# 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 a rich system combination with a diverse set of noise-robust modules and audio characterization. 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 spotting system in a 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). Then, after aligning keyword hits from different systems, we employ system combination at the score level using a logistic-regression-based classifier. When available, side information (such as signal-to-noise ratio or the output of an acoustic condition identification module) is used to guide system combination that is trained on separate held-out data. 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 Audio Transcription (RATS) Phase-I evaluation data that contains highly degraded channel recordings (SNR as low as 0 dB) and different channel characteristics.

***Index Terms*—** Keyword spotting, channel degradation, 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 [1], spoken term detection [2], and spoken document retrieval [3]. 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.

The effect of acoustic condition mismatch and variety on spoken document retrieval performance has been heavily observed in audio search applications for digital archive projects such as [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, we employ diverse 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 side information (such as signal-to-noise ratio or output of an acoustic condition identification module).

In Section 2, we provide an overview of the system architecture. In Section 3, details for the system components are presented. Evaluation of the proposed system combination for keyword spotting is presented in Section 4. Discussion and future work are presented in Section 5, with conclusions in Section 6.

Murat Akbacak and Lukas Burget were at SRI International when this work was done.

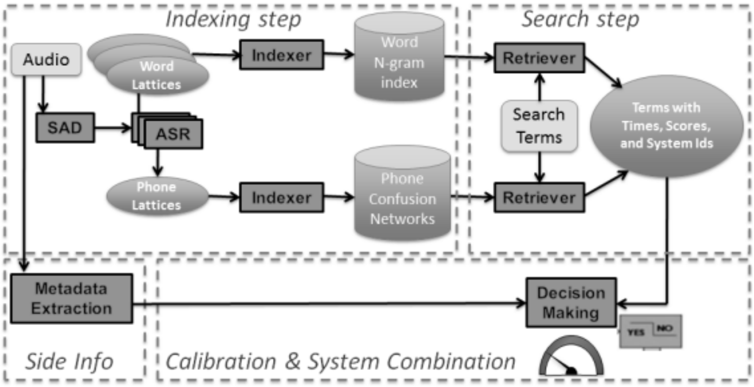

**Fig. 1**. System overview of SCENIC keyword spotting system.

## 2. SYSTEM OVERVIEW

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. During the frontend feature extraction step, we employ conventional features like MFCC, as well as noise-robust features such as PLP and PNCC features [13]. During modeling, we employ conventional diagonal GMM-based acoustic models as well as subspace GMM [15]-based acoustic models which are more robust to channel degradations, under Decipher and Kaldi [14] recognition systems, respectively. We also employ acoustic condition specific, in this case channel-specific, models as well during the recognition step to diversify our system outputs with the expectation that channel-specific models might perform better than multi-style trained models. During the indexing and search step, we employ word units as well as subword units (e.g., phones) to increase the robustness of the system. For all keyword spotting systems, indexing and search is done at the lattice level. All these system outputs go into a system combiner. During system combination, side information is extracted from the audio file to represent acoustic characteristics, and guide effective system combination and calibration. We consider two types of side information: (1) keyword specific side information, and (2) acoustic condition side information. The first of these includes features such as language model score of the keyword, keyword length, in terms of number of phones as well as number of words in the keyword N-gram. The second includes features like signal-to-noise ratio (SNR), and channel identification score vector. The following section provides more details on these system components.

## 3. SYSTEM COMPONENTS

Here, we present the choices made for different components of the keyword spotting system.

### 3.1. 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 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 are trained with Kneser-Ney smoothing.

#### *3.1.1. Robust Front-end Features*

In addition to conventional front-end features, such as MFCC, we employ noise-robust features such as PLP, NMCC, PNCC, and CSAWH. Further details of these front-end features can be found in [13].

#### *3.1.2. Robust Acoustic Models*

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 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 it allows a large model size for small training data. We use the KALDI speech recognition toolkit [14] for SGMMs. For the GMM-based system, we use SRI's Decipher engine. We explore both multi-style training where all channels are pulled into one model, and the channel-specific acoustic models employed in parallel for keyword spotting.

### 3.2. Lattice Indexing and Search

During indexing, audio input is run through the recognition system that produces word or phone recognition hypotheses and lattices. These are converted into a candidate term index with times and detection scores (posteriors). During the retrieval step, first the search terms are extracted from the candidate term list, and then a decision function is applied to accept or reject a candidate based on its detection score. Since the lattice structure provides additional information about the correct hypothesis, to avoid misses (which are more likely to occur in noisy recordings such as RATS data) several studies have used the whole hypothesized word or phone lattices to obtain the searchable index. We used the *lattice-tool* in SRILM toolkit to extract the list of all word N-grams (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 0.5 second time tolerance to merge the same N-grams with different times. Further details can be found in [16]. For the phonetic system, we employ the UTD Phone Confusion Network (PCN)-based keyword spotting system [17].

### 3.3. 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. This system is based on the work in [8] where a universal audio characterization system was developed to characterize audio characteristics of an audio file using the i-vector framework. 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.

### 3.4. System Combiner

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 hits in the system outputs. Figure 3 shows how this is done. After this alignment step, we explore two approaches: (1) max-posterior filtering, (2) logistic-regression score combination with and without side information. Different configurations of detections in the floating window forms detection candidates. We ignore clearly suboptimal candidates (e.g., subset of detections from other configurations). In the first approach, only keyword hit with the maximum posterior score is kept, and all others are discarded during system combination. Although this approach reduces number of false alarms, it does not perform calibration on the scores. In the second approach, we create a vector of best scores (logit posteriors) from each system. We use zero score for systems with no detection in the window and set corresponding indicators of missing scores to 1. 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). We use the final vectors to train a binary linear logistic regression classifier as a fuser.

## 4. EXPERIMENTAL RESULTS

We evaluated the proposed approaches on the keyword spotting portion of RATS Phase-I evaluation data which is in Levantine Arabic. The keyword set contains 200 keywords with at least three syllables. As an evaluation metric specified under the RATS program, we look at P(Miss) at a specific P(False-Alarm) (which is set at 4%), and also provide ROC curves. 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 hits, negative samples (false alarms), and positive samples (correct hits), to train system combination parameters. Figure 3 shows the results for individual systems as well as the combined system without side information.

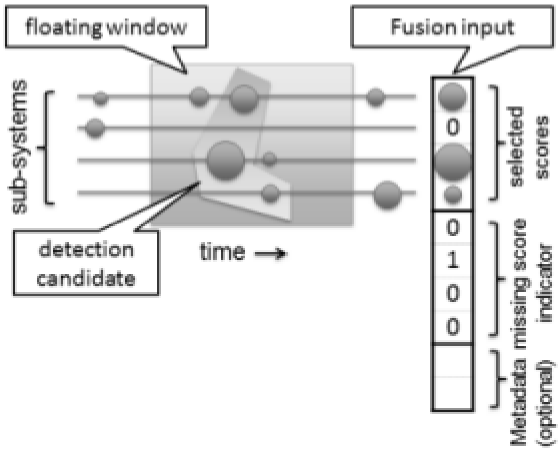

**Fig. 2**. System combination of different keyword spotting systems.

The channel identification system was run at the conversation level on both the training and the test data. Its output is an 8-dimensionnal feature vector that models the likelihood of the given conversation to originate from all of the eight RATS channels. This 8-dimensional feature vector is then appended to the 12-dimensionnal feature vector characterizing each triplet of segment, keyword and time. This total vector size of 20 is used to train a logistic regression model. A second logistic regression model is trained using an 8-dimensionnal 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.

The results presented in Figure 4 show that using knowledge of channel information is very beneficial to the fusion of our six keyword spotting systems, especially to achieve miss rates below 30%. In this range of the DET curve, the two approaches we developed to use channel information bring about 2% to 3% improvements in P(Miss) for a False-Alarm rate per word in the range from 3% to 6%. It is very encouraging that keyword spotting score fusion using estimated channel information performs as well as fusion using the true channel labels since the latter is not available in practice.

An interesting observation is that maximum-posterior-approach yields reasonable performance compared to the logistic-regression based system combiner, yet the latter approach is a more principled way and provides a solution where side information can be used. The logistic-regression-based system combiner does calibration internally, where at the same time maximum-posterior-based system combination requires an extra step of calibration. For word-based systems, calibration on this noisy testset 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 as the posterior ranges and the number of hits are very different between these two types (word vs. phone) of system.

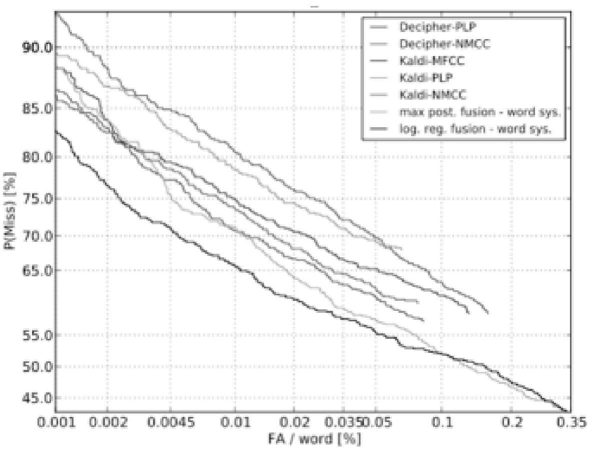

**Fig. 3**. ROC curves for word-based systems, as well as, combination of word-based systems.

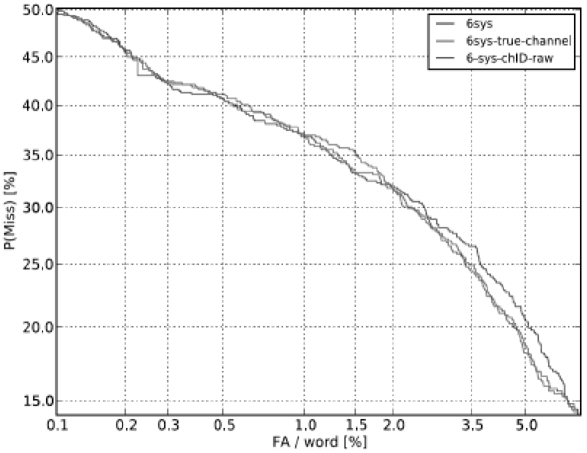

**Fig. 4**. ROC curves for the combination of 5 word-based systems and 1 phonetic system, with and without side information.

## 5. 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 [18], 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. In addition to acoustic side information, we would like to extract other types of nonacoustic side information, such as exploring topic models, which will help to reduce potential mismatches on the language modeling side.

## 6. CONCLUSION

We address 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 representations during audio indexing and search (words vs. subword units). At the end, we employ system combination at the score level, after aligning keyword hits among different systems, and if available use side information (such as signal-to-noise ratio or output of an acoustic condition identification module) to guide a system combination module that is trained on separate held-out data. Lattice-based indexing and search is used in all keyword spotting systems. We present improvements in P(Miss) at a fixed P(False-Alarm) by employing our proposed rich system combination approach on DARPA Robust Audio Transcription (RATS) Phase-I evaluation data containing highly degraded channel characteristics (SNR levels as low as 0 dB) and consisting different channel characteristics.

## 7. 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 discussion 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.

## 8. REFERENCES

[1] 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.

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

[3] 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.

[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] 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.

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

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

[16] 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.

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

[18] 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.