Evaluating robust features on Deep Neural Networks for speech recognition in noisy and channel mismatched conditions

Vikramjit Mitra, Wen Wang, Horacio Franco, Yun Lei, Chris Bartels, Martin Graciarena

Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, USA.

{vikramjit.mitra, wen.wang, horacio.franco, yun.lei, chris.bartels, martin.graciarena}@sri.com

Abstract

Deep Neural Network (DNN) based acoustic models have shown significant improvement over their Gaussian Mixture Model (GMM) counterparts in the last few years. While several studies exist that evaluate the performance of GMM systems under noisy and channel degraded conditions, noise robustness studies on DNN systems have been far fewer. In this work we present a study exploring both conventional DNNs and deep Convolutional Neural Networks (CNN) for noise- and channel-degraded speech recognition tasks using the Aurora4 dataset. We compare the baseline mel-filterbank energies with noise-robust features that we have proposed earlier and show that the use of robust features helps to improve the performance of DNNs or CNNs compared to mel-filterbank energies. We also show that vocal tract length normalization has a positive role in improving the performance of the robust acoustic features. Finally, we show that by combining multiple systems together we can achieve even further improvement in recognition accuracy.

Index Terms: deep neural networks, convolutional neural networks, noise-robust speech recognition, continuous speech recognition, modulation features, damped oscillators.

1. Introduction

Recent advances in neural network technology have redefined the common strategies used in acoustic modeling for automatic speech recognition (ASR) systems, where Gaussian Mixture Model (GMM)-based Hidden Markov Models (HMM) traditionally have been the state of the art. Several studies [1, 2, 3] have demonstrated a significant improvement in speech recognition performance from deep neural networks compared to their GMM-HMM counterparts.

GMM-HMM systems have traditionally been susceptible to background noise and channel distortions. For these systems, a small mismatch between training and testing conditions can make speech recognition a futile effort. To counter such degradation in performance, the speech research community made a significant effort to reduce the mismatch between training and testing conditions by processing the speech signals, either by doing speech enhancement [4, 5] or by using robust signal processing techniques [6, 7, 8, 9]. Studies have also explored introducing robustness into the acoustic models, either by introducing a wide array of noise contaminated data to those models or by implementing a reliability mask [11, 12, 13].

The emergence of Deep Neural Network (DNN) architecture has resulted in a significant boost in speech recognition performance. However, there remains the question if the traditionally known signal processing techniques used in GMM-HMM architectures are still at all relevant to this new paradigm. Given the versatility of the DNN systems, it has been stated [14] that speaker normalization techniques such as vocal tract length normalization (VTLN) [15] do not improve speech recognition accuracy significantly, as the DNN architecture's rich multiple projections through multiple hidden layers allow it to learn a speaker-invariant representation of the data. The current state-of-the-art architectures have also deviated significantly from the traditional cepstral representation to simpler spectral representations. While the basic assumptions in GMM-HMM architectures necessitated uncorrelated observations due to their widely used diagonal covariance design (which in turn forced the observation to undergo a decorrelation step using the widely popular discrete cosine transform (DCT)) the current paradigm makes no such assumption. In fact, the Neural Network architectures are known to benefit from cross-correlations [16] and hence demonstrate better performance using spectral features rather than their cepstral versions [17]. Recent studies [17, 32] have demonstrated that DNNs work very well for noisy speech and improve performance significantly compared to GMM-HMM systems.

Recently, Convolutional Neural Networks (CNNs) [18, 19] have been proposed and are often found to outperform fully connected DNN architectures [20]. CNNs are also expected to be noise-robust [18], especially in the cases where noise/distortion is localized in the spectrum. Studies [13] have shown improvement in speech recognition performance when VTLN is used on acoustic features using a deep CNN acoustic model.

In this work we show that the use of robust features can appreciably improve the performance of CNN acoustic models. We present an exhaustive study on the use of robust acoustic features as observations for CNN/DNN architectures for a noisy English continuous speech recognition task of Aurora4 [21]. We revisited VTLN in our experiments and observed some improvement in performance under noise- and channel-degraded conditions. We also compared DNN acoustic models with their CNN counterparts and observed a consistent gain from the latter. Overall, we found that robust features almost always improved speech recognition performance compared to the mel-filterbank energies.

The paper is structured as follows. First, in Section 2, we briefly describe the Aurora4 dataset, which was used in our experiments. In Section 3 we present the different feature-extraction strategies used in our work. In Section 4, we present the different acoustic models used, followed by their results and discussion in Section 5. Finally, in Section 6, we present our conclusions.

2. Data used for ASR Experiments

For the English LVCSR experiments, the Aurora4 database was used. It contains six additive noise versions with channel matched and mismatched conditions. It is created from the standard 5K Wall Street Journal (WSJ0) database and

has 7180 training utterances of approximately 15 hours total duration, and 330 test utterances, each with an average duration of 7 seconds. The acoustic data (both training and test sets) comes with two different sampling rates (8 kHz and 16 kHz). Two different training conditions were specified: (1) clean training, which is the full SI-84 WSJ train-set without any added noise; and (2) multi-condition training, with about half of the training data recorded using one microphone and the other half recorded using a different microphone (hence incorporating two different channel conditions), with different types of added noise at different SNRs. The noise types are similar to the noisy conditions in test. The Aurora4 test data includes 14 test-sets from two different channel conditions and six different added noises (in addition to the clean condition). The SNR was randomly selected between 0 and 15 dB for different utterances. The six noise types used were (1) car; (2) babble; (3) restaurant; (4) street; (5) airport and (6) train (set07) along with a clean condition. The evaluation set comprised 5K words in two different channel conditions. The original audio data for test conditions 1-7 was recorded with a Sennheiser microphone while test conditions 8-14 were recorded using a second microphone that was randomly selected from a set of 18 different microphones (more details in [21]). The different noise types were digitally added to the clean audio data to simulate noisy conditions. These 14 test sets mentioned above are typically grouped into 4 subsets: clean – matched-channel, noisy - matched-channel, clean with channel distortion, and noisy with channel distortion, which are usually referred to as test sets A, B, C, and D, respectively. A part of the clean training (893 out of 7139 utterances) and the matched channel noisy training (2676 utterances), which were not used in the multi-conditioned training set of Aurora4, were used as the held-out cross-validation set that was used to track the cross-validation error during neural network training.

3. Acoustic Features

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

3.1 Gammatone Filter Coefficients (GFCs)

The gammatone filters are a linear approximation of the auditory filterbank performed in the human ear. In GCC 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.

3.2 Damped Oscillator Coefficients (DOC)

DOC [22] aims to model the dynamics of the hair cells within the human ear. The hair cells detect the motion of incoming sound waves and excite the neurons of the auditory nerves. In DOC processing, the incoming speech signal is analyzed by a bank of gammatone filters (in this work, we used a bank of 40 gammatone filters equally spaced on the ERB scale), which splits the signal into bandlimited subband signals. 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. More details about damped oscillator processing and the DOC pipeline can be obtained in [22]. We analyzed the damped oscillator response by using a Hamming analysis 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 to yield a 40-dimensional DOC feature vector.

3.3 Normalized Modulation Coefficients (NMC)

NMC [23] is motivated by the fact that amplitude modulation (AM) of subband speech signals plays an important role in human speech perception and recognition [24, 25]. These features were obtained from tracking the amplitude modulations of subband speech signals in the 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 34 channels equally spaced on the ERB scale. The subband signals from the gammatone filterbanks were then processed using the Discrete Energy Separation algorithm (DESA) [26], which produced instantaneous estimates of AM signals. The powers of the AM signals were then root compressed using the 15^{th} root. The resulting 40-dimensional feature vector was used as the NMC feature in our experiments.

3.4 Modulation of Medium Duration Speech Amplitudes (MMeDuSA)

Like the NMCs, MMeDuSA [29] features aim to track the subband AM signals of speech, but they use a medium duration analysis window and also track the overall summary modulation. The summary modulation plays an important role in tracking speech activity as well as in locating events such as vowel prominence/stress, etc. [27]. 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 [28] 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 result was used as a 40-dimensional feature set, which we call the MMeDuSA1 feature set. Additionally, the AM signals from the subband channels were bandpass-filtered to retain the modulation information within the range 5 to 200 Hz, then summed across the frequency scale to produce a summary modulation signal. The power signal of the modulation summary was obtained, followed by 15^{th} root compression, resulting in an additional 11 coefficients that were combined with the previous 40-dimensional features to produce the 51-dimensional MMeDuSA2 features.

We also explored vocal tract length normalization (VTLN) of each of the above mentioned acoustic features.

4. Description of the ASR systems used

We used several acoustic models in this study, including traditional GMM-HMM, more recent sGMM-HMM [30], and the more popular DNN and CNN systems. The GMM-HMM is trained using the maximum likelihood criteria using 39-dimensional MFCC features (13 cepstra along with their velocity and acceleration coefficients) followed by segment level mean and variance normalization and fMLLR based

speaker adaptation. The baseline GMM-HMM system consists of context-dependent triphones with roughly 1247 senones and approximately 24K Gaussians. Additionally, we trained an sGMM-HMM system using 2639 senones and roughly 48K Gaussians.

The GMM-HMM model was used to align the training data to produce senone labels for training the DNN and CNN systems. We observed that increasing the number of senones helped to improve the recognition accuracy of the DNN/CNN systems; hence, we selected a final GMM-HMM model of 3162 senones to train our systems. However to have a fair comparison with the GMM-HMM and sGMM-HMM systems, we also trained CNN/DNNs with 1247 senones. Both the DNN and CNN systems were trained with mel-filterbank energy features (with 40 channels), which were treated as the baseline, and then with similar filterbank features using the robust processing outlined in Section 3. The input layer of the CNN/DNN systems was formed using a context window of 15 frames (7 frames on either side of the current frame) to result in 600 input nodes. We also explored different numbers of filter banks in our features and observed 40 to be the near-optimal selection.

The CNN acoustic model was trained using cross entropy on the alignments from the HMM-GMM. 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. The subsequent DNN included five hidden layers, with 1024 nodes per hidden layer, and the output layer, with 3162 nodes representing the senones.

The networks were then discriminatively 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. Experiments and Results

We trained a triphone GMM-HMM system and an sGMM-HMM system with 1247 senones and 2639 senones respectively. Both of those systems used the standard 5K non-verbalized punctuation, closed vocabulary set bigram language model (LM), with model feature space maximum likelihood linear regression (fMLLR) adaptation. We used the GMM-HMM system to generate alignments for the DNN/CNN systems. Five-layered DNN and CNN acoustic models were trained using the mel-filterbank energy features. Table 1 shows the word error rates (WERs) from test sets A, B, C and D of Aurora-4. Note that all the results reported in this paper use the multi-conditioned training data to train the acoustic models. The neural nets in Table 1 had five hidden layers with 1024 neurons in each layer.

We observed a significant reduction in WERs for all the test conditions when we moved from GMM-HMM based systems to deep neural network based systems, but unlike [17] we did not see a substantial reduction of WERs for conditions A and B, that may have happened due to their use of enhanced features based on Log-MMSE noise suppression [33, 34]. We tried increasing the number of senones, number of layers and the number of neurons in each layer for the mel-filterbank features.

Table 1. WER on multi-conditioned training data of Aurora-4 from GMM-HMM, sGMM-HMM, DNN and CNN systems

	A	B	C	D	Avg.
GMM-HMM (MFCC-39)	14.0	22.6	23.7	36.1	27.9
sGMM (MFCC-39)	13.5	20.7	22.4	33.4	25.8
DNN (mel-filterbank, 1247 senones)	10.6	16.4	15.8	26.6	20.3
CNN (mel-filterbank, 1247 senones)	**9.5**	**14.8**	**14.6**	**23.6**	**18.2**

For mel-filterbank features, approximately 3162 senones gave the best recognition performance amongst all other senone models that we have explored. Table 2 presents the results from the experiments using different layers and different number of neurons in each layer for the 3162 senone DNN and CNN model.

Table 2. WER from the DNN and CNN systems using different numbers of layers and neurons for mel-filterbank features

	A	B	C	D	Avg.
DNN (Layer 4, neuron 512)	10.8	17	16.4	26.8	20.71
DNN (Layer 4, neuron 1024)	10.8	16.2	15.6	25.2	19.63
DNN (Layer 4, neuron 2048)	10.4	16.1	15.4	25.1	19.50
DNN (Layer 4, neuron 4096)	10.4	16.0	15.7	25.4	19.61
DNN (Layer 5, neuron 1024)	10.3	15.9	15.2	25.2	19.44
DNN (Layer 5, neuron 2048)	10.3	15.8	15.8	25.2	19.44
DNN (Layer 5, neuron 4096)	10.5	16.0	15.8	25.7	19.75
DNN (Layer 6, neuron 2048)	10.2	15.7	14.9	25.3	19.36
CNN (Layer 4, neuron 1024)	10.0	14.4	13.4	21.9	17.23
CNN (Layer 4, neuron 2048)	**9.8**	**14.2**	**13.2**	**21.6**	**16.99**

Table 2 shows that for DNNs, the depth helped to improve the performance more than the width of the network. It also clearly shows that CNNs are a much better candidate for noisy speech recognition as they demonstrate a significant reduction in mismatched channel conditions compared to their DNN counterpart. It is also worth noting that a four hidden layer CNN with 1024 neurons was able to outperform a six hidden layer DNN with 2048 neurons. For CNNs, we noticed that going deeper than 4 layers did not significantly reduce the WER for CNNs. Increasing the thickness of the neurons to 2048 or more did not substantially lower the WERs either. We compared the performance of the robust features using a 5 hidden layer, 1024 neuron DNN and a 4 hidden layer, 1024 neuron CNN. Tables 3 and 4 show the results from using all the robust features explored in this paper.

Tables 3 and 4 show that the robust features did help in improving the performance of the DNN and CNN systems and the overall noise robustness trend of CNN systems is consistent with Table 3, where we saw that CNN systems always outperform DNN systems in both noisy conditions and in clean and channel-mismatched conditions. We can also observe that the robust features helped overall to reduce the

WER compared to the mel-filterbank features both in clean and noisy conditions, where DOC performed the best in the DNN systems and NMC performed the best in CNN systems.

Next we evaluated the role of VTLN in DNN/CNN performance under noisy conditions. Tables 5 and 6 show the WERs from the robust features using VTLN.

Table 3. WER on multi-conditioned training data of Aurora-4 from the 5 hidden layer DNN with 1024 neurons with different features

	A	B	C	D	Avg.
mel-filterbank	10.3	15.9	15.2	25.2	19.44
GFC	**9.3**	**13.5**	**12.9**	25.8	18.43
NMC	10.2	14.6	13.4	24.8	18.57
DOC	9.7	**13.5**	14.1	25.0	18.20
MMeDuSA1	10.1	14.2	13.3	24.7	18.34
MmeDuSA2	10.0	13.8	**12.9**	**24.1**	**17.88**

Table 4. WER on multi-conditioned training data of Aurora-4 from the 4 hidden layer CNN with 1024 neurons with different features

	A	B	C	D	Avg.
mel-filterbank	10.0	14.4	13.4	21.9	17.23
GFC	8.9	**11.9**	11.4	**21.3**	15.68
NMC	**8.4**	**11.9**	11.4	**21.3**	**15.64**
DOC	9.0	**11.9**	11.8	21.8	15.93
MMeDuSA1	9.3	12.4	**11.1**	21.5	15.99
MmeDuSA2	9.1	12.4	11.6	21.8	16.14

Table 5. WER on multi-conditioned training data of Aurora-4 from the 5 hidden layer DNN with 1024 neurons with different VTLN transformed features

	A	B	C	D	Avg.
GFC	8.8	12.6	**12.7**	24.1	17.26
NMC	**8.7**	**12.5**	12.9	23.7	17.06
DOC	9.0	12.7	13.6	22.7	16.79
MMeDuSA1	9.4	13.5	13.0	23.8	17.59
MmeDuSA2	9.1	13.1	12.8	**22.4**	**16.78**

Table 6. WER on multi-conditioned training data of Aurora-4 from the 4 hidden layer CNN with 1024 neurons with different VTLN transformed features and a 5-way ROVER combination of all 5 CNN systems

	A	B	C	D	Avg.
GFC	**8.3**	**11.4**	**11.3**	21.2	15.37
NMC	**8.3**	11.7	11.4	20.9	15.38
DOC	8.9	11.6	11.8	**20.1**	**15.06**
MMeDuSA1	8.9	12.0	11.4	20.8	15.51
MmeDuSA2	8.8	11.8	11.4	20.8	15.41
5-way ROVER	**8.0**	**10.6**	**10.6**	**19.1**	**14.06**

Tables 5 and 6 show that the performance comparisons between the robust features on DNN and CNN are very similar. The GFC almost always performs the best under clean conditions, which is anticipated, as it is not doing any noise-specific signal processing with the exception of a gammatone analysis filterbank. The GFC also performed reasonably well in almost all the experiments compared to its mel-filterbank counterpart, which shows that it is a promising candidate as a filterbank feature for DNN/CNN architectures. Comparing tables 5 and 6 with tables 3 and 4 we see that VTLN use indeed makes some difference; however, as with previous observations [14, 31], we observed no significant reduction in WER due to VTLN, where we noticed a paltry 1% absolute WER reduction from using the VTLN transformed features from DNNs and an even less WER reduction from CNNs. Hence, VTLN gave more improvement in the DNN system than the CNN system as evident from tables 3, 4, 5 and 6. We believe that owing to the localized spatial pooling of the CNNs from one layer to the next, they have more robustness against any localized frequency warps that may happen due to speaker differences, compared to their DNN counterparts.

Finally we employed a 5-way ROVER [10] combination of the GFC, NMC, DOC, MMeDuSA1 and MMeDuSA2 subsystems, where the individual subsystems were weighted equally. We observed consistent improvement in performance from system combination, where we observed almost 1% absolute reduction in WER across all the different conditions. The result of ROVER combination of the robust-feature based systems is given in Table 6. Thus from these results we can infer that the robust features by themselves gave consistent improvement in performance compared to baseline mel-filterbank energies, and given that these features are sufficiently different from one another, helped to create a diverse set of subsystems whose individual hypotheses combined well with one another to result in further improvement in recognition accuracy.

6. Conclusions

In this paper we presented results from experiments using robust features in a DNN and CNN acoustic model. We observed a consistent improvement in performance from CNN models compared to their DNN counterpart for the different test sets of the Aurora4 speech recognition task. We also witnessed that robust features help to reduce the WERs compared to the baseline mel-filterbank energies. Use of VTLN on the acoustic features was found to be beneficial, especially in the DNN setup, but we did not observe a significant reduction in WERs as typically observed from the GMM-HMM systems. We also observed that a ROVER-based CNN system combination resulted in performance improvement beyond the best individual systems, indicating that the systems provide some degree of complementary information. The experiments reported here use the raw energy coefficients from the robust feature pipeline directly and we have not explored context modeling of those features, in which typically velocity, acceleration or coefficients are appended to the static features. In future we intend to explore context modeling to see if it has any significant contribution to speech recognition performance under noisy and channel degraded conditions.

7. Acknowledgements

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

8. References

[1] 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.

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

[3] B. Kingsbury, T. N. Sainath, and H. Soltau, “Scalable minimum bayes risk training of deep neural network acoustic models using distributed hessian-free optimization,” Proc. of Interspeech, 2012.
[4] 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.
[5] 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.
[6] 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.
[7] 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.
[8] V. Tyagi, “Fepstrum features: Design and application to conversational speech recognition”, IBM Research Report, 11009, 2011.
[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, Japan, 2012.
[10] 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.
[11] B. Kingsbury, G. Saon, L. Mangu, M. Padmanabhan and R. Sarikaya, "Robust speech recognition in noisy environments: the 2001 IBM SPIN Eevaluation system", In Proc. of ICASSP, Vol.1, pp.I-53–I-56, FL, 2002.
[12] S. Fine, G. Saon, and R.A. Gopinath, "Digit recognition in noisy environments via a sequential GMM/SVM system", In Proc. of ICASSP, Vol.1, pp.I-49–I-52, FL, 2002.
[13] M. Cooke, P. Green, L. Josifovski and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data", Speech Comm., 34(3), pp.267-285, 2001.
[14] D. Yu, M. Seltzer, J. Li, J-T. Huang and Frank Seide, "Feature Learning in Deep Neural Networks - Studies on Speech Recognition Tasks", ICLR 2013.
[15] P. Zhan and A Waibel, “Vocal tract length normalization for LVCSR,” in Tech. Rep. CMU-LTI-97-150. Carnegie Mellon University, 1997
[16] V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman and L. Goldstein, "Retrieving Tract Variables from Acoustics: a comparison of different Machine Learning strategies," IEEE Journal of Selected Topics on Signal Processing, Sp. Iss. on Statistical Learning Methods for Speech and Language Processing, Vol. 4, Iss. 6, pp. 1027-1045, 2010.
[17] M. Seltzer, D. Yu, and Y. Wang, "An Investigation Of Deep Neural Networks For Noise Robust Speech Recognition", Proc of ICASSP, 2013.
[18] 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.
[19] G. Hinton, L. Deng, D. Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," IEEE Signal Proc. Mag., 29(6), pp.82-97, 2012.
[20] O. Abdel-Hamid, L. Deng and D. Yu, "Exploring Convolutional Neural Network Structures and Optimization Techniques for Speech Recognition," Proc. of Interspeech, pp. 3366-3370, 2013.
[21] G. Hirsch, “Experimental framework for the performance evaluation of speech recognition front-ends on a large vocabulary task”, ETSI STQ-Aurora DSR Working Group, June 4, 2001.
[22] V. Mitra, H. Franco and M. Graciarena, “Damped Oscillator Cepstral Coefficients for Robust Speech Recognition,” Proc. of Interspeech, pp. 886–890, 2013.
[23] 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.
[24] R. Drullman, J. M. Festen and R. Plomp, “Effect of Reducing Slow Temporal Modulations on Speech Reception,” J. Acoust. Soc. of Am., Vol. 95, No. 5, pp. 2670–2680, 1994.
[25] O. Ghitza, “On the Upper Cutoff Frequency of Auditory Critical-Band Envelope Detectors in the Context of Speech Perception,” J. Acoust. Soc. of America, vol. 110, no. 3, pp. 1628–1640, 2001.
[26] 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.
[27] 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.
[28] H. Teager, “Some Observations on Oral Air Flow During Phonation,” in IEEE Trans. ASSP, pp. 599–601, 1980.
[29] V. Mitra, H. Franco, M. Graciarena, D. Vergyri, “Medium duration modulation cepstral feature for robust speech recognition,” Proc. of ICASSP, Florence, 2014.
[30] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlıcek, Y. Qian, P. Schwarz et al., “The kaldi speech recognition toolkit,” in Proc. ASRU, 2011.
[31] T. Sainath, A. Mohamed, B. Kingsbury and B. Ramabhadran, "Deep convolutional neural network for LVCSR", Proc. of ICASSP, 2013.
[32] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," proc. of ICASSP, 2013.
[33] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean square error log-spectral amplitude estimator,” IEEE Trans. on Acoust., Speech, Signal Processing , vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.
[34] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, and A. Acero, “A Minimum-mean-square-error noise reduction algorithm on mel-frequency cepstra for robust speech recognition,” in Proc. of ICASSP, Las Vegas, NV, 2008.