# Speech Enhancement Using Deep Neural Network

**Pallavi D. Bhamre[1], Hemangi H. Kulkarni[2]**

*[1] Post-graduate Student, Department of Electronics and Telecommunication, R. H. Sapat College of Engineering, Management studies and Research, Nashik , Maharashtra, India*
*[2]Asst. Professor, Department of Electronics and Telecommunication, R. H. Sapat College of Engineering, Management studies and Research, Nashik, Maharashtra, India*

---------------------------------------------------------------------***---------------------------------------------------------------------

**Abstract -** *Speech is the main source of human interaction. The quality and intelligibility of speech signals during communication are generally corrupted by the surrounding noise. Corrupted speech signals therefore to be enhanced to improve quality and intelligibility. In the field of speech processing, much effort has been devoted to develop speech enhancement techniques in order to restore the speech signal by reducing the amount of disturbing noise. Speech enhancement deals with improving the quality and intelligibility of speech degraded in the presence of background noise. In various acoustic environments, especially at low Signal-to-Noise ratios (SNR), the goal of speech enhancement methods is to improving the quality and intelligibility of speech. Regarding intelligibility, different machine learning methods that aim to estimate an ideal binary mask have revealed promising results. This project covers the work of speech enhancement by use of the supervised method Deep Neural Network (DNN). In contrast to the different noise reduction techniques such as MMSE, the supervised method enhance the speech by finding a mapping function between the noisy and clean speech signals based on deep neural networks.*

***Key Words***: **Deep Neural Networks, Dropout, Global Variance equalization, Noise aware training, Speech enhancement.**

## 1.INTRODUCTION

In today's technological era speech is the most important way of communication. Speech is a source through which human can talk.

Now what is Speech Enhancement? Enhancement means the improvement in the value or quality of something. When this applied to speech, this simply means improvement in intelligibility or quality of the degraded speech signal by using signal processing tools.

Speech Enhancement is difficult problem for two reasons. First, the nature and characteristics of the noise signals can change dramatically in time and application to application. Second, the performance measurement is also defined differently for different applications.

In recent years, many researches had being carried out in the field of Speech processing. The main problem with any Speech processing experiment is the background noise. Clean speech is corrupted with the background noisy environment. So the main goal of Speech Enhancement is to improve the quality of noise which gets corrupted by the noisy background environment so that the clean speech can be recognized. The main goal of Speech Enhancement is to improve the intelligibility and quality of the noisy speech signal which gets degraded in adverse conditions. Also, the performance of the Speech Enhancement in real acoustic environment is not always satisfactory. Speech Enhancement has wide number of applications such as mobile communication, hearing aid devices and Speech recognition system.

Different Speech Enhancement techniques are developed in the recent years such as spectral subtraction, Minimum mean square error (MMSE), Log MMSE, Optimally modified log spectral amplitude (OM-LSA), Wiener filtering, etc. The improvement in the Speech Signal-to-noise ratio (SNR) is the main goal of these Speech Enhancement techniques.

Techniques are proposed for enhancement and bandwidth compression of speech degraded by additive background noise. Relationship between these systems, speech enhancement system has been successful in the context of bandwidth compression in presence of background noise leads to higher intelligibility [1]. Importance of short-time spectral Amplitude (STSA) of speech signal in its perception, the system utilizes minimum mean square error (MMSE) STSA estimator which is based on modeling speech and noise spectral components and then compared with other widely used systems based on wiener filtering and spectral subtraction algorithm [2]. STSA estimator minimizes the mean square error of the log-spectra and examines it in enhancing noisy speech and it compared with minimum mean square error STSA estimator which is very effective in enhancing noisy speech which improves its quality. Also, it lowers the residual noise level without affecting the speech itself [3]. Noise suppression technique has been for the restoration of degraded audio recording. Musical noise also eliminated without bringing distortion to the recorded signal. Non-linear smoothing is used to obtain a more consistent estimate of SNR. Attenuation function avoids the appearance of musical noise [4]. High level spectral parameters, such as mel-cepstra or line spectral pairs are taken as the features for hidden markov model (HMM) based parametric speech synthesis is improved by first, distributions of low-level, untransformed spectral envelops are used as parameters for synthesis. Second, instead of using single Gaussian distribution, here graphical models with multiple hidden variables, including restricted Boltzmann machine (RBM) and deep belief networks (DBN). Here focus on the acoustic modeling to tackle the over-smoothing problem [5]. An explicit denoising process in learning the DAE (Deep auto encoder) is included. Greedy layer wised pre-training plus fine tuning strategy in training of DAE which is used as filter for speech estimation [6]. DNN maps the relationship between the noisy and clean speech signals. A large training set of speech and noise files is designed and then DNN model is applied as a non regression model. Also, the strategy called global variance equalization, dropout and noise aware training is used which gives better results [7].

As the noise corruption process is complex, a non-linear model like neural network is used. Neural Network maps the relationship between clean and noisy speech signals. Early, shallow neural network (SNN) is used as a non-linear filter. It predicts the clean speech in the time or frequency domain. SNN takes only one hidden layer with 160 neurons. It was proposed to estimate the Signal-to-Noise ratio (SNR) on the amplitude modulation spectrograms and then noise is suppressed according to the estimated SNR of different channels. The SNR are estimated in the limited frequency resolution and it was not efficient to suppress the noise with sharp spectral peaks. Also, the small size of the network can not able to learn the mapping the relationship between the clean and noisy speech. So Deep neural network is used where the layer used are more in numbers. Here training is done using a large set of dataset which contains clean speech data and noise speech data.

In this paper, DNN based Speech Enhancement is used via training deep and large neural network architecture contains large set of data. Three strategies are also used to improve the quality of enhanced speech and generalization capability of DNN. Firstly, Global Variance Equalization is used. It equalizes between the global variance of the enhanced features and reference clean speech features. It was proposed to alleviate the over-smoothing issue in DNN based Speech enhancement system. Secondly, dropout is used. It deals with the over fitting issue. Third, noise aware training is used which deals to improve the performance. The paper is organized in following sections. Section I gives the basic introduction. Section II explains the System overview. Section III presents Results and discussion. Section IV gives Conclusion.

## 2. PROPOSED SYSTEM

The block diagram of proposed system is given in fig.1. A DNN is used for mapping the input and noisy features. The system is constructed in two parts: 1) Training Stage and 2) Enhancement stage.

In the training stage, the DNN model is trained using the log-spectral features from the pairs of noisy and clean speech. Here a log-spectral feature is used as it is assume to give perceptually relevant parameters. Firstly, short-time Fourier analysis is applied to the input signal to find the discrete Fourier transform of each overlapped window and then the log-spectral features are calculated.

In the enhancement stage, the noisy speech features are processed by trained DNN model to find the clean speech features. After we obtain the log-spectral features of clean speech,$\widehat{X^1}(d)$, the reconstructed spectrum $\widehat{X^f}(d)$is given by:

$$\widehat{X^f}(d) = \exp\left\{\frac{\widehat{X^1}(d)}{2}\right\}\exp\{j\langle Y^f(d)\} \qquad (1)$$

Where $\langle Y^f(d)$ denotes $d^{th}$ dimension phase of the noisy speech. A frame of speech signal is obtained from inverse DFT of the current frame spectrum. Finally, overlap add method is used to synthesize the waveform.

Another two blocks namely noise estimation for noise aware training and post processing with global variance equalization is used to improve the overall performance of the speech enhancement system. The dropout is used to improve the generalization capacity of DNN.

Now, we explain the basic DNN training procedure and then describe the different techniques used to improve the performance.

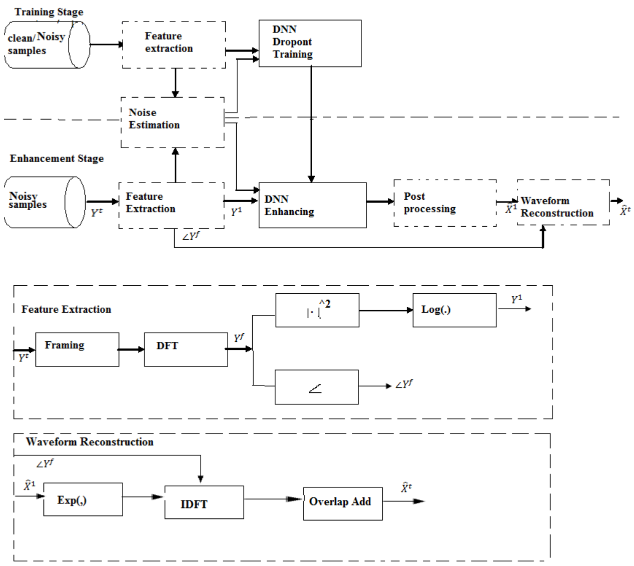

**Fig -1**: Block Diagram of speech enhancement system

## 2.1 Basic DNN Training

The architecture used here is feed forward neural network. It has many levels of non linearities so that it represents the highly non-linear regression model which maps the noisy features to the clean features. Here the training of DNN model consists of unsupervised pre-training part and a supervised fine-tuning part. The hidden unit is of sigmoid type and output unit is linear. We first pre-train a deep generative model with the normalized log-spectra of noisy speech by stacking multiple restricted Boltzmann machines and then the back-propagation algorithm with the MMSE based object function between the normalized log-power spectral features of the estimated and the reference clean speech is used to train the DNN. A mini-batch stochastic gradient descent algorithm is used to improve the error function as given by:

$$E_r = \frac{1}{N}\sum_{n=1}^{N}\left\|\widehat{X_n}(Y_{n-\tau}^{n+\tau}, W, b) - X_n\right\|_2^2 \quad (2)$$

Where $E_r$ is the mean squared error, $\widehat{X_n}(Y_{n-\tau}^{n+\tau}, W, b)$ and $X_n$ is the estimated and reference normalized log-spectral features at sample index n, N is the batch size.

## 2.2 Post-processing with global variance equalization

One of the error residual problems is over-smoothing which causes a muffling effect on the estimated clean speech when compared with reference clean speech. Global variance equalization is used to alleviate this problem. Here the equalization between the global variance of the estimated and reference clean speech features is taken.

The global variance of the estimates clean speech features is defined as:

$$GV(d) = \frac{1}{M}\sum_{n=1}^{M}\left(\widehat{X_n}(d) - 1M\sum_{n=1}^{M}\widehat{X_n}\,(d)\right)^2 \qquad (3)$$

Where $\widehat{X_n}(d)$ is the d-th component of a DNN output vector at n-th frame and M is the total number of speech frames in the training set.

To address the over-smoothing problem, a global equalization factor $\alpha(d)$ is given as follows:

$$\alpha(d) = \sqrt{\frac{GV_{ref}(d)}{GV_{est}(d)}} \qquad (4)$$

Where $GV_{ref}(d)$ and $GV_{est}(d)$ is the d-th dimension of the global variance of the reference features and the estimation features respectively.

A dimension independent global equalization factor β is given by:

$$\beta = \sqrt{\frac{GV_{ref}}{GV_{est}}} \qquad (5)$$

Where $GV_{ref}$ and $GV_{est}$is the dimension independent global variance of the reference and the estimation features respectively.

## 2.3 Dropout Training

One of the challenges in DNN based speech enhancement system is to address the mismatches between the training and testing conditions which is caused by different SNR levels, speaker variabilities, noise type, etc. To address this mismatches issues, a strategy called dropout is used to improve the generalization capability of the DNN.

In the DNN training, dropout randomly omits a certain percentage (says p) of the neurons in the input and each hidden layers during each presentation of sample for training of that each sample. This prevents complex co-adaption where the activations of the multiple nodes are highly correlated. This strategy might cause degradation in performance in matching noise types while improve the robustness in mismatched cases especially that noises which is not seen in the training data. At the enhancement stage, DNN discounts all the weights that are involved in the dropout training (1-p).

### 2.4 Noise aware training

The relationship between the clean speech and noise signal is non linear and complicated so it is difficult to estimate the clean speech spectra especially for non-stationary noises. The noise information of each utterance was not specifically utilized in the basic training of DNN. To enable this noise awareness, DNN is fed with the noisy speech samples augmented with an estimate of the noise. So DNN can use this additional on-line noise information to predict the clean speech.

The input vector of the DNN with the noise estimate is given by:

$$V_n = [Y_{n-\tau}, \dots., Y_{n-1}, Y_n, Y_{n+1}, \dots, Y_{n+\tau,}\widehat{Z_n}] \quad (6)$$

$$\widehat{Z_n} = \frac{1}{T}\sum_{t=1}^{T} Y_t \quad (7)$$

Where $Y_n$ is the log-spectral feature vector of the current noisy speech frame n and $\widehat{Z_n}$ is fixed over the utterance and estimated using the first T frames.

## 3. RESULTS AND DISCUSSION

The database is taken from the Aurora2 database. Here 8 different noise types such as airport noise, street noise, restaurant noise, car noise, exhibition noise, babble noise, station noise and train noise are used. The clean speech file is corrupted with above mentioned noises at different SNR levels, i.e., 20 dB, 15 dB, 10 dB and 5 dB, to build a multi-condition training set of pairs of noisy and clean speech signals. All these noise types are taken for the training of DNN. A speech enhancement method MMSE is taken for performance comparison with our DNN approach.

All the clean speech and noise waveform were down-sampled to 8 KHz. The frame length is 32msec i.e., 256 samples are taken of each speech file and frame shift is 16msec i.e., 128 samples. The dimension of the log-spectral feature vector is 129. Perceptual evaluation of speech quality (PESQ) was used for objective measure. PESQ is calculated by comparing the enhanced speech with the clean reference speech. It ranges from -0.5 to 4.5 [18].Other subjective measures such as Segmental SNR (SSNR), Log spectral distortion (LSD), Log likelihood ratio (LLR) is taken also. All of them are obtained by comparing the enhanced speech with clean reference speech. SSNR denotes the degree of noise reduction, while LSD represents the speech distortion.

The number of epoch for the RBM pre-training in each layer is set to 20. The learning rate of pre-training was set to 0.0005. For the fine-tuning of the baseline, the learning rte was set to 0.1 for the first 10 epochs and then decreased by 10% after each subsequent epoch. The momentum rate is 0.9. The total number of epoch at this stage is 50. The mini-batch size N is set to 128. For the back-propagation algorithm with dropout regularization, the corruption level for input layer is 0.1 and for each hidden layer is 0.2. The initial momentum rate of dropout is 0.5 and then the rate increases to 0.9 in the first 10 epochs, after it kept as 0.9. The first T=6 frames of each utterance is used for a noise estimate in NAT.

Table 1 list for PESQ value for 3 SNR values and 4 noise types without dropout and noise aware training.

**Table -1**: PESQ values for 3 SNR and 4 noise types without dropout and noise aware training

| Noise Type | 0 dB | 5 dB | 10 dB |
|---|---|---|---|
| Airport | 3.0107 | 3.2015 | 3.4066 |
| Babble | 3.0564 | 3.3877 | 3.3999 |
| Car | 3.2061 | 3.3714 | 3.5142 |
| Exhibition | 3.4858 | 3.4883 | 3.4709 |

Table 2 lists the DNN with global variance equalization and without global variance equalization.

**Table -2:** DNN value with GV and without GV

| Noise Type | Without GV equalization | With GV equalization |
|---|---|---|
| Airport | 3.4066 | 3.6399 |
| Babble | 3.3999 | 3.4721 |
| Car | 3.5142 | 3.5862 |
| Exhibition | 3.4709 | 3.9982 |

Table 3.lists the DNN value with only DNN and DNN with different three strategy used and DNN with all three techniques.

**Table -3:** PESQ value of DNN of noisy speech and Enhanced speech

| Noise type with SNR value | DNN | DNN with GV | DNN with Drop-out | DNN with NAT | DNN with GV, Dropout and NAT |
|---|---|---|---|---|---|
| Airport_0dB | 2.8303 | 2.7521 | 3.037 | 3.1502 | 3.1668 |
| Babble_0dB | 3.004 | 2.8865 | 2.8944 | 3.0531 | 3.21 |
| Car_0 dB | 2.7778 | 2.8711 | 2.7733 | 2.8786 | 3.212 |
| Exhibition_0 dB | 2.7843 | 2.8714 | 2.9022 | 3.066 | 3.48 |

**Table -4:** Different parameters with different noise types at different SNR

| Noise type | 0 dB | | | | 5 dB | | | |
|---|---|---|---|---|---|---|---|---|
| | SSNR | PESQ | LLR | LSD | SSNR | PESQ | LLR | LSD |
| Airport | -2.08 | 3.06 | 0.11 | 0.37 | -7.73 | 3.18 | 0.11 | 0.38 |
| Babble | -2.30 | 3.23 | 0.13 | 0.39 | -8.02 | 3.06 | 0.11 | 0.37 |
| Car | -2.48 | 3.03 | 0.16 | 0.42 | -8.11 | 3.17 | 0.14 | 0.40 |

Figure 2 describes waveform of clean speech and noisy speech

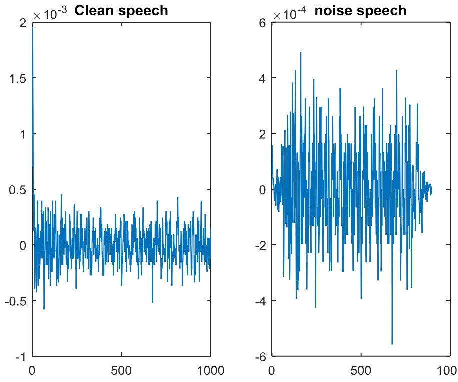

**Fig -2**: Waveform of clean speech and noisy speech

Figure 3 describes the Original speech and reconstructed speech.

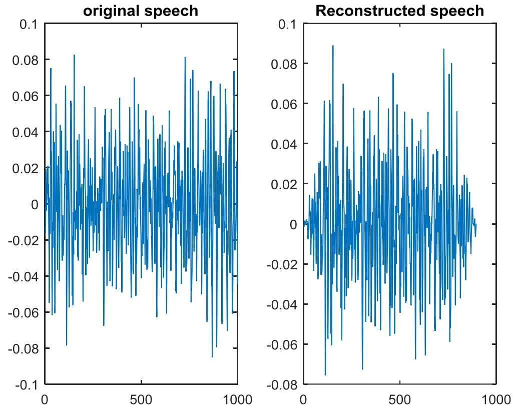

**Fig -3**: Waveform of original speech and reconstructed speech

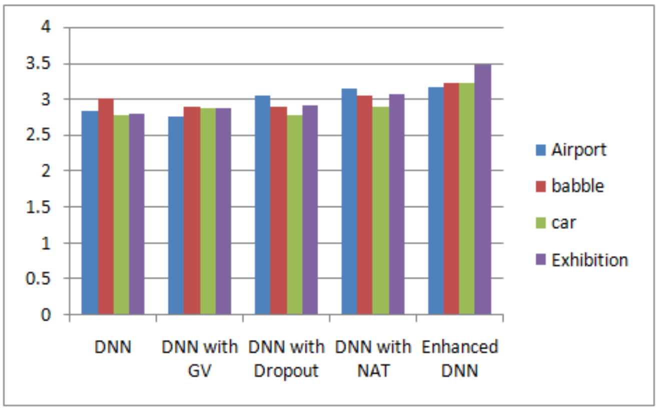

**Chart-1**: PESQ value for different Noise types at SNR=0 dB using different strategies.

Figures 5 and 6 describes the Spectrogram of Noisy speech and DNN enhanced speech respectively.

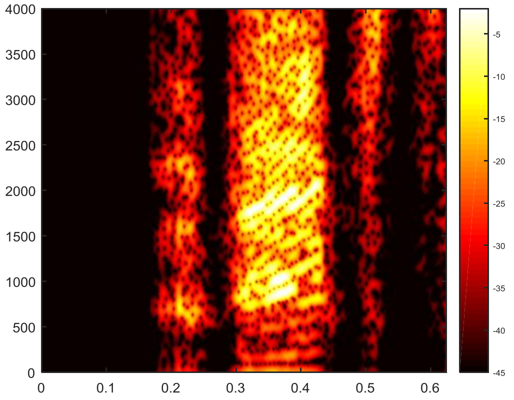

**Fig -5**: Spectrogram of noisy speech

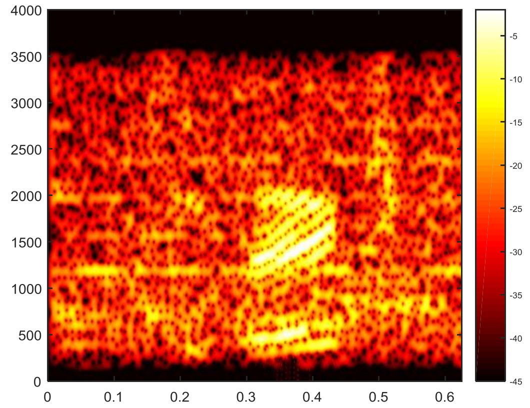

**Fig -5**: Spectrogram of DNN Enhanced speech

Table 5 lists the comparison between DNN and MMSE speech enhancement method. (a) is the value of MMSE and (b) is the value of DNN Enhanced.

**Table -5:** PESQ value of comparison between MMSE and DNN Enhanced

| SNR | Airport | | Babble | | Car | |
|---|---|---|---|---|---|---|
| | (a) | (b) | (a) | (b) | (a) | (b) |
| SNR 0dB | 1.983 | 3.010 | 2.155 | 3.056 | 2.033 | 3.219 |
| SNR 5dB | 2.385 | 3.201 | 2.326 | 3.387 | 2.334 | 3.371 |
| SNR 10dB | 2.631 | 3.278 | 2.499 | 3.836 | 2.440 | 3.355 |
| Average | 2.333 | 3.163 | 2.326 | 3.426 | 2.269 | 3.315 |

## 4. CONCLUSIONS

In this paper, DNN based Speech Enhancement system is proposed. By using various DNN configurations, a large training set is needed to learn the mapping function between the noisy and clean speech features. Also the three strategies used with DNN gives good results. Proposed post-processing technique, Global variance equalization is effective in brightening the formants spectra of the enhanced speech signals. Also two improved training techniques reduce the residual noise and increase the performance. As compared with the typical speech enhancement method called MMSE, our DNN proposed system gives better results.

## REFERENCES

[1] Jae S. Lim, Alan V. Oppenheim, “Enhancement and Bandwidth Compression of Noisy Speech,”Proceedings of the IEEE, VOL. 67, no. 12, December 1979.

[2] Yariv Ephraim, David Malah, “Speech Enhancement Using a- Minimum Mean- Square Error Short-Time Spectral Amplitude Estimator,” IEEE Transactions on Acoustica, Speech and Signal processing, Vol. ASSP-32, no.6, December 1984.

[3] Yariv Ephraim, David Malah, “Speech Enhancement Using a- Minimum Mean- Square Error Short-Time Spectral Amplitude Estimator,” IEEE Transactions on Acoustica, Speech and Signal processing, Vol. ASSP-32, no.6, December 1985.

[4] Olivier Cappe, “Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor,” IEEE transactions 1994.

[5] Zhen-Hua Ling, LiDeng, and Dong yu, “Modeling Spectral Envelopes Using Restricted Boltzmann Machines and Deep Belief Networks for Statistical Parametric Speech Synthesis,” IEEE Transactions on Audio, speech and Language, vol.21. no.10. October 2013.

[6] Xugang Lu, Yu Tsao, Shigeki Matsuda, Choiri Hori, “Speech Enhancement Based on Deep Denoising Autoencoder,” in Proc. Interspeech, 2013,pp 436-440.

[7] Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, “A Regression Approach to Speech EnhancementBased on Deep Neural Networks,” IEEE/ACM transactions on audio, speech language processing, vol.23, no.1, January 2015.

[8] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoustic, Speech, Signal Process.*, vol. ASSP-27,no. 2, pp. 113–120, Apr. 1979.

[9]J. Tchorz and B. Kollmeier, “SNR estimation based on amplitude modulation analysis with applications to noise suppression,” *IEEE Trans.Speech Audio Process.*, vol. 11, no. 3, pp. 184–192, May 2003.

[10] I. Cohen and B. Berdugo, “Speech enhancement for non-stationary noise environments,” *Signal Process.*, vol. 81, no. 11, pp. 2403–2418, 2001.

[11] Y. X. Wang and D. L. Wang, “Towards scaling up classification-basedspeech separation,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013.

[12] S. Rangachari and P. C. Loizou, “A noise-estimation algorithm forchighly non-stationary environments,” *Speech Commun.*, vol. 48, no.2, pp. 220–231, 2006. of feature detectors,” *Arxiv*, 2012.

[13] M. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. ICASSP*, 2013, pp. 7398–7402.

[14] A. Narayanan, D. L. Wang, "Joint noise adaptive training for robust automatic speech recognition," in *Proc. ICASSP*, 2014, pp. 2523–2527.

[15]G. Shi, M. M. Shanechi, and P. Aarabi, "On the importance of phase in human speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1867–1874, Sep. 2006.

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

[17] S. Rangachari, P. C. Loizou, and Y. Hu, "A noise estimation algorithm with rapid adaptation for highly nonstationary environments," in *Proc. ICASSP*, 2004, pp. 305–308.

[18] ITU-T, Rec. P.862, Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs International Telecommunication Union-Telecommunication Standardization Sector, 2001.

[19] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," in *Proc. ISCA ITRW ASR*, 2000, pp. 181–188.

[20] P. C. Loizou*, Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC, 2013.