Feature mapping using deep belief networks for robust speech recognition

Mojtaba Gholamipoor[1], Babak Nasersharif [2]

Receive :2016/04/20 Accepted: 2016/07/26

Abstract

Performance of automatic speech recognition (ASR) systems degrades in noisy conditions due to mismatch between training and test environments. Many methods have been proposed for reducing this mismatch in ASR systems. In recent years, deep neural networks (DNNs) have been widely used in ASR systems and also robust speech recognition and feature extraction. In this paper, we propose to use deep belief network (DBN) as a post-processing method for de-noising Mel frequency cepstral coefficients (MFCCs). In addition, we use deep belief network for extracting tandem features (posterior probability of phones occurrence) from de-noised MFCCs (obtained from previous stage) to obtain more robust and discriminative features. The final robust feature vector consists of de-noised MFCCs concatenated to mentioned tandem features. Evaluation results on Aurora2 database show that the proposed feature vector performs better than similar and conventional techniques, where it increases recognition accuracy in average by 28% in comparison to MFCCs.

Keywords: MFCC, Tandem feature, DBN, Robustness, Speech recognition

1. Introduction

Automatic speech recognition (ASR) , as defined in [1], is : ((the process of converting the speech signal into its corresponding sequence of words or other linguistic entities using algorithms implemented in a device, a computer or computer clusters)) [1]. ASR systems have a wide range of applications nowadays: voice command and control in home entertainment systems (e.g. Smart TVs), content based audio search, voice search and interacting using mobile devices (such as Siri on iPhone)[1]. Due to extension of such real-world applications, robustness of ASR systems in noisy conditions is more important than before. The performance of ASR systems degrade rapidly when there is a mismatch between training and test conditions [1]. This mismatch can be created due to contamination of speech signal with noise, speaker variations and so environment effects on speech signal [1].

Generally, robustness methods can be divided into three main categories: methods which performs in the signal level for removing noise from the speech signal (speech enhancement methods [2]), methods in the feature level which compensate noise effects on speech features and finally model adaptation methods [1].

Robustness methods in the feature level, generally are divided into two main groups. Methods of first group, change the feature extraction process to obtain more robust features such as Phase Auto-Correlation (PAC) features [3]. In the second group, a linear or non-linear transformation is applied to the feature vectors to compensate noise effects on them, such as: Cepstral mean and variance Normalization (CMVN) [2] and MVA processing [4].

One of the robust features extraction methods in noisy condition is tandem method [2, 5]. In this method, a multi-layer perceptron (MLP) is used to map traditional features to posterior probability features which have more discrimination property[5].

Recently, deep neural networks (DNNs) are widely used in speech recognition systems for acoustic modeling [6, 7] and also feature extraction and transformation [1, 8]. DNNs are artificial neural networks with multiple hidden layers between input and output layer [9]. DNNs can model complex structures and they have a high capability for modeling, learning and extracting features [10, 11, and 12].

A DNN with several layers and nonlinear functions in each layer is capable to model complex structure and discover data dependency [13]. Thus, as mentioned before, plus to acoustic modeling [6, 7], DNNs have been used for speech enhancement [14, 15] and also for feature extraction and transformation in ASR. Two types of DNNs, Deep belief network (DBN) [8, 11] and auto-encoder [16, 17], have been used for

[1] MSc in Artificial Intelligence from the School of Computer EngineeringK.N.Toosi University of Technology m.gholamipour@ee.kntu.ac.ir

[2]Assistant Professor Department of Computer Engineering, K.N.Toosi University of Technology.bnasersharif@kntu.ac.ir

dimension reduction and feature extraction from raw data.

In [14], a DNN has been used to directly learn a spectral mapping from the spectrogram of corrupted speech to desired clean speech where authors have been shown that their method leads to significant improvements of predicted speech intelligibility and quality in reverberation noisy conditions. In [15], authors used DNN for regression-based speech enhancement and they have shown that their method compared with logarithmic mean square error (MSE) achieves significant improvements for various objective quality measure.

In [12], authors compared shallow and deep neural network in feature learning and representation for speech recognition. They demonstrated that DNNs can extract more invariant and discriminant features at higher level layers. This property enables DNNs to generalize better than shallow network in mismatched conditions. Also, authors in [12] have shown that these representations become insensitive to small perturbations when network depth is increased.

In this paper, we propose to use DBN in two manners. First, we use DBN to map noisy MFCCs to clean ones. After that, we use DBN to extract tandem features from mapped MFCCs. Finally, we concatenate mapped MFCCs to extracted tandems in order to obtain our proposed discriminative and robust speech feature vector.

The remainder of this paper is organized as follows. Section 2 discusses DBN theory briefly. Section 3, introduces the tandem system framework for robust feature extraction. Section 4, explains our method for extracting robust MFCCs. Section 5 includes our experimental setup conditions. In Section 6, we report our results. Finally, we give our conclusions in Section 7.

2. Deep belief network (DBN)

The DBN is a multi-layer neural network which has a large number of neurons in each layer. The basic problem of DBN is occurred on its training phase. When free parameters of network are randomly initialized, back propagation algorithm can be trapped in a local minimum. To avoid this problem in DBN, instead of random initializing, in a unsupervised pre-training step, each pair of network layers are greedy and separately trained using restricted Boltzmann machine (RBM) [18, 19]. So, DBN is a generative model created by stacking RBMs. After pre-training step, in supervised step, back propagation algorithm is performed to train DBN for classification or estimation where an output layer is added to DBN.

As mentioned earlier, DBN has been used in the feature extraction and mapping [8, 11]. In [8], authors used DBN to reduce mismatch between far-field and near-field speech where they used robust inputs for DBN for this purpose. In [11], authors used DBN to learn a de-noising Mel filter bank where its input is noisy spectral amplitude and its output is de-noised Mel filter bank energies. In [20], authors used DBN for speech enhancement and so robust speech recognition. In this case, DBN inputs includes a central frame and its neighbors where we expect that DBN removes noise from the central frame and reconstruct it.

As mentioned previously, recently DBN was used instead of GMMs in combination with Hidden Markov model (HMM) [6, 7]. In this case, four major techniques are used for DBN training to make it robust to noise [21]. These techniques includes: multi-condition training of speech, training with enhanced features, noise aware training and dropout training [21].

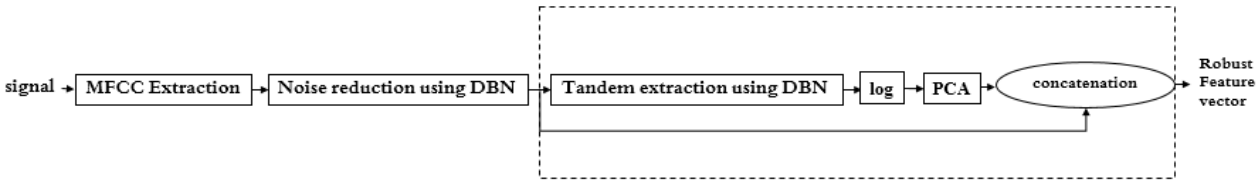


Figure 1: Proposed method structure

3. Tandem system

Most successful ASR systems is based on hidden Markov model (HMM). In HMM based systems, we can use Gaussian mixture model (GMM) or

Artificial neural network (ANN) to estimate the observation probability in HMM states. A way for combining these two systems is the tandem method [5]. In this method, in the first step, conventional features such as perceptual linear prediction (PLP) coefficients) or MFCCs are extracted from speech signal. Then, these features are mapped to posterior probabilities using a MLP where principle component analysis (PCA) is performed on mapped features for feature de-correlation [5].
In tandem method, due to DBN capability in feature mapping, MLP can be replaced by DBN [22]. Results in [22] show that DBN perform better than MLP in extracting tandem features.

4. Proposed method

Due to mentioned properties of DBN such as good training and its power in feature representation and also function approximation, we propose to use DBN for noise reduction and also feature extraction. The overall proposed method has been shown in Figure 1. We describe detail and steps of proposed method in the following subsections.

a) De-noising using DBN

In this part, we propose to use DBN for mapping noisy MFCCs to clean ones in the frame level. Thus, we use DBN as an estimator in the feature level. Figure 2 shows our used structure for DBN. This network includes two hidden layers with 512 neurons in each layer. We extract 12 MFCC and energy plus to their first and second derivatives from each frame. Then, feature vector dimension for each frame is equal to 39. The network input includes a central frame and its neighbors. In Figure 2, τ indicates the number of neighbor frames. If this parameter is equal to zero, only MFCCs of current noisy frame are input of the network.

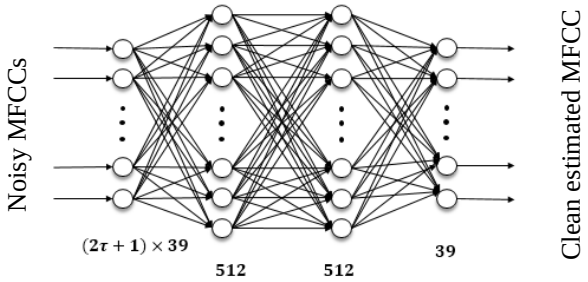


Figure 2: Structure of DBN as noise reducer
τ : the radius of neighborhood for central frame

If τ is equal to 1, in addition to MFCCs in the current frame, MFCCs in one frames after and before the current frame (total: three frames) have been considered as inputs of the network. In the present work, τ is a member of $\{0,1,2,3\}$. In this DBN, the first RBM between first and second layer is a Gaussian RBM and the other RBM between second and third layer is a Bernoulli RBM.

b) Feature extraction and mapping using DBN

The role of the DBN as tandem features extractor is mapping speech features to posterior probabilities. In [22, 23], DBN has been used to extract tandem features from PLP coefficients. In [21], the combination of two DBNs have been used to extract tandem features for phone recognition. In this work, we use DBN by two hidden layers and 512 neurons in each layer to extract tandem features from MFCCs which can be noisy or mapped MFCCs obtained from DBN in Section 3.1. Figure 3 shows our proposed structure for DBN as tandem feature extractor. The network input is similar to mentioned feature vector in previous sub-section where τ is member of {0,1,2,3}. The network outputs contain 18 posterior probabilities corresponding to number of existed phones in Aurora2 dataset.

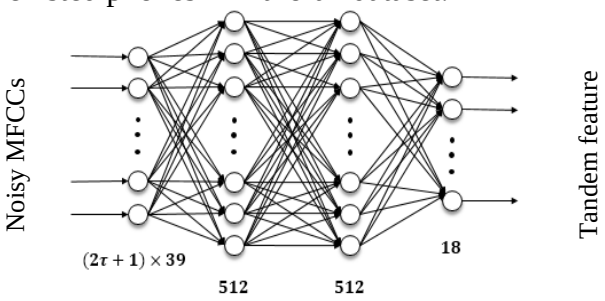


Figure 3: Structure of DBN as tandem extractor
τ : the radius of neighborhood for central frame

c) The overall proposed system

According to DBN power in reducing mismatch between training and test conditions and also its power in extracting robust tandem features, in the overall proposed system, mapped MFCCs obtained from de-nosing DBN (described in Section 3.1) have been fed as input to tandem extractor DBN in order to achieve posterior probabilities. After taking logarithm from posterior probabilities and performing principal component analysis (PCA) on them, we concatenate the obtained result to mapped MFCCs to construct final robust feature vector. We utilize stereo data to train DBN for MFCC

de-noising and to train another DBN for tandem feature extraction. So, two DBNs are separately trained. Figure 1 shows this overall process.

5. Experimental setup

We evaluate our proposed method using Aurora2 dataset [24]. Frame length and frame shift are 25 and 10 msec, respectively.The number of Mel filters is equal to 26. We used HMM with 16 states and 3 Gaussian mixtures per states. We used clean train set for HMM training. The number of features in case of MFCCs is equal to 39 (12 MFCC plus to energy and their first and second derivations) ;in case of posterior probabilities is equal to 18 and in case of concatenating both mentioned features is equal to 57.The DBN in both role of noise reducer and feature extractor has two hidden layers with 512 neurons in each layer. The number of epochs in pre-training stage is 10 epochs and in fine-tune (back propagation) is 200 epochs. The multi-layer perceptron (MLP) has one hidden layer with 512 neurons where has been trained with the same 200 epochs. Both MLP and DBN are trained using clean and noisy speech (multi condition set).

Table 1: Abbreviations for used method in the reported results

Abbreviations	Description	Number of features
NMFCC	mean-variance normalized (MVN) MFCC	39
MLP-DMFCC	De-noised MFCC using MLP	39
DBN-DMFCC	De-noised MFCC using DBN	39
MLP-NDMFCC	De-noised MFCC using MLP and normalized using MVN	39
DBN-NDMFCC	De-noised MFCC using DBN and normalized using MVN	39
MLP-TMFCC	tandem features Extracted from MFCC using MLP	18
DBN-TMFCC	tandem features Extracted from MFCC using DBN	18
DBN-TDMFCC	tandem features extracted from DBN-DMFCC	18
MLP-TMFCC+MFCC	tandem features extracted by MLP concatenated to MFCC	57
DBN-TMFCC+MFCC	tandem features extracted by DBN concatenated to MFCC	57
DBN-TDMFCC+DBN-DMFCC	DBN-TDMFCC concatenated to DBN-DMFCC	57

We used toolbox implemented in [25] for DBN. The used abbreviations in reporting result are defined in Table 1.

6. Results

a) DBN for noise reduction

Table 2 reports average of recognition accuracy on all three Aurora 2 test sets. We used two types of neural networks for mapping noisy features: DBN and MLP. It can be shown from Table 2 that DBN has better results than MLP and so learns a better mapping and estimation in comparison to MLP.

Table 2: Average of recognition accuracy on SNR 0 to 20 dB for all test sets using MLP and DBN as noise reducer

Methods	Average Accuracy
MLP-DMFCC	69.09
MLP-NDMFCC	71.67
DBN-DMFCC	**76.95**
DBN-NDMFCC	**78.15**

b) DBN for extracting tandem features

Table 3 also shows average of recognition accuracy on test sets. Both DBN and MLP are trained to extract tandem features. As can be seen from table 3, DBN has a better performance in extracting tandem features comparing with MLP.

Table 3 Average of recognition accuracy on SNR 0 to 20 dB for all test sets using MLP and DBN as tandem extractor

Methods	Average Accuracy
MLP-TMFCC	73.07
MLP-TMFCC+MFCC	78.04
DBN-TMFCC	**76.05**
DBN-TMFCC+MFCC	**78.38**

c) Selecting number of neighbor frames

Table 4 shows average of recognition accuracy for different numbers of neighbor frames. As shown in Table 4, the appropriate number of frames for DBN in the role of extracting tandem features is 3 frames ($\tau = 1$) and in the role of reducing noise is equal to 7 frames ($\tau = 3$). τ is the radius of neighborhood for the central frame defined and shown in Figure 2.

Table 4: Average of recognition accuracy on SNR 0 to 20 dB for different neighborhood radius frames for all test sets

τ	Average Accuracy	
	DBN-TMFCC	DBN-DMFCC
0	76.05	76.95
1	**76.54**	78.84
2	75.54	79.09
3	75.71	**79.69**

d) DBN for feature extraction and noise reduction

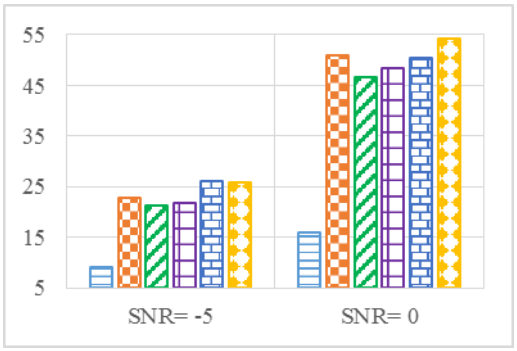
In this section, we evaluate the overall proposed system shown in Figure 1. Table 5 shows the average of recognition accuracy for each test set using this system. According to Table 4, for mapping DBN and tandem extractor DBN, we consider 7 and 3 subsequent frames in DBN inputs, respectively.

As can be seen from table 5, mapped noisy MFCC using DBN (DBN-DMFCC) has better

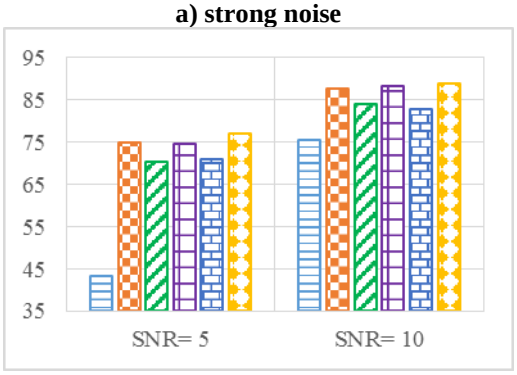
Table 5: Average of recognition accuracy on SNR 0 to 20 dB

Methods	Test sets			
	A	B	C	Average
MFCC	63.06	66.66	60.18	63.30
NMFCC	76.52	79.88	70.61	75.67
DBN-DMFCC	80.70	82.36	76.02	79.69
DBN-NDMFCC	81.38	82.40	77.39	80.39
DBN-TMFCC	77.45	78.54	73.76	76.54
DBN-TMFCC+MFCC	80.86	82.04	76.42	79.77
DBN-TDMFCC	77.47	77.63	75.45	76.85
DBN-TDMFCC+DBN-DMFCC	82.38	83.38	78.28	**81.35**

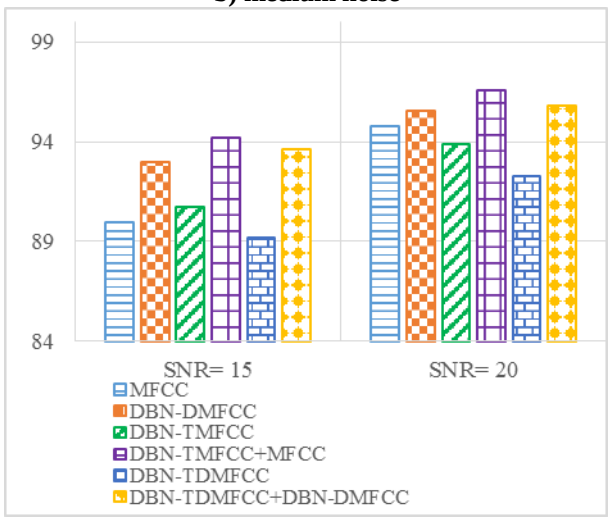
results than MFCC and NMFCC. This shows the capability of DBN in clean features estimation. Also, mean-variance normalization of DBN-DMFCC (DBN-NDMFCC) increases its recognition accuracy. As shown in table 5, tandem features extracted by DBN (DBN-TMFCC) has higher results in comparison to MFCC and NMFCC. However, the best results belong to the method in the last raw of table 5 which is the same proposed method shown in Figure 1.



a) strong noise



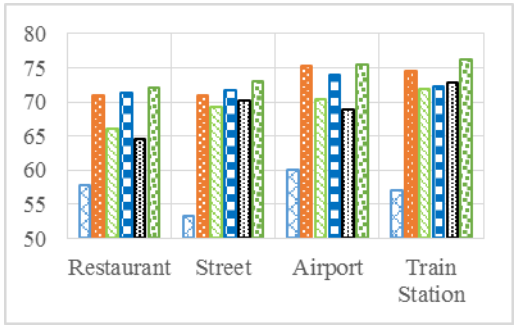
b) medium noise



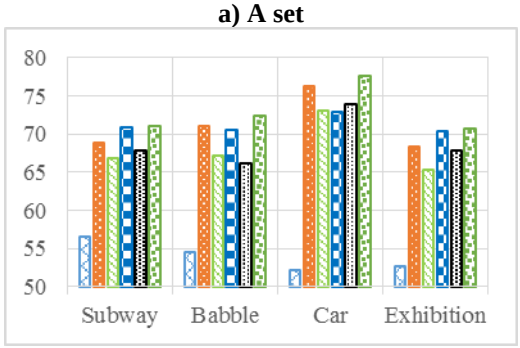
c) weak noise

Figure 4: Avrage of recognition accuracy for various SNRs

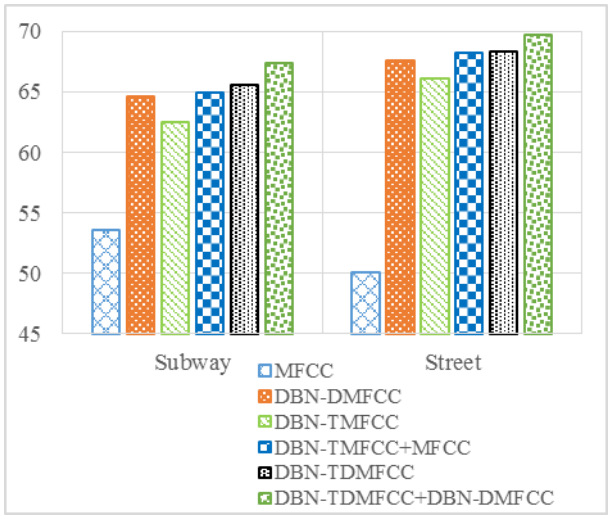
Figure 4 shows the average of recognition accuracy on different noise levels for various SNRs. As shown in the figure, when noise level is high (SNR=-5), tandem features extracted from de-noised MFCC (DBN-TDMFCC) have the highest recognition accuracy. On the other hand, when noise level is low (SNR=15, SNR=20), DBN-TMFCC+MFCC method has the highest recognition accuracy..In the other conditions, the proposed method, shown in Figure 1, works better than the other methods



a) A set



b) B set



c) C set

Figure 5: Avrage of recognition accuracy for different noise types

Figure 5 shows the average of recognition accuracy for different noise types. According to Fig. 5, for all noise types, the proposed method has the highest results among other methods.

7. Conclusions

In this paper, we propose to use DBN for noisy MFCCs mapping to clean ones and also extracting tandem features from mapped MFCCs. Furthermore, we concatenated these two mentioned groups of features to obtain the proposed robust feature vector. Results show that DBN due to its capability in nonlinear mapping and estimation, has a good performance in extracting robust and discriminative features. Thus, our proposed feature vector performs better than traditional and other similar features in noisy conditions.

8. Acknowledgements

We would like to thank Dr. Ahmad Akbari and Audio and Speech Processing Lab (ASPL) of Iran University of Science and Technology for supporting this project.

9. References

[1] J. Li, L. Deng, Y. Gong and R. Haeb-Umbach, “An Overview of Noise-Robust Automatic Speech Recognition”, Audio, Speech and Language Processing, IEEE/ACM Transactions on, Vol. 22, No. 4, pp. 745-777, 2014.

[2] P. C. Loizou, “Speech Enhancement: Theory and Practice”, 2nd edition, Boca Raton, FL, USA: CRC, 2013.

[3] S. Ikbal, H. Misra and H. Bourlard, "Phase Autocorrelation Derived Robust Speech Features", International Conference on Acoustics, Speech and Signal processing, vol. 2, pp. 133-136, 2003.

[4] C-P. Chen and J. Bilmes, "MVA processing of speech features", IEEE Transactions on Audio, Speech, and Lanuage Processing, vol. 15, no. 1, pp. 257-270, January 2007.

[5] H. Hermansky, D.P.W. Ellis and S. Sharma, “Tandem Connectionist Feature Extraction for Conventional HMM Systems”, International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol. 3, pp. 1635-1638, 2000.

[6] G. Dahl, D. Yu, L. Deng, “Context-Dependent Pre-Trained Deep Neural Networks for Large Vocabulary Speech Recognition”, IEEE Transaction on Audio, Speech, and language Processing, Vol. 20, No. 1, pp. 30–42, 2012.

[7] A. Mohamed, G.E. Dahl, G. Hinton, “Acoustic Modeling Using Deep Belief Networks”, Audio, Speech and Language Processing, IEEE Transactions on, Vol. 20, pp. 14-22, 2011.

[8] S. Chang and S. Wegmann, “On the Importance of Modeling and Robustness for Deep Neural Network Feature”, International

Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4530-4534, 2015.

[9] Y. Bengio, "Learning Deep Architectures for AI", Foundations and Trends® in Machine Learning, Vol. 2, No. 1, pp. 1-127, 2009.

[10] L. Deng, D. Yu, "Deep Learning Methods and Applications", Foundations and Trends® in Signal Processing, Vol. 7, No. 3-4, pp. 197-387, 2014.

[11] K. Han, Y. He, D. Bagchi, E. F. Lussier and D. Wang, "Deep Neural Network Based Spectral Feature Mapping for Robust Speech Recognition", Interspeech, pp. 1102-1105, 2015.

[12] D. Yu, L. Seltzer, J. Li, J. Huang, F. Seide, "Feature Learning in Deep Neural Networks – Studies on Speech Recognition Tasks", International Conference on Learning Representations (ICLR), pp. 1-9, 2013.

[13] N. Morgan, "Deep and wide: Multiple layers in Automatic Speech Recognition", IEEE Transaction on Audio, Speech, and language Processing, Vol. 20, No. 1, 2012.

[14] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, "Learning Spectral Mapping for Speech Dereverberation and Denoising", Audio, Speech and Language Processing, IEEE/ACM Transactions on, Vol. 23, No. 6, pp. 982-992, 2015.

[15] Y. Xu, J. Du, L. Dai, and C. Lee, "An experimental study on speech enhancement based on deep neural networks," IEEE Signal Processing Letters, Vol. 21, No. 1, pp. 65–68, 2014.

[16] F. Xue, Y. Zhang, and J. Glass. "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition" IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1759-1763, 2014.

[17] P. Vincent and et al. "Extracting and composing robust features with denoising autoencoders" In Proceedings of the 25th international conference on Machine learning, pp. 1096-1103. ACM, 2008.

[18] G. Hinton, S. Osindero, "A fast learning Algorithm for Deep Belief Nets", Neural Computation, Vol.18, pp. 1527–1554, 2006.

[19] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, "Greedy Layerwise Training of Deep Networks", Advanced in Neural Information Processing Systems (NIPS), Vol. 19, pp. 153-160, 2007.

[20] J. Du, Q. Wang, T. Gao, Y. Xu, L. Dai and C.H. Lee, "Robust Speech Recognition with Speech Enhanced Deep Neural Networks", Interspeech, pp. 616-620, 2014.

[21] M. L. Seltzer, D. Yu and Y. Wang, "An Investigation of Deep Neural Networks for Noise Robust Speech Recognition", International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7398-7402, 2013.

[22] O. Vinyals and S.V. Ravuri, "Comparing Multilayer Perceptron to Deep Belief Network Tandem Features for Robust ASR", International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4596-4599, 2011.

[23] X. Zheng, Z. Wu, B. Shen, H. Meng and L. Cai, "Investigation of Tandem Deep Belief Network Approach for Phoneme Recognition", International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7586-7590, 2013.

[24] H.G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy condition", ISCA ITRW ASR, 2000.

[25] M. A. Keyvanrad and M. M. Homayounpour, "A brief survey on deep belief networks and introducing a new object oriented MATLAB toolbox (DeeBNet V2.0)", ArXiv: 14083264, 2015.