INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Fusion Strategies for Robust Speech Recognition and Keyword Spotting for Channel- and Noise-Degraded Speech

Vikramjit Mitra[1], *Julien VanHout*[1], *Wen Wang*[1], *Chris Bartels*[1], *Horacio Franco*[1], *Dimitra Vergyri*[1], *Abeer Alwan*[2], *Adam Janin*[3], *John Hansen*[5], *Richard Stern*[4], *Abhijeet Sangwan*[5], *Nelson Morgan*[3]

[1]Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, USA
[2]Speech Processing and Auditory Perception Lab., Univ. of California, Los Angeles, CA, USA
[3]International Computer Science Institute (ICSI), Berkeley, CA, USA
[4]Dept. of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburg, PA, USA
[5]Center for Robust Speech Systems (CRSS), U.T. Dallas, Richardson, TX, USA

vikramjit.mitra@sri.com

Abstract

Recognizing speech under high levels of channel and/or noise degradation is challenging. Current state-of-the-art automatic speech recognition systems are sensitive to changing acoustic conditions, which can cause significant performance degradation. Noise-robust acoustic features can improve speech recognition performance under varying background conditions, where it is usually observed that robust modeling techniques and multiple system fusion can help to improve the performance even further. This work investigates a wide array of robust acoustic features that have been previously used to successfully improve speech recognition robustness. We use these features to train individual acoustic models, and we analyze their individual performance. We investigate and report results for simple feature combination, feature-map combination at the output of convolutional layers, and fusion of deep neural nets at the senone posterior level. We report results for speech recognition on a large-vocabulary, noise- and channel-degraded Levantine Arabic speech corpus distributed through the Defense Advance Research Projects Agency (DARPA) Robust Automatic Speech Transcription (RATS) program. In addition, we report keyword spotting results to demonstrate the effect of robust features and multiple levels of information fusion.

Index Terms: noise-robust speech recognition, large-vocabulary continuous speech recognition, time frequency convolution, feature-map fusion.

1. Introduction

Current large-vocabulary continuous speech recognition (LVCSR) systems demonstrate high levels of recognition accuracy under clean condition or at high signal-to-noise ratios (SNRs). However, these systems are very sensitive to environmental degradations such as background noise, channel mismatch, or distortion. Hence, enhancing robustness against noise and channel degradations has become an important research area for automatic speech recognition (ASR), keyword spotting (KWS), speech-activity detection, speaker identification, etc.

Traditionally, ASR systems use mel-frequency cepstral coefficients (MFCCs) as the acoustic observation. However, MFCCs are susceptible to noise; as a consequence, their performance can degrade dramatically with increases in noise levels and channel degradations. To counter this vulnerability, researchers have actively explored alternative feature sets that are robust to background degradations. Research on noise-robust acoustic features typically aims to generate a relatively invariant speech representation, such that background distortions minimally impact the features. To reduce the effect of noise, speech-enhancement-based approaches have been explored, where the noisy speech signal is enhanced by reducing noise corruption (e.g., spectral subtraction [1], computational auditory scene analysis [2], etc.). Noise-robust signal-processing approaches have also been explored, where noise-robust transforms and/or human-perception-based speech-analysis methodologies were investigated for acoustic-feature generation, such as the ETSI [European Telecommunications Standards Institute] advanced [3] front-end, power-normalized cepstral coefficients [PNCC] [4], modulation-based features [5-7], and many more.

Recently, the introduction of deep learning techniques [8] enabled significant improvement in speech recognition performance [9]. In this vein, convolutional deep neural networks (CNNs) [10, 11] have proved to often perform better than fully connected deep neural networks (DNNs) [12, 13], specifically for features having spatial correlations across their dimensions. CNNs are also expected to be noise robust [11], especially when the noise or distortion is localized in the spectrum. Speaker-normalization techniques, such as vocal tract length normalization (VTLN) [14], have been found to have less impact on speech recognition accuracy for CNNs than for DNNs. With CNNs, the localized convolution filters across frequency tend to normalize the spectral variations in speech arising from vocal tract length differences, enabling CNNs to learn speaker-invariant data representations. Recent results [12-13] have also shown that CNNs are more robust to noise and channel degradations than DNNs. Typically for speech recognition, a single layer of convolution filters is used on the input contextualized feature space to create multiple feature maps that, in turn, are fed to fully connected DNNs.

In [15, 16], convolution across time was applied over windows of acoustic frames that overlap in time to learn classes such as phone, speaker, and gender. In 1980s, the notion of weight sharing over time was first introduced through the time-delay neural network (TDNN) [17]. Recent DNN/CNN architectures use a hybrid topology, in which

http://dx.doi.org/10.21437/Interspeech.2016-279

DNN/CNNs produce subword unit posteriors, and a hidden Markov model (HMM) performs the final decoding. As the HMMs typically model time variations well, time convolution is usually ignored in current CNN architectures. However, with environmental degradation such as reverberation, the introduced distortion typically corrupts time-scale information. Recent work [18] has shown that employing temporal convolution along with spatial (frequency-scale) convolution using a time-frequency CNN (TFCNN) can add to the robustness of speech recognition systems.

Here, we present speech recognition and keyword spotting (KWS) results for acoustic models trained on highly channel-degraded speech data collected as part of the U.S. Defense Advanced Research Agency's (DARPA's) Robust Automatic Transcription of Speech (RATS) program. State-of-the-art KWS systems [19-21, 29] mostly focus on training multiple KWS systems and then fusing their outputs to generate a highly accurate final KWS result. Fusion of multiple systems is usually observed to provide better KWS performance than that of the individual systems. Studies [22] have explored feature-combination approaches and demonstrated that they can improve KWS system accuracy. In this work, we show gains from feature-level combination, feature-map-level fusion, and system-level fusion, and we investigate if such techniques result in better KWS performance. This paper focuses on a Levantine Arabic (LAR) KWS task.

The DARPA RATS program aims to develop robust speech-processing techniques for highly degraded speech signals with emphasis on four broad tasks: (1) speech activity detection (SAD); (2) language identification (LID); (3) key word spotting (KWS); and (4) Speaker Identification (SID). The data was collected by the Linguistic Data Consortium (LDC) by retransmitting conversational telephone speech through eight different communication channels [23]. Note that RATS rebroadcasted data is unique in the sense that the noise and channel degradations were not artificially introduced by performing simple mathematical operations on the speech signal, but rather by transmitting clean source signals through different radio channels [23], where variations among the different channels introduced an array of distortion modes. The data also contained distortions, such as frequency shifting, speech modulated noise, non-linear artifacts, no transmission bursts, etc., which made robust signal-processing approaches even more challenging compared to those for the traditional noisy corpora available in the literature.

2. Dataset and Task

The speech dataset used in our experiments was collected by the Linguistic Data Consortium (LDC) under DARPA's RATS program, which focused on speech in noisy or heavily distorted channels in two languages: LAR and Farsi. The data was collected by retransmitting telephone speech through eight communication channels [23], each of which had a range of associated distortions. The DARPA RATS dataset is unique in that noise and channel degradations were not artificially introduced by performing mathematical operations on the clean speech signal; instead, the signals were rebroadcast through a channel- and noise-degraded ambience and then rerecorded. Consequently, the data contained several unusual artifacts, such as nonlinearity, frequency shifts, modulated noise, and intermittent bursts—conditions under which traditional noise-robust approaches developed in the context of additive noise may not have worked so well.

For LAR acoustic model (AM) training, we used approximately 250 hours of retransmitted conversational speech (LDC2011E111 and LDC2011E93); for language model (LM) training we used various sources: 1.3M words from the LDC's EARS (Effective, Affordable, Reusable Speech-to-Text) data collection (LDC2006S29, LDC2006T07); 437K words from Levantine Fisher (LDC2011E111 and LDC2011E93); 53K words from the RATS data collection (LDC2011E111); 342K words from the GALE (Global Autonomous Language Exploitation) Levantine broadcast shows (LDC2012E79); and 942K words from web data in dialectal Arabic (LDC2010E17). We used a held out set for LM tuning, which was selected from the Fisher data collection containing about 46K words. To evaluate KWS performance for LAR, we used two test sets—referred to as dev-1 and dev-2. Each set consisted of 10 hrs of held-out conversational speech. A set of 200 keywords was pre-specified for the LAR test set, with each keyword composed of up to three words and at least three syllables, and appearing at least three times on average in the test set.

2.1. Acoustic features

We used several different acoustic features to parameterize speech. We briefly outline the features explored in this section.

2.1.1 Damped Oscillator Coefficients (DOC)

DOCs use forced damped oscillators to model the hair cells found within the human ear [24]. DOC tracks the dynamics of the hair cell oscillations to auditory stimuli and uses that as the acoustic feature. In the human auditory system, the hair cells detect the motion of incoming sound waves and excite the neurons of the auditory nerves, which then transduce the relevant information to the brain. For our DOC processing, a bank of gammatone filters that produces 40 bandlimited subband signals analyzed the incoming speech signal. The gammatone filters were equally spaced on the equivalent rectangular bandwidth (ERB) scale. The outputs of the gammatone filters were used as the forcing functions to an array of 40 damped oscillators, whose response was then used as the acoustic feature. We analyzed the damped oscillator response by using a Hamming window of 26 ms with a frame rate of 10 ms. The power signal from the damped oscillator response was computed and then root compressed by using the 15th root, resulting in the 40 dimensional features that comprised the DOC feature in our experiments.

2.1.2 Normalized Modulation Coefficients (NMC)

The NMC [6] feature captures and uses the amplitude modulation (AM) information from bandlimited speech signals. NMC is motivated by AMs of subband speech signals playing an important role in human speech perception and recognition. NMCs are obtained by using the approach outlined in [6], with which the features are generated from tracking the AM trajectories of subband speech signals in a time domain by using a Hamming window of 26 ms with a frame rate of 10 ms. For our processing, a time-domain gammatone filterbank with 40 channels equally spaced on the ERB scale was used to analyze the speech signal. We then processed the subband signals by using a modified version of the Discrete Energy Separation algorithm (DESA) that produced instantaneous estimates of AM signals. The powers of the AM signals were then root compressed by using the 15th root. The resulting 40-dimensional feature vector was used as the NMC feature in our experiments.

2.1.3 Gammatone Filtebank energies (GFBs)

Gammatone filters are a linear approximation of the auditory filterbank found in the human ear. For the GFBs, the power of the bandlimited time signals within an analysis window of 26 ms was computed at a frame rate of 10 ms. The subband powers from 40 filters were root compressed using 15th root.

2.1.4 Log-Spectrally Enhanced Power Normalized Cepstral Coefficients (LSEN-PNCC)

The LSEN feature was initially introduced in [25] for enhancement of the mel-spectrum with applications to noise-robust speech recognition. It was adapted in [26] to enhance the gammatone power-normalized spectra of noisy speech obtained with the power-normalized cepstral coefficients (PNCC) features pipeline and renamed LSEN-PNCC.

PNCC [4, 27] was developed with the goal of providing a computationally efficient representation of speech that attempts to emulate at least in crude form a number of physiological phenomena that are potentially relevant for speech processing. PNCC processing includes (1) traditional pre-emphasis and short-time Fourier transformation (STFT); (2) integration of the squared energy of the STFT outputs by using gammatone frequency weighting; (3) “medium-time” nonlinear processing in each channel that suppresses the effects of additive noise and room reverberation; (4) a power-function nonlinearity with exponent 1/15; and (5) generation of cepstral-like coefficients by using a discrete cosine transform (DCT) and mean normalization.

2.1.5 Gabor-MFCC

The Gabor/Tandem posterior [28] features use a mel-spectrogram convolved with spectro-temporal Gabor filters at different frequency channels. Here, we used a multi-layer perceptron (MLP) to predict monophone class posteriors of each frame, which were then Kurhunen-Loeve transformed to 22 dimensions and appended with standard 39-dimensional MFCCs to yield 64 dimensional features.

In addition to these methods, we used standard mel-filterbank (MFBs) as the baseline feature set.

3. Acoustic Modeling

We pooled training data (multi-condition) from all eight noisy channels to train multi-channel acoustic models that used three-state, left-to-right HMMs to model crossword triphones. The training corpus was clustered into pseudo-speaker clusters by using unsupervised agglomerative clustering.

We have trained DNN-, CNN-, and TFCNN-based acoustic models in our experiments. To generate the training-data alignments necessary for training our acoustic models, we first trained a GMM-HMM model by using multi-condition training to produce senone labels. Altogether, the GMM-HMM system produced ~5K context-dependent (CD) states. The input layer of the CNN and DNN systems was formed by using a context window of 15 frames (7 frames on either side of the current frame); for TFCNNs, the context window was 17 frames.

For the CNN acoustic models, 200 convolutional filters of size 8 were used in the convolutional layer, and the pooling size was set to 3 without overlap. The subsequent fully connected network had five hidden layers, with 2048 nodes per hidden layer, and the output layer included as many nodes as the number of CD states for the given dataset. The networks were trained by using an initial four iterations with a constant learning rate of 0.008, followed by learning-rate halving based on cross-validation error decrease. Training stopped when no further significant reduction in cross-validation error was noted or when cross-validation error started to increase. Backpropagation was performed by using stochastic gradient descent with a mini-batch of 256 training examples. For the DNN systems, we used six layers with 2048 neurons in each layer, with similar learning criteria as the CNNs. The TFCNN architecture was based on [18], where two parallel convolutional layers are used at the input, one performing convolution across time, and the other across the frequency scale of the input filterbank features. That work showed that the TFCNNs gave better performance compared to their CNN counterparts. Here, we used 75 filters to perform time convolution, and 200 filters to perform frequency convolution. For time and frequency convolution, eight bands were used. A max-pooling over three samples was used for frequency convolution, and a max-pooling over five samples was used for time convolution. The feature maps after both the convolution operations were concatenated and then fed to a fully connected neural net, which had 2048 nodes and five hidden layers.

In this work, we present three types of information fusion in deep neural network architecture.

(1) Feature-level fusion: a pair of acoustic feature was concatenated with each other and was used to train on a single network.

(2) Feature-map-level fusion: two separate convolution layers were trained for two different feature sets and the ensuing feature maps were combined before training a fully connected network (fCNN), as shown in Figure 1.

(3) Decision-level fusion: two CNNs were jointly trained sharing their output layer (pCNN), as shown in Figure 2.

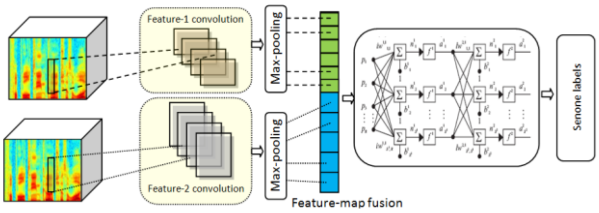

Figure 1. CNN feature-map-level fusion (fCNN).

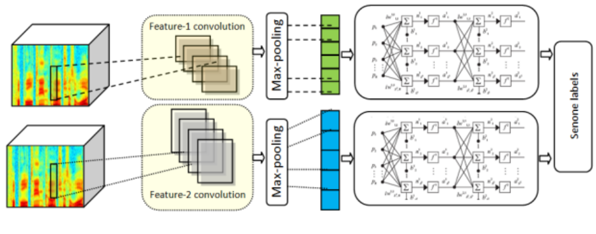

Figure 2. CNN decision-level fusion (pCNN).

4. Results

We trained the different DNN, CNN, and TFCNN acoustic models by using all the features described in section 2. We report speech recognition performance in terms of word error rates (WERs). First, we compared performance using the DNN system. Table 1 shows the WERs for the different features for clean, channel C and G from dev-1 individually and an averaged WER from dev-1 all channels. Note that our experimental observations showed that channels G and C were the best and the worst out of the eight channels with respect to ASR performance; hence, we have selected these two channels to present our results.

Table 1. WER from DNN models trained with different features, for dev-1 clean, channels C and G data and dev-1 all.

Features	Clean	C	G	dev-1
MFB	45.9	75.7	51.1	62.8
GFB	45.8	76.0	51.5	63.3
Gabor-MFCC	48.9	79.0	54.4	67.2
LSEN-PNCC	45.4	76.3	51.0	63.1
NMC	45.9	76.0	51.6	63.1
DOC	**45.3**	**75.6**	**50.5**	**62.3**

Table 1 demonstrates that the DOC robust feature helped to reduce the WER compared to the MFB and GFB baseline. Next, we investigated CNNs and TFCNNs, and report the results in Table 2. In table 2 we present the result for DOC which is one of the better-performing features; for the other features, the trend was similar as obtained from the DOCs. Note that the Gabor-MFCC features have their individual dimensions uncorrelated with respect to each other, and, as a consequence, using a convolution layer did not result in performance improvement for these features.

Next, we explored feature combination, feature-map fusion (fCNN), and output layer fusion (pCNN) for DOC, LSEN-PNCC, and NMC. We report the results in Table 3.

Table 2. WER from DNN, CNN, and TFCNN models trained with DOC features, for dev-1 clean channels C and G data and dev-1 all.

	Clean	C	G	dev-1
DNN	45.3	75.6	50.5	62.3
CNN	44.4	74.3	48.9	60.4
TFCNN	44.1	74.2	48.7	**60.2**

Table 3. WER from different fusion experiments using DOC, LSEN-PNCC, and NMC features, for dev-1 clean, channels C, G and dev-1 all data.

	Model	clean	C	G	dev-1
DOC-NMC	DNN	46.0	76.2	51.1	63.3
DOC-LSENPNCC	DNN	46.1	76.5	51.3	63.3
NMC-LSENPNCC	DNN	46.5	76.7	51.6	64.1
DOC-NMC	pCNN	44.5	74.6	49.2	60.7
DOC-LSENPNCC	pCNN	44.0	74.7	49.2	60.5
NMC-LSENPNCC	pCNN	44.9	74.8	50.0	61.4
DOC-NMC	fCNN	44.2	74.1	49.1	60.3
DOC-LSENPNCC	fCNN	43.8	74.0	48.5	**60.2**
NMC-LSENPNCC	fCNN	44.2	74.5	49.3	60.6

Table 3 shows that fusing the feature maps from the convolution layers was more effective than simple feature combination or fusion at the output layer. Next we performed KWS experiments where we evaluated performance by considering the average P(miss) over a P(FA) range of 0.0001 and 0.0005. Note that the "misses" are instances where the hypothesis failed to detect a keyword and the "false alarms" (FA) are instances where the hypothesis falsely detected a keyword. We used the KWS setup described in [29] to perform our KWS tasks, and we report the results obtained in table 4 which gives the avg. P(miss) from baseline (MFB and GFB) systems, best single feature systems and top 3 multi-feature systems, where both RATS LAR dev-1 and dev-2 data were used and dev-1 data was used for ranking and calibration. It should be noted that unlike our earlier reported system [29], this paper report results on a simple word-based KWS system without any rank-based score normalization or sub-word search. As shown before, rank-based normalization and the use of fuzzy keyword matching at the phonetic level can drastically improve performance, specifically in high-recall operating points.

Table 4. Average P(miss) for P(FA) between 0.0001 and 0.0005, obtained from baseline (MFB and GFB) systems, best single feature systems and best three fused feature systems based on dev-1 performance.

Feature(s)	Model	Avg. Pmiss @ (Pfa = 0.0001-0.0005)	
		dev1	dev2
MFB	DNN	46.7	34.4
GFB	CNN	44.2	31.0
DOC	CNN	43.9	31.0
DOC	TFCNN	43.6	30.9
DOC-NMC	fCNN	42.8	30.6
NMC-LSENPNCC	fCNN	43.0	30.1
DOC-NMC	pCNN	43.5	30.3

The NMC-LSENPNCC fCNN system gave the best result on dev-2, while the DOC-NMC fCNN system gave the best result on dev-1. Fusing features helped to improve the performance, but that gain was not significant with respect to the individual features. The WERs from the single and fused feature systems were quite similar, for example both DOC-TFCNN and DOC-NMC fCNN system gave similar overall WER for dev-1, however the benefit of fused feature systems were apparent from the KWS results, where the fused feature systems gave better performance than individual feature systems.

5. Conclusions

In this work, we presented different robust features and demonstrated their performance for speech recognition and a KWS task using the Levantine Arabic KWS dataset distributed through the DARPA RATS program. Our results indicate that feature map fusion of robust acoustic features can reduce the WER by 4.1% relative with respect to the MFB-DNN baseline, however the performance from a single feature (DOC) TFCNN system was found to be equally good. The relative reduction of P(miss) at P(fa) between 0.0001 to 0.0005 on dev-2 from the NMC-LSENPNCC fCNN system was found to be 12.5% compared to the MFB-DNN system. Our results indicate that feature fusion and fusion of DNN systems at the output layer can not only improve KWS performance but also improve robustness against heavily channel- and noise-degraded speech.

6. Acknowledgements

This material is based upon work partly supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-15-C-0037. The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

7. References

[1] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system", IEEE Trans. Speech Audio Process., 7(2), pp. 126–137, 1999.

[2] S. Srinivasan and D. L. Wang, “Transforming binary uncertainties for robust speech recognition”, IEEE Trans Audio, Speech, Lang. Process., 15(7), pp. 2130–2140, 2007.

[3] Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Adv. Front-end Feature Extraction Algorithm; Compression Algorithms, ETSI ES 202 050 Ver. 1.1.5, 2007.

[4] C. Kim and R. M. Stern, “Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring”, in Proc. ICASSP, pp. 4574–4577, 2010.

[5] V. Tyagi, “Fepstrum features: Design and application to conversational speech recognition”, IBM Research Report, 11009, 2011.

[6] V. Mitra, H. Franco, M. Graciarena and A. Mandal, “Normalized amplitude modulation features for large vocabulary noise-robust speech recognition”, in Proc. of ICASSP, pp. 4117-4120, Japan, 2012.

[7] V. Mitra, H. Franco, M. Graciarena, and D. Vergyri, “Medium duration modulation cepstral feature for robust speech recognition,” Proc. of ICASSP, Florence, 2014.

[8] A. Mohamed, G.E. Dahl, and G. Hinton, “Acoustic modeling using deep belief networks,” IEEE Trans. on ASLP, vol. 20, no. 1, pp. 14 –22, 2012.

[9] F. Seide, G. Li, and D. Yu, “Conversational speech transcription using context-dependent deep neural networks,” Proc. of Interspeech, 2011.

[10] T. Sainath, A. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural network for LVCSR," Proc. of ICASSP, 2013.

[11] O. Abdel-Hamid, A. Mohamed, H. Jiang, and G. Penn, “Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition,” Proc. of ICASSP, pp. 4277–4280, 2012.

[12] V. Mitra, W. Wang, H. Franco, Y. Lei, C. Bartels, and M. Graciarena, “Evaluating robust features on deep neural networks for speech recognition in noisy and channel mismatched conditions,” in Proc. of Interspeech, 2014.

[13] V. Mitra, W. Wang, and H. Franco, “Deep Convolutional Nets and Robust Features for Reverberation-robust Speech Recognition,” in Proc. of SLT, pp. 548–553, 2014.

[14] P. Zhan and A Waibel, “Vocal tract length normalization for LVCSR,” in Tech. Rep. CMU-LTI-97-150. Carnegie Mellon University, 1997

[15] H. Lee, P. Pham, Y. Largman, and A. Ng, “Unsupervised feature learning for audio classification using convolutional deep belief networks,” Proc. of Adv. Neural Inf. Process. Syst. 22, pp. 1096–1104, 2009.

[16] D. Hau and K. Chen, “Exploring hierarchical speech representations using a deep convolutional neural network,” Proc. of 11th UK Workshop Comput. Intell. (UKCI ’11), 2011.

[17] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, “Phoneme recognition using time-delay neural networks,” IEEE Trans. Acoust., Speech, Signal Process., 38(3), pp. 328–339, 1989.

[18] V. Mitra, and H. Franco, “Time-frequency convolution networks for robust speech recognition,” Proc. of ASRU, 2015.

[19] A. Mandal, J. van Hout, Y-C. Tam, V. Mitra, Y. Lei, J. Zheng, D. Vergyri, L. Ferrer, M. Graciarena, A. Kathol, and H. Franco, "Strategies for High Accuracy Keyword Detection in Noisy Channels," in Proc. of Interspeech, pp. 15-19, 2013.

[20] T. Ng, R. Hsiao, L. Zhang, D. Karakos, S.H. Mallidi, M. Karafiat, K. Vesely, I. Szoke, B. Zhang, L. Nguyen, and R. Schwartz, “Progress in the BBN Keyword Search System for the DARPA RATS Program,” Proc. of Interspeech, pp. 959-963, 2014.

[21] L. Mangu, H. Soltau, H-K. Kuo, B. Kingsbury, and G. Saon, "Exploiting diversity for spoken term detection," in Proc. of ICASSP, pp. 8282-8286, 2013.

[22] V. Mitra, J. van-Hout, H. Franco, D. Vergyri, Y. Lei, M. Graciarena, Y-C. Tam and J. Zheng, “Feature fusion for high-accuracy keyword spotting,” in Proc. of ICASSP, pp. 7193-7197, Florence, 2014.

[23] K.Walker and S. Strassel, “The RATS radio traffic collection system,” in Proc. of ISCA, Odyssey, 2012.

[24] V. Mitra, H. Franco, and M. Graciarena, “Damped oscillator cepstral coefficients for robust speech recognition,” Proc. of Interspeech, pp. 886–890, 2013.

[25] J. van-Hout and A. Alwan, “A novel approach to softmask estimation and log-spectral enhancement for robust speech recognition,” Proc. of ICASSP, pp. 4105–4108, 2012.

[26] J. van Hout, “Low Complexity Spectral Imputation for Noise Robust Speech Recognition,” Master’s thesis, University of California, Los Angeles, USA, 2012.

[27] C. Kim and R. M. Stern, “Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition,” IEEE Trans. on Audio, Speech, and Language Processing, accepted for publication, 2016.

[28] B. T. Meyer, S. V. Ravuri, M. R. Sch¨adler, and N. Morgan, “Comparing different flavors of spectro-temporal features for ASR,” in Proc. of Interspeech, 2011, pp. 1269–1272.

[29] J. van-Hout, V. Mitra, Y. Lei, D. Vergyri, M. Graciarena, A. Mandal and H. Franco, “Recent improvements in SRI’s Keyword Detection System for Noisy Audio,” in Proc. of Interspeech, pp. 1727-1731, Singapore, 2014.