1 and 3.8% on fas dev. We added to this vocabulary the pre-specified keyword terms so that there were no OOV keywords during ASR search. Multi-term keywords were added as multi-words (treated as single words during recognition). The final LM was an interpolation of individual LMs trained on each of the corpora for each language as noted in Sec. 2. The interpolation weights were optimized to minimize perplexity on the LM tuning set. The individual LMs were retrained to *boost* the probability of keywords by repeating twice the sentences in the LM training corpora that contained any keywords and also by repeating the keyword terms twice. The same interpolation weights that were computed before boosting were used to interpolate the boosted language models, in order to avoid any additional bias towards a subset with more sentences containing keywords.

3.5. ASR Performance

We trained a separate ASR system for each combination of front end feature type and acoustic model type[1]. Table 1 shows the word error rates (WER) for the Levantine and Farsi ASR systems. Each system was cross adapted using adaptation hypotheses from other systems that use different modeling criterion and front-end features as described in Sec. 3.3.

[1]The Decipher GMM and KALDI SGMM systems are not directly comparable and the KALDI SGMM was better than KALDI GMM

Front End	alv dev-1		fas dev	
	GMM	SGMM	GMM	SGMM
MFCC	73.5	73.8	74.6	75.1
PLP	73.5	73.5	75.1	75.8
NMCC	75.5	n/a	75.0	75.1
MDMC	n/a	73.5	-	-
Gabor/Tandem	76.1	n/a	77.3	78.5

Table 1: WER (%) Decipher GMM & KALDI SGMM for Levantine & Farsi

4. Keyword Search

4.1. Keyword Aware Pruning

We developed a keyword-aware pruning scheme similar to [13] in our GMM ASR systems, in order to retain within ASR lattices rare or unintelligible keywords (in the presence of noise), which may be "lost" during ASR search pruning due to low LM and AM scores. For each frame, we sorted the partial hypotheses by their log likelihood and applied a different pruning beam depending on whether a hypothesis ended at a keyword state to decide if a hypothesis should be removed. We decided on a pruning beam for keywords that is twice as large as that for non-keywords, which provided a good trade-off between speed of ASR search and KWS performance. For the SGMM systems, we found that the standard approach to generate determinized state-level lattices did not have rich alternative hypotheses to ensure the span of decision error tradeoff (DET) curves to include low $p(miss)$[2] regions in KWS detection. We added a special pruning step, before determinization, that applied an aggressive pruning beam to non-keywords, but ensured survival of at least one path in the lattice for each existing keyword (including multiwords), which significantly improved the coverage of keywords in the lattices.

4.2. Keyword Search Index

ASR-based KWS search is performed in lattice structures since they contain rich alternative hypotheses. We use word lattices and phone confusion networks (PCNs) to generate a keyword search index. ASR word lattices of each system is used to create a candidate term index by listing all words in the lattice along with their start/end time and posterior scores. We used a time tolerance of 0.5 seconds to merge multiple occurrences of a word with different times. The KWS output of each system is obtained by taking the subset of words in the index that are keywords. Since n-gram keywords were added to the LM these are treated as single words in the lattices and therefore appear in the index. We added links in the word lattices where two or three consecutive nodes form a keyword. This allowed recovery of multiword keywords for which ASR search hypothesized the sequence of words forming the keyword instead of the keyword itself. We also converted the word lattices of each system to phone lattices and then into PCNs using SRILM. A less constrained search is carried out allowing phone deletions using a PCN-based KWS search[14].

4.3. System Combination & Selection

We used a logistic regression-based classifier for combining the KWS search indices of multiple systems as we reported in [15]. Given N systems, we created a vector of length $2N$ for a keyword hit, where the first N dimensions corresponded to each system's score converted from $[0, 1]$ to $[-\infty, +\infty]$ via the logit function and the second N dimensions were binary indicator variables to indicate if each *separate* system has a missing score. This vector x is used to train a binary classifier using logistic regression using positive and negative examples (i.e. correct hits and false alarms of keywords) on development data, which learns the fusion weights w and a bias b and produces the final system score $s = w^T x + b$. We also experimented with adding $N(N+1)/2$ indicator variables to x for indicating all possible *pairs* of N systems missing scores at the same time instant. Since our phonetic KWS search produces up to an order of magnitude more keyword hits than the word systems, we experimented with training a separate fusion model when the hits only came from phone KWS search. For selecting the best systems, we considered $2N$ systems equally split between word and phone systems for a given language. Then using a two step greedy search, we first selected W word systems leaving one out a time till we stop seeing KWS performance gains, then to these we add the best P phone systems.

[2] $p(miss)$ = probability of miss; $p(fa)$ = probability of false alarm

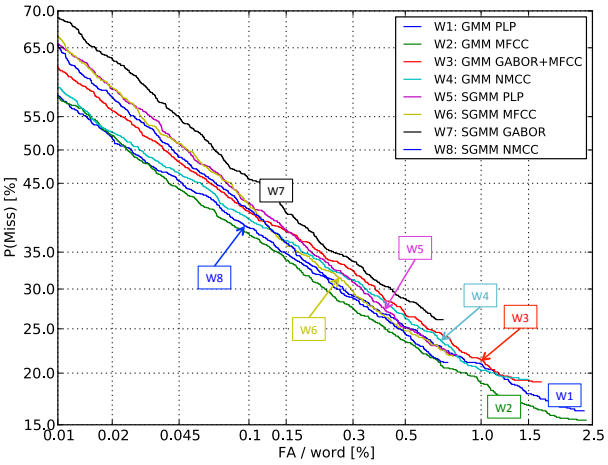


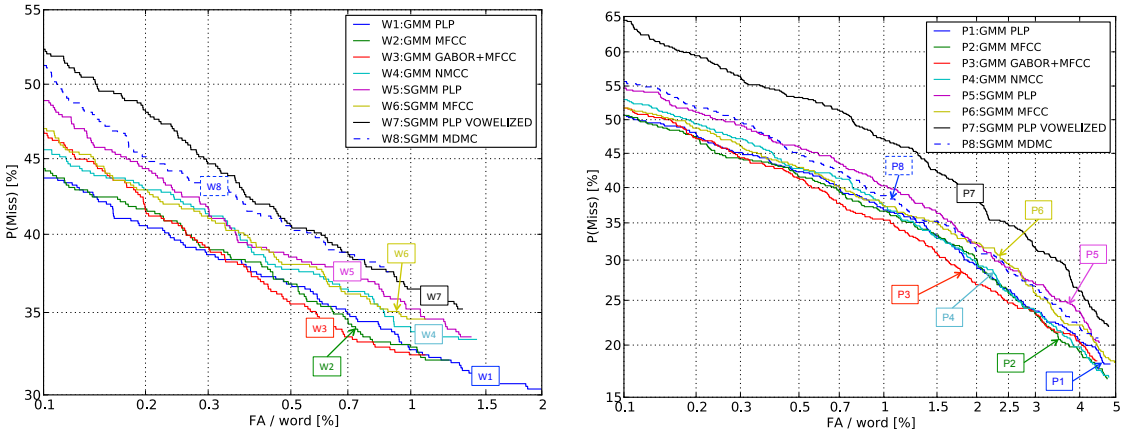
Figure 2: Farsi word-based systems.

5. Results & Discussion

In this part, we present our results in terms of the $p(miss)$ and $p(fa)$ KWS metrics used under the RATS program. Evaluation was performed on the **alv dev2** set for Levantine and on the **fas dev** set for Farsi with 200 keywords. The fusion was trained on separate audio with 1000 extra keywords that we selected for each language.

First, we found that our various keyword-aware strategies brought significant gains. Preliminary experiments on a GMM-PLP system showed that the two keyword search techniques described in Sec 4.2, either adding multiwords to the LM and creating a 1-gram index, or using a single-word LM but adding multiword links in the lattice, provide similar performance and a lowest operating point of 43% $p(miss)$ at 1% $p(fa)$. Combining the two approaches, by adding multiword links on top of lattices created with a multiword LM brought an extra 3.5% gain in $p(miss)$ at 1% $p(fa)$. Also, we found that keyword boosting in the LM brought gains of 1 to 2% in $p(miss)$ at 1% $p(fa)$. While keyword-aware pruning in ASR search brought no gains in $p(miss)$ at a given $p(fa)$, it allowed us to extend the range of operating points of our system by an additional 1% in $p(miss)$. This is because using a larger pruning beam for keywords allows more keywords with very low scores to be hypothesized, therefore lowering the lowest achievable miss rate.

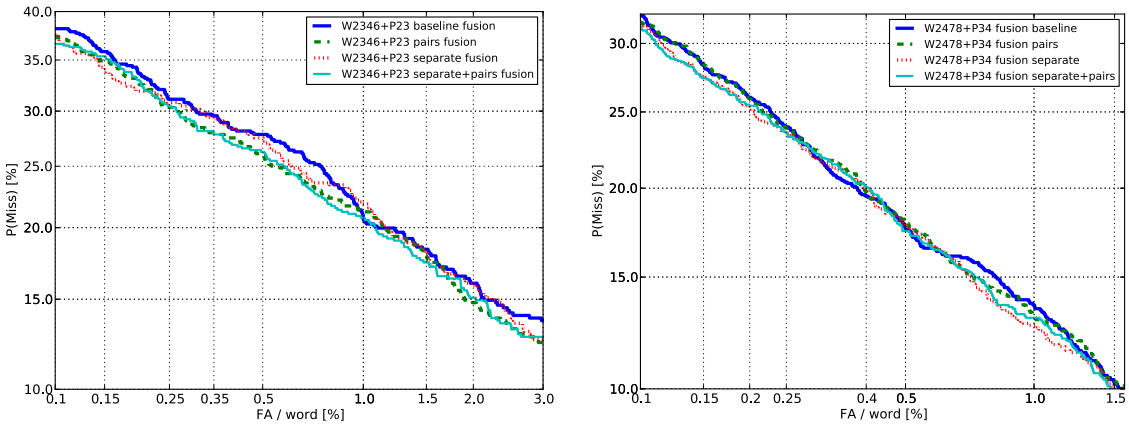
Second, we found our word-based systems and PCN-based



(a) Eight word-based systems

(b) Eight PCN-based systems

Figure 3: Levantine Arabic individual systems performance on the **alv dev2** set



(a) System combination performance

(b) System combination performance

Figure 4: Levantine and Farsi system combination performance on the **alv dev2** and **fas dev** sets

systems to have different characteristics. The word-based systems generally exhibit a lower $p(fa)$ at a given $p(miss)$, they are limited to a lowest $p(miss)$ of 30% to 35% for Levantine and of 15% to 25% for Farsi (see DET curve on Figures 2 and 3a). The more flexible KWS search performed on PCNs can achieve much lower $p(miss)$ rates at the cost of increased false-alarm rates (Figure 3b). For both languages and indexing approaches (word-based and PCN-based), we observed that the best performing systems were GMM-based, and used Gabor/Tandem posteriors, MFCC or PLP front-end features.

Third, we found system selection and fusion to bring very large gains compared to using individual systems. After applying the system selection procedure of Sec. 4.3, the best 6 systems for combination were found to be W2346+P23[3] for Levantine, and W2478+P34 for Farsi. For Levantine, the addition of the two PCN-based systems to the 4 selected word systems brought gains up to 2% in p(miss) for a fixed p(fa) and extends the span of the DET curve to a lowest achievable p(miss) of less than 5%. Fig. 4a shows that the novel *pairs* and the *separate* modeling approaches brought gains of the order of 2% in $p(miss)$ at $p(fa)$ below 1% over the baseline approach and their combination lowered $p(miss)$ by up to 1% in some parts of the DET curve. For Farsi, Figure 4b shows that the *separate* approach shows about 1% gain in $p(miss)$, while the *pairs* modeling does not show gains. Improvements from system combination can be observed by comparing the best Levantine word and phone systems in Fig. 3a & 3b (W3 and P3) and the 6-system fusion in Fig. 4a. At 35% $p(miss)$ the fused system achieves 0.15% $p(fa)$ while W3 and P3 achieve 0.5% and 1% $p(fa)$ respectively. At a $p(miss)$ of 20% the best phone system (P2) achieves 3.5% $p(fa)$ while the combined system achieves 1% $p(fa)$.

Our future work will focus on improved noise-robust features, novel subword units for ASR and improvements to phonetic KWS to reduce $p(fa)$.

Acknowledgements This material is based upon work supported by 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. Approved for Public Release, Distribution Unlimited.

[3]The system ids start with W for word-based systems and P for phone-based systems

6. References

[1] J. R. Rohlicek, W. Russell, S. Roukos, and H. Gish, "Continuous hidden markov modeling for speaker-independent word spotting," in *International Conference on Acoustics, Speech, and Signal Processing, 1989*. IEEE, 1989, pp. 627–630.

[2] H. Bourlard, B. D'hoore, and J.-M. Boite, "Optimizing recognition and rejection performance in wordspotting systems," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 1994*, vol. 1. IEEE, 1994, pp. I–373.

[3] M. Weintraub, "LVCSR log-likelihood ratio scoring for keyword spotting," in *International Conference on Acoustics, Speech, and Signal Processing, 1995*, vol. 1. IEEE, 1995, pp. 297–300.

[4] J. Keshet, D. Grangier, and S. Bengio, "Discriminative keyword spotting," *Speech Communication*, vol. 51, no. 4, pp. 317–329, 2009.

[5] I. Szoke, P. Schwarz, P. Matejka, L. Burget, M. Karafiát, M. Fapso, and J. Cernocky, "Comparison of keyword spotting approaches for informal continuous speech," in *Proc. of Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, 2005.

[6] K. Walker and S. Strassel, "The RATS radio traffic collection system," in *Odyssey 2012-The Speaker and Language Recognition Workshop*, 2012.

[7] B. T. Meyer, S. V. Ravuri, M. R. Schädler, and N. Morgan, "Comparing different flavors of spectro-temporal features for ASR," in *Proc. of Interspeech*, 2011, pp. 1269–1272.

[8] S. Sadjadi and J. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *Signal Processing Letters, IEEE*, vol. 20, no. 3, pp. 197–200, 2013.

[9] V. Mitra, H. Franco, M. Graciarena, and A. Mandal, "Normalized amplitude modulation features for large vocabulary noise-robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4117–4120.

[10] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. Gadde, M. Plauché, C. Richey, E. Shriberg, K. Sönmez, F. Weng *et al.*, "The SRI March 2000 Hub-5 conversational speech transcription system," in *Proc. NIST Speech Transcription Workshop*, 2000.

[11] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlıcek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *Proc. ASRU*, 2011.

[12] A. Venkataraman and W. Wang, "Techniques for effective vocabulary selection," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[13] B. Zhang, R. Schwartz, S. Tsakalidis, L. Nguyen, and S. Matsoukas, "White listing and score normalization for keyword spotting of noisy speech," in *Proceedings of Interspeech 2012*, 2012.

[14] A. Sangwan and J. H. Hansen, "Keyword recognition with phone confusion networks and phonological features based keyword threshold detection," in *Signals, Systems and Computers (ASILOMAR), 2010 Conference Record of the Forty Fourth Asilomar Conference on*. IEEE, 2010, pp. 711–715.

[15] M. Akbacak, L. Burget, W. Wang, and J. van Hout, "Rich system combination for keyword spotting in noisy and acoustically heterogeneous audio streams," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013.