A Modified Speech Enhancement Model Based on Deep Neural Networks

C. H. Lin, Bo-Wei Chen, De-Sheng Chen and Yiwen Wang

Department of Information Engineering and Computer Science
Feng Chia University,
No. 100 Wenhwa Rd., Seatwen, Taichung, Tawian

Corresponding Author: Yiwen Wang, ywang@fcu.edu.tw

Abstract - *Speech intelegibility is more essential than before due to more mobile devices need to improve speech quality. This paper develops a more efficient speech enhancement model based upon deep neural networks (DNNs). A DNN is composed by unrolling a stack of restricted Boltzmann machines (RBMs) and adding on a layer to the last stage of the multilayer perceptron (MLP). In the last layer, each neuron has a linear activation function with initial identity weights. Instead of the back-propagation, the resilient propagation learning is employed to train our DNN. The key characteristics among noise intensities, noise types, sentences and human gender, are also utilized to reduce the size of training dataset. In order to effectively control parameters of our DNN speech enhancement model, correlations between learning results of DNNs and qualities of enhanced speech are analyzed. These modifications of our DNN can speed up the learning process by 1.07~80 times that verified by NOIZEUS speech dataset.*

Keywords: Deep Neural Networks, Speech Enhancement, Resilient Propagation

1 Introduction

Speech enhancement is a continuing challenge in the field of speech and signal processing [1][2]. Many different speech enhancement algorithms under various assumptions have been proposed in the speech communication literature, which can be grouped into three categories: filtering [3][4] , spectral restoration [5][6], and speech model techniques [7][8][9].

Speech enhancements based on filtering and spectral restoration methods remain popular due to their reasonable performance and low computational complexity. The filtering speech enhancement techniques need strong assumptions about the interference between the speech and noisy signals, that limit their applications in dealing with a general acoustic background. The spectral restoration techniques are based on composite source modeling to model each source using a set of Gaussian sub-sources and a soft mask filter is derived using various estimation for separating the sources. Rather than modeling each individual source, the relationships between sources should be learned using a discriminant method. The various model-based approaches proposed so far use a parsimonious nonlinear physical model to describe the environmental distortion, and the clean speech is estimated from the noisy observation without any prior information on the noise type or speaker identity. However, the main difficulty of most of these methods is estimation of the noise power spectral density (PSD).

In contrast to traditional signal processing based approaches, neural network based methods build probabilistic or deterministic interaction models of speech and/or noise using stereo training targets and show promising results in challenging real-world speech applications. The basic idea of using neural nets for speech enhancement is to build a Gaussian mixture model (GMM) for the joint distribution of the clean and noisy speech by using DNNs[11] and then to learn the mapping function between clean speech and noisy speech.

The recent progress of deep neural network (DNN) for stereo mapping in speech recognition area can be found in][9]10][12][13][14]. G. Wang proposed to use DNN as the log posterior probabilities to form a logistic regression context-dependent to solve the data sparse problem [10]. The definition of a deep neural network (DNN) can be found in Wikipedia or [11]. The contrastive divergence (CD) algorithm is normally used to learn RBM weights, that provides an approximation to the maximum likelihood method; see [12] for more details.

The proposed speech enhancement model is based on Y. Xu's work [15] to map noisy speech into cleaner speech by using auto encoders to extract the best representations or features, using RBM to obtain generatively pre-training model, and using DNN to continuously fine-tuning the system automatically. However, each neuron in the last layer is modified as a linear activation function with initial identity weights. In the DNN learning process, the resilient

This research was supported by MOST grant 105-2221-E-035-083 -

propagation learning is employed to train our DNN instead of Xu's back-propagation learning algorithm.

2 Methodology

2.1 Proposed Speech Enhancement Model

Fig. 1 shows the block diagram of the proposed speech enhancement model. A short-time Fourier analysis is applied to the input noisy signal, Y^t, by computing the DFT of each overlapping windowed frame to obtain the phase information, $\angle Y^f$, as well as its log-power spectra, Y^l. After we obtain the estimation of the log-power spectra of clean speech, X^l, from DNN module, an overlap-add method is utilized with noisy signal phase, $\angle Y^f$, to reconstruct the waveform of the estimated clean signal. A detailed description of the feature extraction module and the waveform reconstruction module can be found in [16].

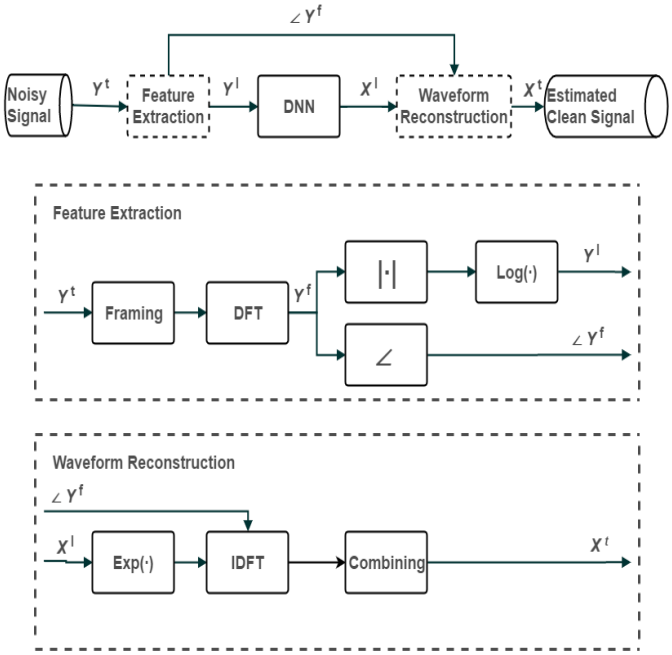

Fig. 1. The proposed speech enhancement model.

2.2 Deep Neural Network Model

A special case of DNN as illustrated in Fig. 2, where the left part is a stacking three RBMs for DNN unsupervised pre-training to obtain the noisy features and the right part is the three RBMs cascaded with the unrolled RBMs that can be considered as a MLP for DNN supervised training.

2.2.1 Pretraining DNNs with Noisy Features

The left part of Fig. 2 is a cascading of multiple RBMs to show the noisy feature pre-training. The first RBM has one visible layer of noisy signals which is connected to a hidden layer. The contrastive divergence (CD) is employed to train the weights of each RBM.

2.2.2 Training DNNs with Noisy Features and Clean Features

The right part of Fig. 2 is a MLP consisting of the three RBMs in the pretraining phase cascaded with the unrolled RBMs and plused a set of linear activation function neurons as the last layer. The initial weights of RBM are obtained from the results of the contrastive divergence training of RBMs. The initial weights of unrolled RBM are transpose of the weights of the previous RBMs. The initial weights of the last layer is a unit matrix to preserve the effectiveness results of contrastive divergence training. Becide the neurons of ipunt and output layers have linear activation functions, all other neurons own logistic activation functions.

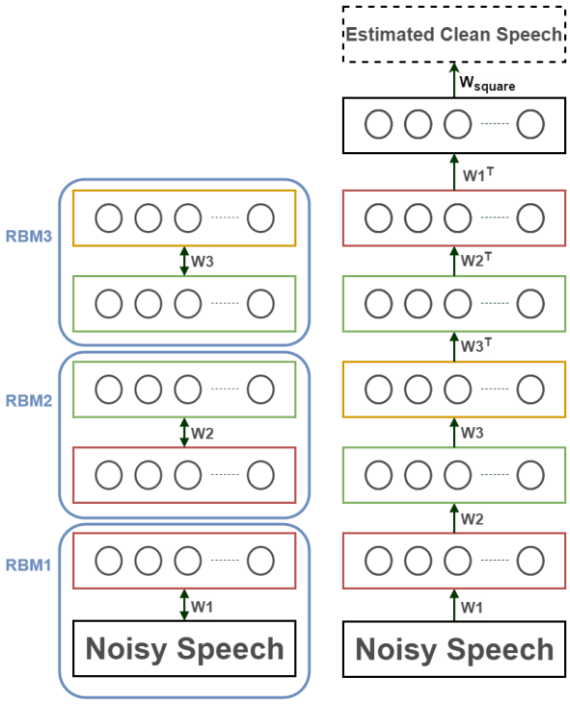

Fig. 2. The proposed DNN model.

In the whole DNN learning process, instead of the back-propagation learning, the resilient propagation learning, RPROP [17] is employed to train our DNN with minimum mean squared error (MMSE) loss function in the log domain which is more consistent with the human auditory system. The RPROP performs a local adaptation of the weight-updates according to the behaviour of the MMSE to overcome the inherent disadvantages of pure gradient-descent. Our DNN learning process is only dependent on the temporal behaviour of the derivative sign rather than the size of the derivative.

In RPROP, each weight is updated as follows :

$$w_{ij}^{t+1} = w_{ij}^{t} + \Delta w_{ij}^{t}$$

$$\text{where } \Delta w_{ij}^{t} = \begin{cases} -\Delta_{ij}^{t}, & if \ \frac{\partial MMSE^{t}}{\partial w_{ij}} > 0 \\ -\Delta_{ij}^{t}, & if \ \frac{\partial MMSE^{t}}{\partial w_{ij}} < 0 \\ 0, & else \end{cases} \text{ based on}$$

$$\Delta_{ij}^{t} = \begin{cases} \eta^{+} * \Delta_{ij}^{t-1} \ , \ if \ \dfrac{\partial MMSE^{t-1}}{\partial w_{ij}} * \dfrac{\partial MMSE^{t}}{\partial w_{ij}} > 0 \\ \eta^{-} * \Delta_{ij}^{t-1} \ , \ if \ \dfrac{\partial MMSE^{t-1}}{\partial w_{ij}} * \dfrac{\partial MMSE^{t}}{\partial w_{ij}} < 0 \\ \Delta_{ij}^{t-1} \ , \ else \end{cases}$$

According to the backtracking condition:

$$If \ \frac{\partial MMSE^{t-1}}{\partial w_{ij}} * \frac{\partial MMSE^{t}}{\partial w_{ij}} < 0, \ \Delta w_{ij}^{t} = -\Delta w_{ij}^{t-1}$$

3 Experimental Result and Analysis

The above modifications of our DNN can speed up the learning process by 1.07~80 times that verified with NOIZEUS dataset [18] for clean speech and AURORA2 database [19] for noisy environments.

3.1 Experimental setup

The clean samples from NOIZEUS data set which has 3 individual male voices and 3 individual femal voices were added with eight types of noise (*Airport, Babble, Car, Exhibition, Restaurant, Station, Street and Train*) and four levels of SNR, at 15 dB, 10 dB, 5 dB, 0 dB from AURORA2 database [19] to build a multi-condition training set. As for signal analysis, speech waveform was down-sampled to 8KHz, and the corresponding frame length was set to 32 msec. There are 2816 neurons for both input and output layers and 352 neurons for a RBM hidden layer.

The number of epoch for each layer of RBM pre-training was 500. Learning rate of DNN MLP was fixed at 0.01. The stop criterion is set the MMSE to be less than 0.1. We have implemented our DNN speech enhancement model in C++ running on a PC with Intel(R) Core(TM) i7-4710HQ CPU @ 2.5GHZ, on 64bit Windows 8. Table 1. shows the RPROP is about 83 times faster than BP.

Gradient Decent Algorithm	BP	RPROP
MLP's Epochs	2491	30
MLP's Learning Time	327.9376 minutes	3.95817 minutes

Table 1. Epochs and run-time for RPROP and BP.

Table 2. Shows the initial weights use the identity matrix is about 3 times faster than random matriices.

Gradient Decent Algorithm	Random Matrix	Identity Matrix
MLP's Epochs	92	30

Table 2. Epochs and run-time for RPROP and BP.

We evaluate the quality of tested signals using the value of Robust_Test$_{Ts}$, which is defined as follows:

$$\text{Robust\_Test}_{Ts} \triangleq \frac{\sum_{X\in(TRAIN-\{Ts\})} Value_X}{||(TRAIN-\{Ts\})||} - ORG_{Ts}, \ Ts \in TEST.$$

TEST is a set of all signals to be tested when *TRAIN* is a set of all signals to be trained. *Value* is *MMS*E after DNN is trained when ORG_X is the original *MMS*E of the *X* signal.

Similarly we evaluate the quality of trained signals using the value of Robust_Train$_{Tr}$, which is defined as follows:

$$\text{Robust\_Train}_{Tr} \triangleq \frac{\sum_{X\in(TEST-\{Tr\})}(Value_X - ORG_X)}{||(TEST-\{Tr\})||}, \ Tr \in TRAIN.$$

The average value of all tested signals under a give set of trained signals will be :

$$\text{Average}_{\ all} \triangleq \frac{\sum_{X\in TEST} Value_X}{||TEST||}.$$

We will use "*no*" in the rest of tables to represent the ORG_X, for the corresponding *X* signal. "*notes*" in the rest of tables are used to illustrate the rank of corresponding Robust_ values. The relative quality of the signals in the corresponding signal set is better when the value of Robust_ is smaller.

3.2 Corelation of Noise Intensities for DNNs

Firstly we investigate the correlations among noise intensities that affect our DNN learning quality. In this experiment, *TEST* is a set of signals using one male voice additive *airport* noise with four levels of SNR, at 15 dB, 10 dB, 5 dB, 0 dB. Table 3 uses conditions *TRAIN*={a single signal in *TEST* } to learn only one noise intensity but to test all four noise intensity levels.

Tested NI / Trained NI	0dB	5dB	10dB	15dB	Robust $_{trained\ info}$	Average $_{all}$	Notes	
0dB	0.01	27.94	24.01	33.66	12.10	21.41	4	2
5dB	40.36	0.01	22.51	27.00	0.06	22.47	2	3
10dB	37.38	28.36	0.01	14.19	-9.21	19.98	1	1
15dB	55.93	44.33	22.19	0.01	1.02	30.61	3	4
no	72.68	32.28	14.43	2.59		30.50	from noise only	
Robust $_{tested\ info}$	-28.13	1.26	8.47	22.36				
Notes	1	2	3	4				

Table 3. DNN quality after only one noise intensity learned.

The results of Table 3 show that if only one noise intensity can be chosen, the 10dB will be the best to learn the characteristics of the noise environment. Table 4 uses conditions *TEST*={all four intensity levels of one male voice additive *airport* noise} and *TRAIN*={3 different noise intensity levels from TEST} to learn 3 different noise intensity levels but to test all four noise intensity levels.

Tested NI / Trained NI	0dB	5dB	10dB	15dB	Robust $_{trained\ info}$	Average $_{all}$	Notes	
no 0dB	23.262	0.005	0.008	0.007	-49.42	5.82	1	4
no 5dB	0.006	19.371	0.010	0.010	-12.91	4.85	2	3
no 10dB	0.009	0.012	8.377	0.007	-6.06	2.10	3	1
no 15dB	0.010	0.008	0.009	12.491	9.90	3.13	4	2
no	72.683	32.280	14.433	2.593		30.50	from noise only	
Robust $_{tested\ info}$	-49.42	-12.91	-6.06	9.90				
Notes	1	2	3	4				

Table 4. DNN quality after three noise intensities learned.

The results of Table 4 show that if multiple noise intensities can be chosen, the 0dB and 15dB will be the best to

learn the characteristics of the noise. The 0dB signal helps the 15dB signal to distinguish the similarity of the male voice and discrepancy of the noise.

3.3 Corelation of Noise Types for DNNs

Secondly we investigate the correlations among noise types that affect our DNN learning quality. In this experiment, *TEST* is a set of signals using one male voice additive eight noise types using SNR levels at 10dB and 0dB. Table 5 uses conditions *TRAIN*={a single noise type in *TEST*} to learn only one noise type but to test all eight noise types.

Tested NT / Trained NT	Airport	Babble	Car	Exhibition	Restaurant	Station	Street	Train	Robust $_{tested\ info}$	Average $_{all}$	Notes	
Airport	0.01	19.61	23.16	28.83	18.18	16.70	23.69	28.69	-28.80	19.86	4	6
Babble	19.82	0.01	20.36	22.72	15.38	18.29	23.25	26.29	-30.47	18.26	2	3
Car	24.77	21.63	0.01	25.28	22.09	20.31	24.65	26.34	-25.33	20.63	7	7
Exibition	23.65	21.20	16.88	0.01	20.77	19.23	19.00	19.82	-27.95	17.57	5	2
Restaurant	16.08	17.54	17.66	22.10	0.01	15.60	24.54	25.06	-32.12	17.32	1	1
Station	19.42	21.74	19.19	25.22	19.54	0.01	25.38	21.23	-29.86	18.97	3	5
Street	22.95	20.77	21.34	23.90	21.91	19.48	0.01	20.91	-25.95	18.91	6	4
Train	27.28	32.98	29.35	26.04	33.81	27.59	28.77	0.01	-16.45	26.98	8	8
no	37.64	38.69	55.73	61.87	34.64	37.31	65.13	67.07		49.76	from noise only	
Robust $_{tested\ info}$	-14.21	-16.48	-34.59	-37.00	-12.97	-17.71	-40.95	-43.02				
Notes	7	6	4	3	8	5	2	1				

Table 5. DNN quality after only one noise type learned.

The results of Table 5 show that if only one noise type can be chosen, the *Restaurant* noise type will be the best to learn the characteristics of the noise environment. Table 6 uses conditions, *TEST*={all eight noise types of one male voice using SNR levels at 10dB and 0dB} and *TRAIN*={7 different noise types from TEST}, to learn 7 different noise types but to test all eight noise types..

Tested NT / Trained NT	Airport	Babble	Car	Exhibition	Restaurant	Station	Street	Train	Robust $_{tested\ info}$	Average $_{all}$	Notes	
No Airport	13.741	0.012	0.008	0.005	0.012	0.009	0.009	0.005	-23.90	1.73	8	7
No Babble	0.011	11.554	0.012	0.008	0.014	0.010	0.008	0.005	-27.13	1.45	5	3
No Car	0.010	0.011	13.042	0.008	0.011	0.007	0.010	0.006	-42.69	1.64	4	5
No Exibition	0.009	0.011	0.007	14.295	0.011	0.007	0.006	0.008	-47.58	1.79	3	8
No Restaurant	0.011	0.007	0.012	0.008	9.782	0.007	0.013	0.006	-24.86	1.23	7	1
No Station	0.011	0.008	0.009	0.005	0.009	11.529	0.009	0.005	-25.78	1.45	6	2
No Street	0.019	0.008	0.007	0.007	0.013	0.007	13.547	0.007	-51.61	1.70	2	6
No Train	0.010	0.009	0.009	0.006	0.009	0.008	0.015	12.452	-54.62	1.56	1	4
no	37.638	38.686	55.728	61.873	34.642	37.307	65.132	67.069		49.76	from noise only	
Robust $_{tested\ info}$	-23.90	-27.13	-42.69	-47.58	-24.86	-25.78	-51.61	-54.62				
Notes	8	5	4	3	7	6	2	1				

Table 6. DNN quality after seven noise types learned.

The results of Table 6 show that if multiple noise types can be chosen, the *Restaurant* and *Exhibition* types will be the best to learn the characteristics of the noise types. These two noise types have inherent human voices that help to distinguish the similarity of the male voice and discrepancy of the noise.

3.4 Corelation of Phonetic Alphabets for DNNs

In this section, we investigate the correlations among phonetic alphabets that affect our DNN learning quality. In this experiment, *TEST* is a set of five sentences using one female voice additive *airport* noise with two levels of SNR, at 10dB and 0dB. Table 7 depicts the American phonetic alphabets corresponding to these five sentences.

Sentence Info / Sentence Order	Text	American Phonetic Alphabet
1	The stray cat gave birth to kittens.	[ðə][stre][kæt][ɡev][bɜθ][tu][ˈkɪtn]
2	The lazy cow lay in the cool grass.	[ðə][ˈlezi][kaʊ][le][ɪn][ðə][kul][ɡræs]
3	The friendly gang left the drug store.	[ðə][ˈfrɛndli][ɡæŋ][left][ðə][drʌɡ][stɔr]
4	We talked of the sideshow in the circus	[wi][tɔkt][əv][ðə][ˈsaɪdʃoʊ][ɪn][ðə][ˈsɝkəs]
5	The set of china hit the floor with a crash.	[ðə][set][əv][ˈtʃaɪnə][hɪt][ðə][flɔr][wɪð][ə][kræʃ]

Table 7. The phonetic alphabets of five sentences to be tested.

Table 8 uses conditions to learn only one sentence type but to test all five sentences when *TRAIN*={a single sentence in *TEST*}.

Tested S / Trained S	1	2	3	4	5	Robust $_{trained\ info}$	Average $_{all}$	Notes	
1	0.01	60.78	57.28	65.27	74.24	18.92	51.52	1	1
2	70.47	0.01	61.54	74.70	66.42	19.88	54.63	2	2
3	79.88	73.62	0.01	85.06	89.84	33.88	65.68	5	5
4	72.12	75.37	66.39	0.01	69.08	26.54	56.59	4	3
5	72.01	67.05	78.50	74.93	0.01	25.57	58.50	3	4
no	51.97	40.25	40.96	57.03	43.63		46.77	from noise only	
Robust $_{tested\ info}$	21.65	28.96	24.96	17.96	31.27				
Notes	2	4	3	1	5				

Table 8. DNN quality after only one sentence learned.

The results of Table 8 show that if only one sentence can be chosen, the second sentence may be the better but not the best. Table 9 uses conditions to learn 3 different sentences but to test all five sentences, where *TEST*={all five sentences of one female voice additive *airport* noise using SNR levels at 10dB and 0dB} and *TRAIN*={3 different sentences from TEST}.

Tested S / Trained S	1	2	3	4	5	Robust $_{trained\ info}$	Average $_{all}$	Notes	
1~3	0.01	0.01	0.01	60.39	70.00	14.86	26.08	2	3
2~4	62.13	0.01	0.01	0.01	64.92	15.72	25.42	3	2
3~5	64.30	53.71	0.01	0.01	0.01	12.89	23.61	1	1
no	51.97	40.25	40.96	57.03	43.63		46.77	from noise only	
Robust $_{tested\ info}$	11.24	13.46		3.36	23.83				
Notes									

Table 9. DNN quality after three sentences learned.

According to the results of Table 8 and 9, we can not find any common rules for the corelation among these limited amount of phonetic alphabets.

3.5 Co-relations of Speakers for DNNs

Final we investigate the correlations among speakers that affect our DNN learning quality. In this experiment, *TEST* is a set of signals using three male 5-sentence voices and three female 5-sentence voices additive *airport* noise with SNR levels at 10dB and 0dB. Table 10 uses conditions *TRAIN*={a single speaker in *TEST*} to learn only one speaker but to test all six speakers.

Tested People / Trained People	CH	DE	JE	KI	SI	TI	Robust $_{trained\ info}$	Average $_{all}$	Notes	
CH	0.01	57.68	71.44	77.22	51.43	69.62	25.59	54.57	2	2
DE	56.78	0.01	71.22	76.96	55.33	69.54	26.18	54.97	3	3
JE	63.73	65.37	0.01	71.50	57.43	56.23	23.42	52.38	1	1
KI	65.11	72.54	66.20	0.01	65.44	67.89	28.45	56.20	5	4
SI	58.32	64.10	75.93	88.79	0.01	75.65	29.78	60.46	6	6
TI	66.56	71.26	62.32	76.75	63.21	0.01	26.21	56.69	4	5
no	42.45	42.98	44.72	46.99	31.91	32.85		40.32	from noise only	
Robust $_{tested\ info}$	19.65	23.21	24.71	31.25	26.65	34.93				
Notes	1	2	3	5	4	6				

Table 10. DNN quality after only one speaker learned.

According Table 10, the DNN learning results of a single speaker may not be beneficial for other speakers. Table 11 uses conditions, *TEST*={ three male 5-sentence voices and three female 5-sentence voices additive *airport* noise with SNR levels at 10dB and 0dB} and *TRAIN*={two different speakers from TEST}, to learn two different speakers but to test all speakers.

Tested People / Trained People	CH	DE	JE	KI	SI	TI	Robust $_{trained\ info}$	Average $_{all}$	Notes	
CH, DE	0.01	0.01	66.22	73.57	48.26	66.85	24.61	42.49	3	3
JE, KI	57.99	59.80	0.01	0.01	54.89	51.43	18.48	37.35	2	1
SI, TI	54.60	59.50	58.57	72.71	0.01	0.01	17.06	40.90	1	2
no	42.45	42.98	44.72	46.99	31.91	32.85		40.32	from noise only	
Robust $_{tested\ info}$	13.84	16.67	17.68	26.15	19.66	26.28				
Notes	1	2	3	5	4	6				

Table 11. DNN quality after two speakers learned.

The results of Table 11 show that if multiple speakers can be chosen, the different gender speakers will be the better to learn the characteristics of the noise environments. All above experiments show that the more training data may obtain the better quality of DNN learning results, but preprocessing of training data according to the key characteristics among noise intensities, noise types, sentences and human gender can reduce the size of training dataset to achieve comparable or even better learning results..

4 Conclusions

This paper presented an efficient speech enhancement model to speed up the DNN learning by 1.07~80 times verified using NOIZEUS speech dataset. The key characteristics among noise intensities, noise types, sentences and human gender, are analyzed and extracted to reduce the size of training dataset. The future work of this paper will theoretically investigate the co-relations between magnitude and phase of noise features as well as the appropriate loss functions for the DNN last MLP layer so that our speech enhancement model can achieve the better real-time noise reduction in various noise environments.

5 References

[1] M. Berouti, R. Schwartz, and J. Makhoul, “Enhancement of speech corrupted by additive noise,” in Proc. ICASSP, 1979, pp. 208–211.

[2] P. C. Loizou, Speech Enhancement: Theory and Practice, Boca Raton, FL, USA: CRC, 2007.

[3] J. Hao, H. Attias, S. Nagarajan, T.-W. Lee and T.J. Sejnowski, “Speech enhancement, gain, and noise spectrum adaptation using approximate Bayesian estimation,” IEEE Trans. on Audio, Speech and Language Processing, vol. 17, no. 1, pp. 24-37, 2009.

[4] S. Nemala, K. Patil, andM. Elhilali, “A multistream feature framework based on bandpass modulation filtering for robust speech recognition,” IEEE Trans. Audio, Speech, Lang. Process., vol. 21, no. 2, pp. 416–426, Feb. 2013.

[5] S. Ganapathy, S.H. Mallidi, and H. Hermansky, "Robust Feature Extraction Using Modulation Filtering of Autoregressive Models", Audio, Speech, and Language Processing, IEEE/ACM Transactions on, On page(s): 1285 - 1295 Volume: 22, Issue: 8, Aug. 2014.

[6] J. Taghia, J. Taghia, N. Mohammadiha, J. Sang, V. Bouse, and R. Martin, “An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments,” in Proc. ICASSP, 2011, pp. 4640–4643.

[7] B. Li, Y. Tsao, and K. C. Sim, “An investigation of spectral restoration algorithms for deep neural networks based noise robust speech recognition,” in Proc. Interspeech, 2013, pp. 3002–3006.

[8] G. Hinton, L. Deng, D. Yu, G. Dahl, A.Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition,” IEEE Signal Process. Mag., vol. 29, no. 6, pp. 82–97, Nov. 2012.

[9] A. Mohamed, G. Dahl, and G. Hinton, “Acoustic modeling using deep belief networks,” IEEE Trans. Audio Speech Lang. Processing, vol. 20, no. 1, pp. 14–22,Jan. 2012.

[10] G. Wang, K. Sim "Regression-Based Context-Dependent Modeling of Deep Neural Networks for Speech Recognition", Audio, Speech, and Language Processing, IEEE/ACM Transactions on, On page(s): 1660 - 1669 Volume: 22, Issue: 11, Nov. 2014

[11] J. Schmidhuberv, “Deep Learning in Neural Networks: An Overviews,” Technical Report IDSIA-03-14 / arXiv:1404.7828 v4 [cs.NE] http://arxiv.org/pdf/1404.7828v4

[12] G. Dahl, D. Yu, L. Deng, and A. Acero, “Context-dependent pre-trained deep neural networks for large vocabulary speech recognition,” IEEE Trans. on Audio, Speech and Language Processing, Vol. 20, No. 1, pp.30-42, 2012.

[13] Li Deng; Xiao Li "Machine Learning Paradigms for Speech Recognition: An Overview", Audio, Speech, and Language Processing, IEEE Transactions on, On page(s): 1060 - 1089 Volume: 21, Issue: 5, May 2013

[14] J.T. Geiger, F. Weninger, J.F Gemmeke, M. Wollmer, B. Schuller and G. Rigoll, "Memory-Enhanced Neural Networks and NMF for Robust ASR", Audio, Speech, and Language Processing, IEEE/ACM Transactions on, On page(s): 1037 - 1046 Volume: 22, Issue: 6, June 2014

[15] Y. Xu, J. Du, L. Dai and C. Lee "An Experimental Study on Speech Enhancement Based on Deep Neural Networks", Signal Processing Letters, IEEE, On page(s): 65 - 68 Volume: 21, Issue: 1, Jan. 2014.

[16] J. Du and Q. Huo, “A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions,” in Proc. Interspeech, 2008, pp. 569–572.

[17] Martin Riedmiller and Heinrich Braun. “A direct adaptive method for faster backpropagation learning: The RPROP algorithm,” Proceedings of the IEEE International Conference on Neural Networks, San Francisco, CA, April 1993.

[18] IEEE Subcommittee, “IEEE Recommended Practice for Speech Quality Measurements,” IEEE Trans. Audio and Electroacoustics, AU-17(3), 225-246, 1969.

[19] H. Hirsch, and D. Pearce. “The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions,” ISCA ITRW ASR2000, Paris, France, September 18-20, 2000.