DEEP CONVOLUTIONAL NETS AND ROBUST FEATURES FOR REVERBERATION-ROBUST SPEECH RECOGNITION

Vikramjit Mitra, Wen Wang, Horacio Franco

[1]Speech Technology and Research Laboratory, SRI International, Menlo Park, CA
{vikramjit.mitra,wen.wang,horacio.franco}@sri.com

ABSTRACT

While human listeners can understand speech in reverberant conditions, indicating that the auditory system is robust to such degradations, reverberation leads to high word error rates for automatic speech recognition (ASR) systems. In this work, we present robust acoustic features motivated by human speech perception for use in a convolutional deep neural network (CDNN)-based acoustic model for recognizing continuous speech in a reverberant condition. Using a single-feature system trained with the single channel data distributed through the REVERB 2014 challenge on ASR in reverberant conditions, we show a substantial relative reduction in word error rates (WERs) compared to the conventional filterbank energy-based features for single-channel simulated and real reverberation conditions. The reduction is more pronounced when multiple features and systems were combined together. The proposed system outperforms the best system reported in REVERB-2014 challenge in single channel full-batch processing task.

Index Terms—*deep convolutional networks, feature combination, robust speech recognition, reverberation robustness, robust features.*

1. INTRODUCTION

Current state-of-the-art automatic speech recognition (ASR) systems are quite sensitive to speech signal degradations such as reverberation, noise and channel mismatch, which can result in significantly reduced speech recognition accuracy. Reverberation is one of the major sources of performance degradation for ASR systems [1] and in the past few years there has been an increased surge in exploring reverberation robust strategies for ASR systems. The environment from which speech is captured primarily defines the character of the reverberation and its effect on speech. Reverberation is typically caused by multiple reflections of the source sound from the ambient enclosure; such distortions seriously degrade speech signal quality. Approaches to circumvent reverberation effects on speech have become an important research topic lately, with microphone array processing [2]; echo cancellation [3]; robust signal processing [4]; and speech enhancement [5] acting as major research directions.

Typically, acoustic mismatch between training and testing conditions is a key contributor of performance degradation for ASR systems. ASR systems trained on clean data (i.e., data without artifacts) usually suffer a huge performance loss when deployed in reverberant conditions. This effect can be mitigated by training the ASR system with reverberant data, which helps to reduce the mismatch between the training and testing data [6].

Robust signal processing techniques and de-reverberation strategies have been explored in [7, 8, 9, 10, 11], demonstrating that the use of suitable acoustic features can improve the reverberation robustness of ASR systems. Recent advances in neural network technology have redefined the common strategies used in ASR systems, where Gaussian Mixture Model (GMM)-based Hidden Markov Models (HMM) are replaced with a more accurate neural network-based acoustic model. Deep Neural Network (DNN)-based acoustic models have simplified a lot of steps in ASR systems; for example, cepstral features are no longer a necessity, and one can build a highly accurate ASR system using only the spectral energies [12]. Once widely used, vocal tract length normalization (VTLN)-based speaker normalization [13] no longer seems to have a significant impact on speech recognition accuracy, as DNN's rich projections through multiple hidden layers allows it to learn a speaker-invariant representation of the data. Recently, we observed [14] that VTLNs make a much lesser impact on ASR accuracy for Convolutional DNNs (CDNNS) [27, 28] compared to traditional DNNs. [14] also observed that that the CDNNs showed more noise and channel robustness than DNNs.

In this work, we explore a set of robust acoustic features in a CDNN framework and compare different robust acoustic features with respect to baseline mel-filterbank energies. We also explored the role of deltas and feature fusion in our experiments. Recent studies [15] have shown that the use of i-vectors [16] as features seem to benefit DNN-based ASR systems in performing speaker adaptation. We explored both single i-vectors and i-vector based fusions in our experiments, and report the observations in this paper. We used the data distributed through the REVERB (REverberant Voice Enhancement and Recognition Benchmark) 2014 challenge [1] to train and evaluate our systems. We trained individual feature-based systems and observed that robust features almost always outperformed mel-filterbank features in all reverberant conditions.

The paper is structured as follows. First, in Section 2, we briefly describe the REVERB 2014 dataset used in our experiments. In Section 3 we present the different feature-extraction strategies used in our work. In Section 4, we present the ASR system used in our work. In Section 5, we show the results from our experiments. Finally, in Section 6, we present our conclusions.

2. DATASET AND TASK

In this work we used the REVERB 2014 challenge dataset, which contains single-speaker utterances recorded with one-channel, two-channel, or eight-channel circular microphone arrays. The dataset includes training, development and evaluation sets. The training set consists of the clean WSJCAM0 [17] dataset, which was convolved with room impulse responses (with reverberation times from 0.1 sec to 0.8 sec) and then corrupted with background noise. The evaluation and development data contain both real recordings

and simulated data. The real data is borrowed from the MC-WSJ-AV corpus [18], which consists of utterances recorded in a noisy and reverberant room. For the simulated data, reverberation effects were artificially introduced. Detailed information about the dataset is given in [1]. In all of our experiments reported here we used the channel-1 training data, which contained altogether 7861 utterances (5699 unique utterances) to build our acoustic models,. The simulated dev. set had 742 utterances in each of far and near microphone conditions spread almost equally in three room types (1, 2, and 3) and the real dev. set had 179 utterances spread almost equally into near and far microphone conditions. The simulated evaluation set contained 1088 utterances in each of the far and near microphone conditions, each of which were split into three room conditions (1, 2 and 3). Each room condition reflected a different reverberation time. The real evaluation set contained 372 utterances split equally between near and far microphone conditions.

In our work, we used SRI's DECIPHER large-vocabulary, continuous speech recognition (LVCSR) system as the baseline system to train and test the acoustic models. We noticed that our baseline system is slightly better than the baseline system provided by the REVERB 2014 challenge organizers. In all our experiments no speaker information was used during acoustic model training and testing, and all processing was independent of the room impulse responses and the relative position of the speakers with respect to the recording device. We report our results in terms of word error rates (WER) using conditions identical to those of the baseline system distributed with the REVERB 2014 challenge data.

3. ACOUSTIC FEATURES

We explored an array of robust features for our experiments, motivated by human auditory perception and speech production. The features explored are briefly outlined in this section.

3.1 Damped Oscillator Coefficients (DOC)

DOC [19] tries to model the dynamics of the hair cells within the human ear as forced damped oscillators. 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. In DOC processing, the incoming speech signal is analyzed by a bank of gammatone filters which splits the signal into bandlimited subband signals. In this work, we used a bank of 40 gammatone filters equally spaced on the equivalent rectangular bandwidth (ERB) scale. In turn, these subband signals are used as the forcing functions to an array of damped oscillators whose response is 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, then root compressed using the 15^{th} root and the resulting 40 dimensional features were used as the DOC feature in our experiments.

3.2 Normalized Modulation Coefficients (NMC)

NMC [20] is motivated by the fact that amplitude modulation (AM) of subband speech signals plays an important role in human speech perception and recognition. These features were obtained from tracking the amplitude modulations of subband speech signals in a time domain using a Hamming window of 26 ms with a frame rate of 10 ms. In this processing, the speech signal was analyzed using a time-domain gammatone filterbank with 40 channels equally spaced on the ERB scale. The subband signals were then processed using the Discrete Energy Separation algorithm (DESA) [21], which produced instantaneous estimates of AM signals. The powers of the AM signals were then root compressed using 15^{th} root. The resulting 40-dimensional feature vector was used as the NMC feature in our experiments.

3.3 Modulation of Medium Duration Speech Amplitudes (MMeDuSA)

MMeDuSA [22, 23] track the subband AM signals of speech using a medium duration analysis window. They also track the overall summary modulation information. The summary modulation plays an important role in both tracking speech activity and locating events such as vowel prominence/stress, etc.. The MMeDuSA-generation pipeline used a time-domain gammatone filterbank with 40 channels equally spaced on the ERB scale. It employed the nonlinear Teager energy operator [24] to crudely estimate the AM signal from the bandlimited subband signals. The MMeDuSA pipeline used a medium duration Hamming analysis window of ~51 ms with a 10 ms frame rate and computed the AM power over the analysis window. The powers were root compressed and the resultant information was used as the acoustic feature in our experiments.

3.4 Gammatone Filter Coefficients (GFCs)

The gammatone filters are a linear approximation of the auditory filterbank performed in the human ear. In GFC processing, speech is analyzed using a bank of 40 gammatone filters equally spaced on the equivalent rectangular bandwidth (ERB) scale. The power of the band limited time signals within an analysis window of ~26 ms was computed at a frame rate of 10 ms. Subband powers were then root compressed using the 15^{th} root and the resulting 40-dimensional feature vector was used as the GFCs.

4. ACOUSTIC MODEL

We used both traditional Gaussian mixture model (GMM) hidden Markov model (HMM)-based acoustic model and CDNN-based acoustic model in our experiments. For the GMM-HMM acoustic model training, we used SRI International's DECIPHER® LVCSR system, which employs a common acoustic frontend that computes 13 MFCCs (including energy) and their Δs, Δ^2s, and Δ^3s. Global mean and variance normalization was performed on the acoustic features prior to acoustic model training. The acoustic models were trained as crossword triphone HMMs with decision-tree-based state clustering that resulted in 2048 fully tied states, and each state was modeled by a 64-component Gaussian mixture model. The GMM-HMM model was trained with maximum likelihood estimation and used the 5K non-verbalized punctuation, closed vocabulary set language model (LM) during decoding. A bigram LM is used in the initial pass of decoding to generate the first-pass hypotheses. In the second pass, model-space MLLR adaptation of the cross-word acoustic models was conducted based on the first-pass hypotheses. The adapted acoustic models were used for generating bigram HTK lattices, which were then rescored by the trigram LM. Both bigram and trigram are the 5K closed vocabulary non-verbalized-punctuation LMs trained on the WSJ CSR LM training data [25]. The HTK baseline system provided with the REVERB-2014 dataset used MFCC acoustic features to

train crossword triphone models where cMLLR adaptation was used during decoding.

To generate alignments to train the CDNN systemm a GMM-HMM model was used to produce the senones' labels. There were altogether 3276 senones produced by the GMM-HMM system. The input layer of the CDNN systems was formed using a context window of 15 frames (7 frames on either side of the current frame). We also explored different numbers of filter banks in our features and observed 40 to be the near-optimal selection.

The CDNN acoustic model was trained using cross entropy on the alignments from the GMM-HMM system. The input features are filterbank energy coefficients with a context of 7 frames from each side of the center frame for which predictions are made. Two hundred convolutional filters of size 8 were used in the convolutional layer, and the pooling size is set to three without overlap. Note that only one convolutional layer was used in our CDNN. The subsequent CDNN included four hidden layers, with 1024 nodes per hidden layer, and the output layer, with 3276 nodes representing the senones. The networks were trained 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. Back propagation was performed using stochastic gradient descent with a mini-batch of 256 training examples.

5. RESULTS

In all our experiments we used full based batch processing, where no prior information about the speakers, room conditions, or background noise was employed. We present our results for three baseline systems: (1) the MFCC-HTK system distributed through the REVERB 2014 challenge website and (2) the DECIPHER-MFCC system that was trained in our REVERB 2014 submission [11] and the mel-filterbank (MFB) feature-based CDNN system that we trained in this work. In table 1 presents the results from all the baseline mel-filterbank based systems. Table 1 shows that our CDNN-based system performed much better than either of the GMM-HMM systems. Note that both the GMM-HMM systems used some form of adaptation (cMLLR and MLLR), while the CDNN system had none. Also note that the far and near WERs correspond to the averaged WER from the three rooms present in each of those conditions.

Table 1. WERs from different baseline systems using simulated eval. data for decoding.

	WER (sim. eval. data)		
	Far (avg.)	Near (avg.)	Avg.
GMM-HMM HTK [1]	19.34	31.64	25.49
GMM-HMM DECIPHER [11]	14.54	23.40	18.97
CDNN (MFB)	12.67	19.47	16.07

Given the results in table 1 we explored all of our robust feature experiments to CDNN-based acoustic models only. We initially explored the role of delta features in CDNN performance by picking the DOC feature and training three CDNN models using (1) DOC features only, (2) DOC + Δ and (3) DOC + Δ + Δ^2. Table 2 shows the WERs from dev. lists for these three models. Results show that WERs from DOC + Δ features were lowest for all the conditions. Henceforth, we always used our features along with their first Δ coefficients. Note that adding Δ^2 increased the dimensionality of the feature space, which may have negatively impacted the performance of the CDNN system. Thus table 2 shows that use of Δ features is beneficial in lowering the WER from the CDNN acoustic model, but the benefit is constrained by the dimensionality of the resulting feature set.

As mentioned earlier, the real and simulated dev. data contained multiple conditions based on the position of the microphone and room types. In table 2 we present the averaged WER computed over all the sub-conditions present in both real and simulated data conditions. As a next step we also explored the number of filter-banks used in DOC feature processing to observe if the frequency resolution of the feature set has an impact on the performance of the feature on the dev. set. We observed that optimal number of channels to be around 40, hence all of our acoustic features presented here used 40 filterbank coefficients along with their Δ-coefficients. All the results shown in tables 1 and 2 were obtained by decoding the acoustic models using the standard bigram non-verbalized punctuation WSJ0 LM.

Table 2. WERs from different DOC systems using dev. data for decoding.

	Dev. WER	
	Simulated	Real
DOC	14.68	32.60
DOC + Δ	13.63	30.95
DOC + Δ + Δ^2	13.72	32.45

Table 3 shows the WERs from using all the different acoustic features on the simulated and real parts of the development data, which indicates that all the robust features performed significantly better than the baseline MFB feature for all given conditions. Note that the results in table 3 were obtained from using the standard 5K word WSJ0 trigram LM (with non-verbalized punctuation) for decoding. We explored combining different features and using the combined feature as input to the CDNN models. Table 3 shows the WERs from using different feature combinations, where we combined features with their Δ coefficients to other features. In general we observed that feature fusion helps in lowering the WER compared to the individual feature-based systems. We tried both two-way and three-way feature fusions and noticed that three-way feature fusion does not provide a significant gain over the two-way fusion. Note that during feature fusion we simply concatenated multiple feature streams together, hence the more features we combined the larger the dimension of the feature space became. We have not explored any dimensionality reduction techniques in our experiments, which can potentially help in achieving better WERs for feature fusions using more than two features. Table 3 shows that feature fusion helps to reduce the WER consistently for the real data compared to the single feature system, where by "single feature" we mean a specific feature type, e.g., MFB, NMC etc. Note that room 3's reverberation time was the most different from the training compared to the other two rooms, and hence room 3 in the simulated data always gave higher WERs compared to the other two rooms. Feature fusion seems to lower the WER for room 3 more compared to the other two rooms, indicating that feature fusion may be helping to reduce the mismatch between the training and testing data by providing the acoustic model with more details about the observation space.

We evaluated the single feature and the fused feature systems using the evaluation data. Table 4 shows the results from the all the

Table 3. WERs on the development set from the different feature-combination based CDNN systems.

FEATURES	WER (%)									
	simulated data							real data		
	Room 1		Room 2		Room 3		Ave.	Room 1		Ave.
	Near	Far	Near	Far	Near	Far		Near	Far	
MFB	8.1	9.5	9.5	15.1	10.4	17.5	11.7	30.5	31.1	30.8
DOC	7.2	8.5	8.0	16.5	9.1	17.4	11.1	27.1	28.5	27.8
NMC	6.4	7.9	7.9	15.6	8.9	17.0	10.6	24.9	28.8	26.8
GFC	6.4	7.8	7.5	15.0	9.0	16.9	10.4	23.9	26.9	25.4
MMeDuSA	6.0	8.2	7.6	15.5	8.9	16.3	10.4	25.5	30.6	28.0
GFC+NMC	6.3	8.0	7.7	15.5	8.6	16.6	10.4	24.3	27.5	25.9
MMeDuSA+DOC	5.9	7.6	7.9	15.0	8.9	16.8	10.3	25.8	26.6	26.2
MMeDuSA+NMC	6.0	7.5	7.3	15.8	8.6	16.3	10.2	24.8	26.9	25.8
DOC+NMC	6.4	7.8	7.7	14.8	9.2	16.6	10.4	23.3	26.3	24.8
GFC+NMC+DOC	6.3	7.7	7.4	14.8	9.6	16.0	10.3	23.8	26.1	24.9
MMeDuSA+DOC+NMC	5.7	7.0	6.9	14.5	8.6	16.2	9.8	24.6	26.3	25.4

Table 4. WERs on the evaluation set from the different feature-combination based CDNN systems.

FEATURES	WER (%)									
	simulated data							real data		
	Room 1		Room 2		Room 3		Ave.	Room 1		Ave.
	Near	Far	Near	Far	Near	Far		Near	Far	
DOC	7.00	7.20	8.30	15.00	10.50	17.20	10.87	31.70	32.40	32.05
NMC	6.60	7.10	7.80	14.30	9.90	17.30	10.50	31.00	29.80	30.40
GFC	6.30	7.60	8.20	14.50	10.10	17.40	10.68	31.70	31.50	31.60
MMeDuSA	6.80	7.40	8.00	13.90	9.80	17.80	10.62	32.10	31.00	31.55
GFC+NMC	6.50	7.80	8.10	14.40	9.50	17.00	10.55	29.40	27.90	28.65
MMeDuSA+DOC	6.40	7.10	7.70	14.40	9.20	16.10	10.15	29.60	28.80	29.20
MMeDuSA+NMC	6.40	7.30	7.90	14.30	9.50	17.20	10.43	29.50	29.40	29.45
DOC+NMC	5.90	6.80	7.70	13.90	10.10	16.40	10.13	28.30	28.80	28.55
GFC+NMC+DOC	6.40	7.50	8.10	14.30	9.50	17.00	10.47	29.80	28.20	29.00
MMeDuSA+DOC+NMC	6.40	6.90	7.60	14.40	9.30	16.20	10.13	29.60	28.50	29.05
m-way ROVER between top 4 CDNN systems	5.83	6.47	7.03	12.48	9.11	15.15	9.35	27.37	26.98	27.18
m-way ROVER between top 4 CDNN systems and top 3 GMM systems from [14]	5.40	5.96	7.04	12.83	9.04	15.44	9.29	23.34	24.49	23.92

CDNN systems explored. Similar to table 3 we witnessed no substantial improvement in performance in 3-way feature fusion compared to 2-way feature fusion; however 2-way feature fusion did show reduction in WERs compared to single feature based systems. We also explored speaker adaptation using utterance based i-vector as features. To control the dimension of the input features in fused feature setup, we used a small universal background model (UBM) with 16 Gaussian components and generated 30 dimension i-vectors from each for each utterance. We concatenated the i-vectors with the input features similar to [9, 15], and observed some improvement in performance beyond the systems without i-vectors only in single feature systems; that observation did not hold for fused feature systems. One possible explanation is that the UBM size and i-vector dimension was too small to make an impact. The strength of the systems presented here can be attributed to the feature fusion for each utterance, assuming that each utterance came from a different speaker, whichis not the case in the given data. However the individual system performance presented in this work performs as good as and sometimes better than the utterance-based batch approach, which provides a rich set of observation to the CDNN model.

Finally, we performed m-way ROVER [26] combination of the top 4 CDNN subsystems with the top 3 GMM-HMM subsystems from our REVERB2014 submission [14] and the results are shown in the last row of table 4. Note that the big gain in performance from combining GMM systems with the CDNN systems (as seen in table 4) shows the complementary nature of the GMM hypothesis, where MLLR adaptation may have been the key contributor. While the combination of the GMM systems only showed a meager less than a percent reduction in WER on the simulated data, the same for real data was almost 12% and as we know the real data was significantly mismatched from the simulated reverb data used for training, hence we can assume that the MLLR adaptation in GMM system may have been a key contributor in improving the WERs during system combination. Note that the top systems were selected based on their WER from the dev. set and the ROVER-based fusion used equal weight for all the subsystems.

The ROVER results in table 4 shows a further relative reduction of WER by 16% for real evaluation data and 8% for simulated evaluation data with respect to the best performing fused feature system (DOC+NMC). Also note that this WER surpasses the best WER reported in REVERB2014 challenge for 1-channel multi-condition training full batch processing without using any additional training data. Finally, the results show that fusion of GMM-HMM systems helped to lower the WERs further compared to the ROVER-fused CDNN system, indicating that GMM-HMM systems contain sufficient complementary information and hence can help to improve system performance further when combined with deep neural network based ASR systems.

6. CONCLUSION

In this work, we used several robust acoustic features in a CDNN based ASR setup and demonstrated that such a system showed sufficient robustness against reverberation compared to baseline MFB features. We demonstrated that the use of delta features can improve ASR performance and feature fusion can enhance the word recognition performance compared to the individual feature based systems. The gain from feature fusion is restricted by the number of features used in fusion because increased dimensionality adds a substantial burden to neural network training and, hence, can be counterproductive. We observed that a 2-way feature fusion is usually more beneficial than single feature systems. Furthermore, system-level fusion using ROVER showed substantial reduction in WER.

In future we want to explore dimensionality reduction techniques after feature fusion and explore feature engineering techniques for DNN architectures. We also aim to explore i-vector based speaker adaptation using speaker clusters to obtain the i-vectors and expect to obtain further gains beyond feature fusion.

7. ACKNOWLEDGMENT

This research was partially supported by NSF Grant # IIS-1162046.

8. REFERENCES

[1] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot and B. Raj, "The REVERB Challenge: A Common Evaluation Framework for Dereverberation and Recognition of Reverberant Speech," *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013.

[2] M. S. Brandstein and D. B. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. New York: Springer Verlag, 2001.

[3] R. Martin and P. Vary, "Combined Acoustic Echo Cancellation, Dereverberation and Noise Reduction: A Two Microphone Approach," *Journal of Annales des Télécommunications*, Vol. 49, Iss. 7–8, pp. 429–438, 1994.

[4] K. Ohta and M. Yanagida, "Single Channel Blind Dereverberation Based on Auto-Correlation Functions of Frame-Wise Time Sequences of Frequency Components," *Proc. of IWAENC*, pp. 1–4, 2006.

[5] M. Wu and D.L. Wang, "A Two-Stage Algorithm for One-Microphone Reverberant Speech Enhancement," *IEEE Trans. Aud. Speech & Lang. Process.*, Vol. 14, No. 3, pp. 774–784, 2006.

[6] L. Couvreur and C. Couvreur, "Robust Automatic Speech Recognition in Reverberant Environments by Model Selection," *Proc. of HSC*, pp. 147–150, 2001.

[7] A. Sehr and W. Kellermann, "A New Concept for Feature-Domain Dereverberation for Robust Distant-Talking ASR," *Proc. of ICASSP*, pp. 369–372, 2007.

[8] M. Delcroix and S. Watanabe, "Static and Dynamic Variance Compensation for Recognition of Reverberant Speech with Dereverberation Preprocessing," *IEEE Trans. on Aud. Speech & Lang. Process.*, Vol. 17, No. 2, pp. 324–334, 2009.

[9] Md. J. Alam, V. Gupta, P. Kenny, P. Dumouchel, "Use Of Multiple Front-Ends And I-Vector-Based Speaker Adaptation For Robust Speech Recognition," in Proc. of REVERB Challenge, 2014.

[10] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, I. Nobutaka, K. Kinoshita, M. Espi, T. Hori, T. Nakatani, "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the REVERB challenge," in Proc. of REVERB Challenge, 2014.

[11] V. Mitra, W. Wang, Y. Lei, A. Kathol, G. Sivaraman, C. Espy-Wilson, "Robust features and system fusion for reverberation-robust speech recognition," in Proc. of REVERB Challenge, 2014.

[12] 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.

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

[14] V. Mitra, W. Wang, H. Franco, Y. Lei, C. Bartels, M. Graciarena, "Evaluating robust features on Deep Neural Networks for speech recognition in noisy and channel mismatched conditions," in Proc. of Interspeech, 2014.

[15] G. Saon, H. Soltau, D. Nahamoo and M. Picheny, "Speaker Adaptation of Neural Network Acoustic Models using I-vectors," in Proc. ASRU, 2013.

[16] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, "Front-end factor analysis for speaker verification," IEEE Trans. on Speech and Audio Processing, 2011, 19, 788-798.

[17] T. Robinson, J. Fransen, D. Pye, J. Foote and S. Renals, "WSJCAM0: A British English Speech Corpus for Large Vocabulary Continuous Speech Recognition," *Proc. ICASSP*, pp. 81–84, 1995.

[18] M. Lincoln, I. McCowan, J. Vepa and H.K. Maganti, "The Multi-Channel Wall Street Journal Audio Visual Corpus (MC-WSJ-AV): Specification and Initial Experiments," proc. of *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2005.

[19] V. Mitra, H. Franco and M. Graciarena, "Damped Oscillator Cepstral Coefficients for Robust Speech Recognition," *Proc. of Interspeech*, pp. 886–890, 2013.

[20] V. Mitra, H. Franco, M. Graciarena, and A. Mandal, "Normalized Amplitude Modulation Features for Large

Vocabulary Noise-Robust Speech Recognition," *Proc. of ICASSP*, pp. 4117–4120, 2012.

[21] P. Maragos, J. Kaiser and T. Quatieri, "Energy Separation in Signal Modulations with Application to Speech Analysis," *IEEE Trans. Signal Processing*, Vol. 41, pp. 3024–3051, 1993.

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

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

[24] H. Teager, "Some Observations on Oral Air Flow During Phonation," in *IEEE Trans. ASSP*, pp. 599–601, 1980.

[25] D. B. Paul and J. M. Baker, "The Design for the Wall Street Journal-based CSR Corpus," *Proc. of HLT*, pp 3

[26] J. G. Fiscus, "A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction. (ROVER)," *Proc. of ASRU*, pp. 347–354, 1997.

[27] 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.

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