FEATURE FUSION FOR HIGH-ACCURACY KEYWORD SPOTTING

Vikramjit Mitra, Julien van Hout, Horacio Franco, Dimitra Vergyri, Yun Lei, Martin Graciarena, Yik-Cheung Tam, Jing Zheng

[1]Speech Technology and Research Laboratory, SRI International, Menlo Park, CA

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

ABSTRACT

This paper assesses the role of robust acoustic features in spoken term detection (a.k.a keyword spotting—KWS) under heavily degraded channel and noise corrupted conditions. A number of noise-robust acoustic features were used, both in isolation and in combination, to train large vocabulary continuous speech recognition (LVCSR) systems, with the resulting word lattices used for spoken term detection. Results indicate that the use of robust acoustic features improved KWS performance with respect to a highly optimized state-of-the art baseline system. It has been shown that fusion of multiple systems improve KWS performance, however the number of systems that can be trained is constrained by the number of frontend features. This work shows that given a number of frontend features it is possible to train several systems by using the frontend features by themselves along with different feature fusion techniques, which provides a richer set of individual systems. Results from this work show that KWS performance can be improved compared to individual feature based systems when multiple features are fused with one another and even further when multiple such systems are combined. Finally this work shows that fusion of fused and single feature bases systems provide significant improvement in KWS performance compared to fusion of single-feature based systems.

Index Terms— *feature combination, noise robust keyword spotting, large vocabulary speech recognition, robust acoustic features, system combination.*

1. INTRODUCTION

KWS entails detecting keywords that are either single-word or multi-word terms in the acoustic speech signals. The most common KWS approach (also called "spoken term detection") uses an LVCSR system to hypothesize words or subword-units from the speech signal and generates a word lattice with indexed words. Next, a search performed within the indexed data for the key words generates a list of keyword occurrences, each with a corresponding time at which it was hypothesized to exist in the speech data. A detailed survey of KWS approaches are given in [1, 2]

The performance of a KWS system is evaluated using different measures, which count the number of (1) "hits"— instances where a correct hypothesis was made; (2) "misses"— instances where the hypothesis failed to detect a keyword; and (3) "false alarms"— instances where the hypothesis falsely detected a keyword. These measures can be used to generate Receiver Operating Characteristic (ROC) curves that depict the overall performance of the KWS system.

In work conducted under the U.S. Defense Advanced Research Projects Agency's (DARPA's) Robust Automatic Speech Transcription (RATS) program, we performed KWS experiments on conversational speech that was heavily distorted by transmission channel and noise, resulting in very low signal-to-noise ratios (SNRs). This paper focuses on the Levantine Arabic (LAR) KWS task that we conducted.

State-of-the-art KWS systems proposed so far [3, 4, 5] have mostly focused on training multiple KWS systems and then fusing their outputs to generate a highly accurate final KWS result. It is usually observed that fusion of multiple systems provides better KWS performance than their individual counterparts; however the realization of the number of individual systems is constrained by the number of acoustic frontends. This paper explores: (1) different ways to fuse multiple acoustic features for training robust KWS systems and compare their performance with respect to the individual feature based systems and finally (2) compare KWS performance between fusion of individual feature based systems and fusion of multi-feature fusion based KWS systems. Although score level fusion is conventionally used in the KWS community for ensuring robust and high-accuracy KWS systems, to the best of our knowledge our work is the first that proposes feature combination and demonstrates that such a combination can effectively produce high-accuracy candidate KWS systems, that results in highly robust KWS systems after system-level fusion.

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 [6], 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 hrs 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 is 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 here; each consisted of 10 hrs of held-out conversational speech. While dev-1 was used to tune and optimize the system fusion parameters, dev-2 was used to measure the KWS performance. A set of 200 keywords was pre-specified for the LAR 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.

3. THE LAR SPEECH RECOGNITION SYSTEM

We used a Gaussian Mixture Model (GMM)-Hidden Markov Model (HMM) based speech activity detection (SAD) system to segment the speech signals from dev-1 and dev-2. More details about the SAD system are provided in [5, 7].

The AM of the LAR LVCSR system was trained using different acoustic features: (1) traditional Perceptual Linear Prediction features using RASTA processing (RASTA-PLP) [8], (2) Normalized Modulation Cepstral Coefficients (NMCC) [9], and (3) Modulation of Medium Duration Speech Amplitude (MMeDuSA) features [10, 18]. We also explored a combination of these acoustic features, followed by their dimensionality reduction using traditional principal component analysis (PCA), heteroscedastic linear discriminant analysis (HLDA) and a nonlinear autoencoder (AE) network.

3.1 NMCC

NMCC [9] was obtained from tracking the amplitude modulations of subband speech signals in a time domain by using a hamming window of 25.6 ms with a frame rate of 10 ms. NMCC features are obtained by analyzing speech using a time-domain gammatone filterbank with 40 channels equally spaced in the equivalent rectangular bandwidth (ERB) scale. Each of the subband signals was then processed using the Discrete Energy Separation algorithm (DESA) [11], which produces instantaneous estimates of amplitude and frequency modulation of the bandlimited subband signals. The amplitude modulation signals in the analysis window of 25.6 ms were used to compute the amplitude modulation power, which were then power compressed using $1/15^{th}$ root compression. Discrete cosine transform (DCT) was performed on the resulting powers to generate cepstral features (for additional details, see [9]). We used 13 cepstral coefficients and their Δ, Δ^2, and Δ^3 coefficients, which yielded a 52-dimensional feature vector.

3.2 MMeDuSA

The MMeDuSA feature generation pipeline used a time-domain gammatone filterbank with 30 channels equally spaced in the ERB scale. It used the nonlinear Teager energy operator [12] to estimate the amplitude modulation signal from the bandlimited subband signals. The MMeDuSA pipeline used a medium duration hamming analysis window of 51.2 ms with 10 ms frame rate and computed the amplitude modulation power over the analysis window. The powers were root compressed and then their DCT coefficients were obtained, out of which the first 13 coeffcients were retained. These 13 cepstral coefficients along with their Δ, Δ^2, and Δ^3 coefficients resulted in a 52-dimensional feature set. Additionally, the amplitude modulation signals from the subband channels were bandpass filtered to retain information in the 5 to 200 Hz range, with that information then summed across the frequency channels to produce a summary modulation signal. The power signal of the modulation summary was obtained, followed by $1/15^{th}$ root compression. The result was transformed using DCT and the first three coefficients were retained and combined with the previous 52-dimensional features to produce the 55-dimensional MMeDuSA features.

3.3 Feature combination and dimensionality reduction

This paper explores the role of feature combination in KWS performance. Note that combination of multiple features result in large dimensional feature sets that are not suitable for GMM-HMM based AM training. To obtain better control over the dimensionality of the features, we explored different ways of dimensionality reduction. In the first approach, we performed a PCA transform on the resulting features, thereby ensuring that at least 90% of the information was retained.

In the second approach we explored HLDA based dimensionality reduction directly on the individual features before concatenating them. Each of the features, NMCC, PLP and MMeDuSA were HLDA transformed to 20 dimensions and then were combined. In this case a combination of two features, such as NMCC+PLP and NMCC+MMeDuSA produced a final feature vector of 40 dimensions, but a 3-way fusion of NMCC+PLP+MMeDuSA results in a final feature dimension of 60. In the latter case we performed another level of HLDA to reduce the 60 dimensional features to 40.

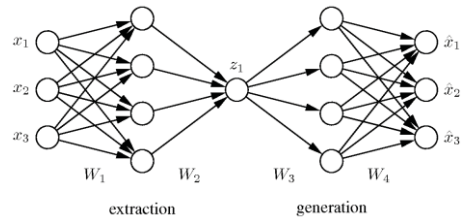

Figure 1. An AE network

Finally, we explored the use of an AE network. An AE network consists of two parts (see Figure 1): (1) an *extraction* part that projects the input to an arbitrary space (in this work it was a lower dimensional space than the input space, represented by z in Figure 1) and (2) a *generation* part that projects back from the intermediary arbitrary space to the output space, where the outputs are an estimate of the input. In essence, the AE maps the input to itself and in the process of doing so its hidden variables (representing the arbitrary intermediate space) learn the acoustic space as defined by the input acoustic features. Once the AE is trained, the generation part of the network can be discarded with only the outputs from the extraction part used as features. In one sense the network then performs a nonlinear transform (assuming the network uses a nonlinear activation function, in our experiments we used a tan-sigmoid function) of the input acoustic space to generate broad acoustic separation of the data and then in the process perform dimensionality reduction if the dimension of z was lower than the input acoustic space. Note that this strictly acoustic-data-intensive approach does not require phone labels or other form of textual representation as do artificial neural network (ANN) based tandem features [14]. In our experiments we selected the dimension of z to be 39 for two-way feature fusion and 52 for three-way feature fusion, where in the latter case the dimension was further reduced to 39 through HLDA.

Table 1 shows the naming convention of the combined features with their dimensionality reduction techniques. Note that all the candidate features used in our experiments have speaker level vocal tract length normalization (VTLN) as we observed that VTLN helped to bring the ROC curve down compared to their non-VTLN counterparts.

Table 1. Different combination of features and their dimensionality reductions used in the experiments

Input Features		Dimensionality Reduction		
	Dim.	Type	Feature name	Dim.
NMCC(52), MMeDuSA(55)	107	PCA	NMCC+MMeDuSA_pca	40
		HLDA	NMCC+MMeDuSA_hlda	40
		AE	NMCC+MMeDuSA_AE	39
NMCC(52), PLP(52)	104	PCA	NMCC+PLP_pca	40
		HLDA	NMCC+PLP_hlda	40
		AE	NMCC+PLP_AE	39
NMCC(52), PLP(52), MMeDuSA(55)	159	PCA+ HLDA	NMCC+PLP+MMeDuSA-pca_hlda	39
		HLDA	NMCC+PLP+ MMeDuSA_hlda	40
		AE+ HLDA	NMCC+PLP+MMeDuSA-AE_hlda	39

3.4 Acoustic Modeling (AM)

For AM training, we used data from all the eight noisy channels available in the LAR RATS-KWS training data to train multi-channel AMs that used three-state left-to-right HMMs to model crossword triphones. The training data was clustered into speaker clusters using unsupervised agglomerative clustering. Acoustic features used for training the HMM were normalized using standard cepstral mean and variance normalization. The AMs were trained using SRI International's DECIPHER™ LVCSR system [15]. We trained speaker-adaptive maximum likelihood (ML) models, where the models were speaker-adapted using ML linear regression (MLLR).

3.5 Language Modeling

The LM was created using SRILM [16], with the vocabulary selected as described in the approach in [17]. Using a held-out tuning set we selected a vocabulary of 47K words for LAR, which resulted in an out of vocabulary (OOV) rate of 4.3% on dev-1. We added to this vocabulary the prespecified keyword terms so that no OOV keywords occurred during the 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 the RATS-KWS LAR corpora. More details about the LM used in our experiments are provided in [5].

4. KWS

We used the ASR lattices generated from our LAR LVCSR system as an index to perform the KWS search. ASR word lattices from the LAR LVCSR system were used to create a candidate term index by listing all words in the lattice along with their start/end time and posterior probabilities. A tolerance of 0.5 s was used to merge the multiple occurrences of a word at different times. The KWS output of each system was obtained by taking the subset of words in the index that were keywords. The n-gram keywords added to the LM were treated as single words in the lattices and therefore appeared in the index. We added links in the word lattices where two or three consecutive nodes formed a keyword. These links allowed recovery of multiword keywords for which the ASR search hypothesized the sequence of words forming the keyword instead of the keyword itself. More details about the KWS system used in our experiments can be obtained in [5].

Fusion of keyword detections from multiple systems is done in two steps: first, the detections were aligned across systems using a tolerance of one second to create a vector of scores for each fused detection. The fused scores were obtained by linearly combining the individual scores in the logit domain using logistic regression. More details on this approach was presented in [5]

5. RESULTS

We present the KWS performance in terms of two metrics, (1) False Alarm (FA) rate at 34% P(miss), and (2) P(miss) at 1% FA. These two metrics provide information about the ROC curve from the KWS experiment at a region that is critical to the DARPA RATS KWS task, whose main goal is to obtain a system with a lower FA rate. Table 2 provides these two metrics for the individual acoustic features and their fusion, while Figure 2 presents their ROC curves.

Table 2. KWS performance for the individual feature based systems on RATS LAR dev-2 dataset

Features	FA(%) at 34% P(miss)	P(miss)(%) at 1% FA
PLP	1.06	34.12
NMCC	0.76	32.86
MMeDuSA	0.97	33.33
Fusion	0.39	26.42

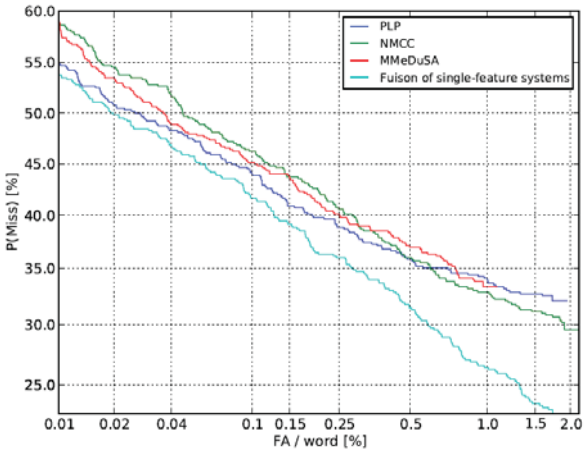

Figure 2. KWS ROC curves from the individual feature-based systems and their 3-way system-fusion

Table 2 and Figure 2 clearly indicate that system fusion significantly improves KWS performance by lowering the ROC curve appreciably. The FA rate at 34% P(miss) and P(miss) at 34% FA was reduced by 48.7% and 20.6% compared to the best performing system (NMCC) at that operating point. The ROC curve shows that while PLP gave lower P(miss) for FA less than 0.5, but for higher FA rates NMCC performed better than PLP.

Table 3 shows the KWS performance of the fused feature systems. The AE based dimension-reduced features clearly didn't perform as good as its PCA and HLDA counterparts but interestingly they contributed well during system fusion and helped to reduce the FA at the operating point. Note that the ROC curve for NMCC+PLP+MMeDuSA-AE_hlda based system did not go down to 34% P(miss) hence FA at that operating point is not reported in the table. The last two rows in table 3 shows the result from the fusion of 3-best feature-combined systems and fusion of all systems (which included both single- and combined-feature systems). Comparing Table 2 and Table 3 we can see that the 3-way single feature system fusion is worse than the fusion of 3-best combined-feature systems indicating that feature-fusion based system may be providing richer KWS systems for system combination compared to the individual feature based systems.

Table 3. KWS performance for the fused feature based systems on RATS LAR dev-2 dataset

Features	FA(%) at 34% P(miss)	P(miss)(%) at 1% FA
NMCC+MMeDuSA_pca	0.88	32.23
NMCC+MMeDuSA_hlda	0.97	33.49
NMCC+MMeDuSA_AE	2.02	38.36
NMCC+PLP_pca	0.73	32.08
NMCC+PLP_hlda	0.79	32.08
NMCC+PLP_AE	2.30	41.67
NMCC+PLP+MMeDuSA-pca_hlda	0.78	32.08
NMCC+PLP+MMeDuSA_hlda	0.71	31.29
NMCC+PLP+MMeDuSA-AE_hlda	-	43.55
Fusion of 3-best systems (combined feature systems only)	0.31	25.63
Fusion of all systems (inclusive of single & combined feature systems)	0.26	24.06

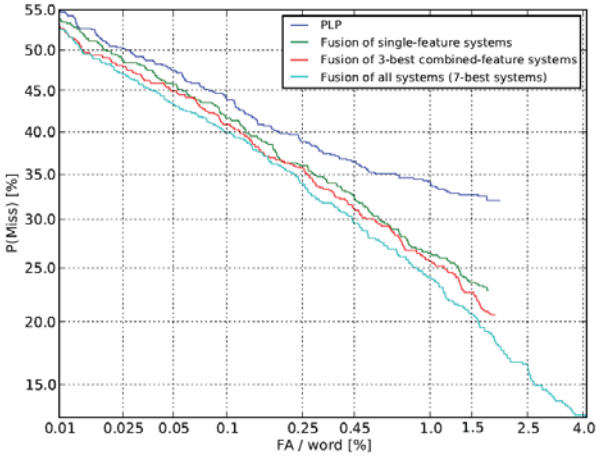

Figure 3. KWS ROC curves from the baseline system (PLP), Fusion of single-feature systems (PLP-NMCC-MMeDuSA), Fusion of 3-best combined-feature systems and fusion of all systems (both single- and combined-feature based systems).

Figure 3 shows the ROC curves for the baseline PLP-system, Fusion of single-feature systems (3-way fusion), fusion of 3-best combined-feature systems and finally the fusion of both single- and combined-feature based systems. The fusion of 3-best combined-feature systems and the 3-way single-feature system fusion results are directly comparable as both of them use three systems only and the same native features. The lowering of the ROC curve from the fusion of 3-best combined-feature system indicates that the feature combination approach is able to exploit the complementary information amongst the features and resulting in better and richer lattices. We observed that feature combination typically results in an increased lattice size, where we observed a maximum relative lattice size increase of 38% compared to a single feature system. The ROC curve from the fusion of all systems show that the generation of multiple candidate systems through feature fusion provides us with richer systems and more options for system level fusion. The best fusion of all systems in fact is the fusion of 7 systems which gave the best ROC curve and those 7 systems consists of the following features: NMCC+PLP+MMeDuSA_hlda, NMCC+PLP_hlda, NMCC, NMCC+PLP_pca, NMCC+PLP+MMeDuSA-pca_hlda, NMCC+PLP+MMeDuSA-AE_hlda and NMCC+MMeDuSA_AE. The fusion of 3-best combined-feature systems consisted of the following candidate systems: NMCC+PLP+ MMeDuSA_hlda, NMCC+PLP_pca and NMCC+PLP+MMeDuSA-pca_hlda.

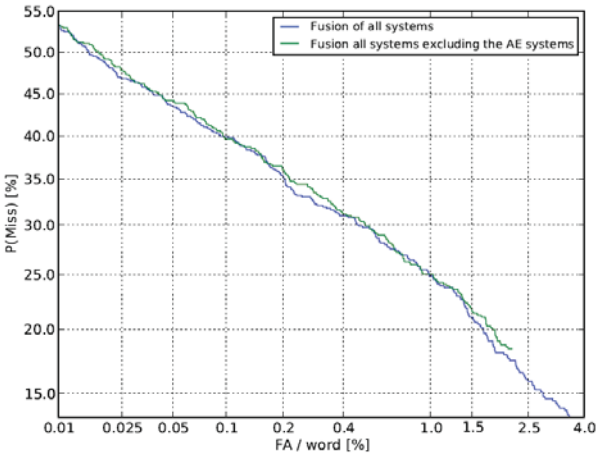

Figure 4. KWS ROC curves from the fusion of all systems and fusion of all systems excluding the AE-based systems.

The interesting aspect of this study is that it uses the same candidate set of acoustic features (NMCC, PLP and MMeDuSA) and assuming that AM and LM remain the same, we observed that by different ways of feature combination we can improve the KWS performance appreciably. Also we observed that even if the AE based feature combination did not result in better individual KWS systems, but such systems captured sufficient complimentary information and hence contributed in slightly lowering the ROC curve as shown in Figure 4, where we show the ROC curve from the fusion of all systems and fusion of all systems except the AE based systems. Figure 4 also shows that the fusion of AE based systems helped the ROC curve to achieve a P(miss) lower than 15% and this happens because the AE-based systems produces more false-alarms compared to others resulting in an extend the ROC curve going beyond the 15% FA region.

6. CONCLUSION

In this work we presented different ways to combine multiple features for training acoustic models for DARPA-RATS LAR KWS task. Our results show that the relative P(miss) can be reduced by 2.4% at 1%FA and the relative FA can be reduced by 6.6% at 34% P(miss) by using feature combination compared to single-feature bases systems. Combining systems using single features and different feature combinations reduces the relative P(miss) at 1% FA by approximately 29.5% compared with the PLP baseline system and by 8.9% compared to the fusion of the single-feature systems. Our results indicate that judicious selection of feature fusion, advanced dimensionality reduction techniques, and fusion of multiple systems can appreciably improve the accuracy of KWS task on heavily channel and noise degraded speech. Future studies will explore similar strategies in a deep neural network acoustic modeling setup.

7. ACKNOWLEDGMENT

This material is based upon 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.
Disclaimer: Research followed all DoD data privacy regulations.
Approved for Public Release, Distribution Unlimited

8. REFERENCES

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

[2] I. Szoke, P. Schwarz, P. Matejka, L. Burget, M. Karafiat, 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.

[3] M.S. Seigel, P.C. Woodland, and M.J.F. Gales, "A confidence-based approach for improving keyword hypothesis scores", in Proc. of ICASSP, pp. 8565-8569, 2013.

[4] 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

[5] 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.

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

[7] M. Graciarena, A. Alwan D. Ellis, H. Franco, L. Ferrer, J.H.L. Hansen, A. Janin, B-S. Lee, Y. Lei, V. Mitra, N. Morgan, S.O. Sadjadi, T.J. Tsai, N. Scheffer, L.N. Tan, and B. Williams, "All for One: Feature Combination for Highly Channel-Degraded Speech Activity Detection," in Proc. of Interspeech, pp. 709-713, 2013.

[8] H. Hermansky and N. Morgan, “RASTA processing of speech,” IEEE Trans. Speech Audio Process., Vol. 2, pp. 578–589, 1994.

[9] 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, 2012.

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

[11] A. Potamianos and P. Maragos, “Time-frequency distributions for automatic speech recognition”, in IEEE Trans. Speech & Audio Proc., vol. 9(3), pp. 196–200, 2001.

[12] H. Teager, “Some observations on oral air flow during phonation”, in IEEE Trans. ASSP, pp. 599–601, 1980.

[13] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P-A. Manzagol "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," The Journal of Machine Learning Research, vol. 11, pp. 3371-3408, 2010.

[14] Q. Zhu, B. Chen, N. Morgan, and A. Stolcke, "Tandem Connectionist Feature Extraction for Conversational Speech Recognition," Machine Learning for Multimodal Interaction, Lecture Notes in Computer Science, Vol. 3361, pp. 223-231, 2005.

[15] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. Gadde, M. Plauché, C. Richey, E. Shriberg, K. Sönmez, F. Weng, and J. Zheng, “The SRI March 2000 Hub-5 conversational speech transcription system,” in Proc. NIST Speech Transcription Workshop, College Park, MD, 2000.

[16] A. Stolcke, “SRILM - An Extensible Language Modeling Toolkit”, in Proc. of ICSLP, pp. 901-904, 2002.

[17] A. Venkataraman and W. Wang, “Techniques for effective vocabulary selection,” in Proc. Eighth European Conference on Speech Communication and Technology, pp. 245-248, 2003.

[18] V. Mitra, M. McLaren, H. Franco, M. Graciarena, and N. Scheffer, "Modulation Features for Noise Robust Speaker Identification," in Proc. of Interspeech, pp. 3703-3707, 2013.