DNN-BASED AR-WIENER FILTERING FOR SPEECH ENHANCEMENT

Yan Yang, Changchun Bao

Speech and Audio Signal Processing Lab, Faculty of Information Technology
Beijing University of Technology, Beijing, China
yangyan00800@emails.bjut.edu.cn, baochch@bjut.edu.cn

ABSTRACT

This paper presents a novel approach for estimating auto-regressive (AR) model parameters using deep neural network (DNN) in the AR-Wiener filtering speech enhancement. Unlike conventional DNN that predicts one kind of target, the DNN used in this paper is trained to predict the AR model parameters of speech and noise simultaneously at offline stage. We train this network by minimizing the Euclidean distance between the output of DNN and the AR model parameters of clean speech and noise. At online stage, the acoustic features are first extracted from noisy speech as the input of the DNN. Then, AR model parameters of speech and noise are estimated by the DNN simultaneously. Finally, the Wiener filter is constructed by the AR model parameters of speech and noise. However, the AR model parameters only models the spectral shape not the spectral details, there are still some residual noise between the harmonics. In order to solve this problem, we introduce the speech-presence probability (SPP), that is, in the test stage, the SPP is estimated and is used to update the Wiener filter. The experimental results show that our approach has higher performance compared with some existing approaches.

Index Terms— speech enhancement, deep neural network, auto-regressive model, speech-presence probability, Wiener filter

1. INTRODUCTION

Speech enhancement aims to suppress the background noise while maintaining the quality and the intelligibility of speech. There are a lot of classical approaches, such as spectral subtraction [1], statistical model-based method [2][3], and subspace-based method [4][5]. These methods are very suitable for removing stationary noise but have poor performance in non-stationary noise condition. It is mainly because that these methods cannot follow the rapid changes of non-stationary noise energy.

In order to remove the non-stationary noise, some codebook-driven methods [6][7][8][9] have been proposed in recent years. In these methods, the codebooks containing the spectral shape information are trained offline using LBG [10] algorithm. In these methods, the Sparse Hidden Markov model-based method could estimate the AR gains of speech and noise effectively [8]. In this method, the likelihood criterion for finding the model parameters is augmented with a regularization term, which encourages sparsity in the AR models of speech and noise.

Although codebook-driven speech enhancement methods are suitable for suppressing the non-stationary noise, there are some drawbacks. The first one is that the codebooks only have prior information of spectral shapes but not the spectral details, which result in much harmonic noise. The second one is that codebooks have some quantization error in the training process offline.

Deep neural network (DNN) has played an important role in speech enhancement. For example, reference [11] well applied DNN to remove the background noise. In the training stage, the acoustic features of log spectral power of noisy speech and the training targets of log spectral power of clean speech are fed into DNN. In the test stage, the acoustic features of log spectral power of noisy speech are extracted same as the training stage, and the acoustic features are fed into DNN to obtain the log spectral power of clean speech. Test results show that this method can achieve significant improvements in both objective and subjective measures. In the reference [12], the DNN was used to estimate the ideal ratio mask (IRM) in the Gammatone domain. In the test stage, the estimated IRM is used to mask the time-frequency (T-F) unit. In this way, the target speech's T-F units is preserved in the synthetic speech. The results showed that this mask-based method is better than the feature mapping-based methods [12].

In our proposed method, the DNN is trained to estimate the AR model parameters of speech and noise simultaneously because of the good prediction performance of the DNN. In the training stage, the acoustic features are first extracted as the input of DNN. The training targets of DNN are the connected vectors of AR model parameters of speech and noise. Here, we use the linear spectrum frequency (LSF) parameters [13] of speech and noise as the training targets and connect them to form a target vector. The network is trained by minimizing the Euclidean distance between the output of DNN and the training target. In the test stage, the acoustic features are extracted from the mixture firstly. Secondly, the acoustic features are fed into the DNN to obtain the AR model parameters of speech and noise. Thirdly, the multiplicative update rule [9] is adopted to estimate the AR gains. At last,

the AR-Wiener filter is constructed by the AR model parameters of speech and noise. On this basis, in order to remove residual noise between the harmonics, speech-presence probability (SPP) [14] is adopted to update the AR-Wiener filter.

The paper is organized as follows. The overview of our proposed method is presented in Section 2. Section 3 elaborates the experimental setup and the test results. We draw a conclusion in Section 4.

2. PROPOSED METHOD

In this section, the noisy speech is modeled as follows:

$$\mathbf{y}_n = \mathbf{x}_n + \mathbf{w}_n \tag{1}$$

$\mathbf{y}_n$, $\mathbf{x}_n$ and $\mathbf{w}_n$ denote the noisy speech, clean speech and noise signal, respectively. This section is divided into two parts. The first part denotes the procedure of estimating AR model parameters using DNN and the construction of AR-Wiener filter. The second part shows the details about speech-presence probability.

2.1. Estimation of AR model parameters with DNN

In this section, the DNN is adopted as the mapping function from noisy features to the AR model parameters of speech and noise. Fig. 1 shows the block diagram of the proposed speech enhancement system.

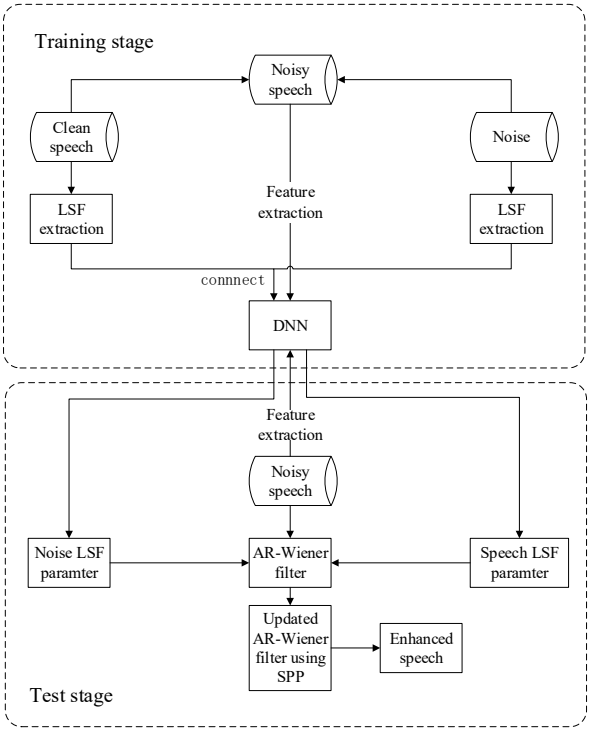

Fig. 1 Block diagram of the proposed method

In the training stage, we obtain noisy speech by combining clean speech and noise from their corpus frame by frame. The acoustic features are extracted from noisy speech and fed into the DNN along with the corresponding desired training target. Here, we use the log-power spectra (LPS) [11] as the acoustic features. The input of DNN combines the LPS of the current frame, five previous frames and five future frames. Thus, we combine the 11 frames of LPS vectors into one vector as the input of DNN:

$$[x_1^{l-5}, x_2^{l-5}, ...x_m^{l-5} ..., x_1^l, ..., x_m^l, ..., x_1^{l+5}, ...x_m^{l+5}] \tag{2}$$

where m denotes the dimension of one frame's LPS and l denotes index of frame. This treatment considers the continuity of time and can improve the performance [15]. The training target is the connected vector that combines the linear spectrum frequency (LSF) [10] parameters of speech and noise:

$$[p_1^x, p_2^x, ..., p_n^x, p_1^w, ..., p_n^w] \tag{3}$$

where p_i^x denotes the speech's LSF parameters and p_i^w denotes the noise's LSF parameters. n denotes the dimension of LSF parameters. In this way, the DNN can be trained to predict the LSF parameters of speech and noise simultaneously. The DNN has three hidden layers, each hidden layer has 512 rectified linear hidden units (Relu) [16]. Here, we use 10 order of LSF parameters of speech and noise (n=10). So the dimension of the output layer is 20. The cost function of the network is Euclidean distance:

$$J(\boldsymbol{v}, \boldsymbol{b}) = \frac{1}{M} \sum_{i=1}^{M} [\boldsymbol{y}^{(i)} - h_{\boldsymbol{v},\boldsymbol{b}}(\boldsymbol{x}^{(i)})]^2 \tag{4}$$

$\boldsymbol{x}^{(i)}$ is the i^{th} input vector and $\boldsymbol{y}^{(i)}$ is the i^{th} expected output vector related to the training target. $h_{\mathbf{v},\mathbf{b}}(\cdot)$ denotes the whole DNN with the connection weights $\boldsymbol{v}$ and biases $\boldsymbol{b}$. M is the total number of training vectors. The training algorithm we used is the stochastic gradient descent (SGD) [17] with the momentum:

$$\boldsymbol{v}(k+1) = \boldsymbol{v}(k) - \eta[(1-\alpha)\frac{\partial J_k(\boldsymbol{v},\boldsymbol{b})}{\partial \boldsymbol{v}} + \alpha \frac{\partial J_{k-1}(\boldsymbol{v},\boldsymbol{b})}{\partial \boldsymbol{v}}] \tag{5}$$

where k is the index of training epochs. η is the learning rate and α is the momentum. The momentum is set to 0.2 in the first 20 training epochs and is raised to 0.5 in the following training epochs. The learning rate is set to 0.03 in the first epoch and decayed by multiplying 0.98 in the following epochs. The SGD with momentum technique can avoid the problem falling in local optima. The dropout rate is set to 0.2, which can prevent the overfitting problem of the DNN [18].

In the test stage, the acoustic features are extracted from noisy speech just like in the training stage. And the output of DNN is a vector that combines the LSF parameters of the estimated speech and noise. The former 10 dimensions of the output represents the estimated speech's LSF parameters and the later 10 dimensions represents the noise's LSF parameters just like in the Eq. (3). Then, we transform the LSF parameters to LPC parameters according to [19]. Thus, the spectral shape of speech, $1/|A_x(k)|^2$, is calculated,
where

$$A_x(k) = \sum_{m=0}^{p} \hat{a}_{x,m} \exp(-\frac{2\pi m}{K} k) \quad (6)$$

$\hat{a}_{x,m}$ is the m^{th} dimension of estimated speech LPC parameter. K presents the Fast Fourier Transform (FFT) size. And the spectral shape of noise, $1/|A_w(k)|$, is calculated in the same way.

At last, the AR-Wiener filter is constructed as following equation:

$$WF_{AR}(k) = \frac{\hat{g}_x}{|A_x(k)|^2} \Big/ \frac{\hat{g}_x}{|A_x(k)|^2} + \frac{\hat{g}_w}{|A_w(k)|^2} \quad (7)$$

$\hat{g}_x$ and $\hat{g}_w$ denote AR gains of speech and noise respectively, which are calculated by multiplicative update rule [9]:

$$\hat{g}_x \bullet \frac{(H_x)^T[(H_y W_y)^{-2} \bullet P_y]}{(H_x)^T (H_y W_y)^{-1}} \to \hat{g}_x \quad (8)$$

$$\hat{g}_w \bullet \frac{(H_w)^T[(H_y W_y)^{-2} \bullet P_y]}{(H_w)^T (H_y W_y)^{-1}} \to \hat{g}_w \quad (9)$$

where

$$\hat{\boldsymbol{P}}_y = [\hat{P}_y(0),...,\hat{P}_y(K-1)]^T \quad (10)$$

$$\boldsymbol{H}_x = [\frac{1}{|A_x(0)|^2},...,\frac{1}{|A_x(K-1)|^2}]^T \quad (11)$$

$$\boldsymbol{H}_w = [\frac{1}{|A_w(0)|^2},...,\frac{1}{|A_w(K-1)|^2}]^T \quad (12)$$

$$\boldsymbol{W}_y = [g_y] \quad (13)$$

$$\boldsymbol{H}_y = [\frac{1}{|A_y(0)|^2},...,\frac{1}{|A_y(K-1)|^2}]^T \quad (14)$$

In Eq. (10), $\hat{P}_y$ is the power spectrum of noisy speech. H_x,H_w and H_y are the spectral shape of speech, noise and noisy speech, respectively. W_y is the AR gain of noisy speech. The symbol "•" represents the point-wise multiplication.

2.2 Updated AR-Wiener filter using SPP

The proposed method in previous section can track the rapid changing of noise energy, so it is very suitable to suppress the non-stationary noise. However, there are also residual noise between the harmonics because the above method is only used to model the spectrum shape but not the spectrum details. In this part, we estimate the speech-presence probability (SPP) [14], which is used to update the AR-Wiener filter above. This method can reduce the noise between the harmonics.

For calculating SPP, H_0^k denotes the state that speech is absent in frequency bin k, while H_1^k denotes the speech is present. So, the SPP is calculated by $P(H_1^k|Y_k)$:

$$P(H_1^k \mid Y_k) = \frac{P(Y_k \mid H_1^k) P(H_1^k)}{P(Y_k)} \quad (15)$$

$$= \frac{P(Y_k \mid H_1^k) P(H_1^k)}{P(Y_k \mid H_1^k) P(H_1^k) + P(Y_k \mid H_0^k) P(H_0^k)} \quad (16)$$

where Y_k denotes the spectrum of noisy speech. P(H_1^k) is the prior probability of speech presence. For the state H_0^k, Y_k has the same spectrum as the noise's spectrum D_k. Because D_k obeys the Gauss distribution with zero mean and $\lambda_d(k)$ variance, $P(Y_k|H_0^k)$ is the same:

$$P(Y_k \mid H_0^k) = \frac{1}{\pi\lambda_d(k)} \exp(-\frac{Y_k^2}{\lambda_d(k)}) \quad (17)$$

for the state H_1^k, $Y_k = D_k + X_k$, where X_k is the spectrum of clean speech. So, the $P(Y_k|H_1^k)$ satisfies the Gauss distribution with zero mean and $(\lambda_d(k) + \lambda_x(k))$ variance:

$$P(Y_k \mid H_1^k) = \frac{1}{\pi(\lambda_x(k) + \lambda_d(k))} \exp[-\frac{Y_k^2}{(\lambda_x(k) + \lambda_d(k))}] \quad (18)$$

where λ_x is the variance of speech power. Then, we put the Eq. (17) and Eq. (18) into Eq. (16), we can get the SPP as follows:

$$\hat{P}(H_1^k \mid Y_k) = \frac{1 - q_k}{1 - q_k + q_k(1 + \xi_k') \exp(-\nu_k')} \quad (19)$$

where

$$\xi_k' = \frac{\xi_k}{1 - q_k} \qquad \nu_k' = \frac{\xi_k'}{\xi' + 1} \gamma_k \quad (20)$$

$q_k = P(H_0^k)$. ξ_k is priori SNR, γ_k is posteriori SNR:

$$\xi_k = \frac{X_k^2}{D_k^2} \approx \frac{P_x(k)}{P_w(k)} \qquad \gamma_k = \frac{Y_k^2}{D_k^2} \approx \frac{Y_k^2}{P_w(k)} \quad (21)$$

where

$$P_x(k) = \frac{\hat{g}_x}{|A_x(k)|^2} \qquad P_w(k) = \frac{\hat{g}_w}{|A_w(k)|^2} \quad (22)$$

$P_x(k)$ and $P_w(k)$ are the power spectrum of speech and noise and can be calculated by the estimated AR-gains and spectral shape in the section 2.1. $\hat{g}_x$ and $\hat{g}_w$ are calculated according to the Eq. (8) and Eq. (9). $A_x(k)$ is calculated according to the Eq. (6).

To obtain the SPP, We first calculate the priori SNR and posteriori SNR by the Eq. (21). And then we obtain the SPP according to Eq. (19). Finally, the AR-Wiener filter in Section 2.1 is updated by the SPP as follow:

$$WF_{updated}(k) = \hat{P}(H_1^k \mid Y_k) WF_{AR}(k) \quad (23)$$

$WF_{updated}(k)$ is the updated AR-Wiener filter. When the frequency bin k is between adjacent harmonic frequencies, the power spectrum of noise is much larger than the power of speech. So the priori SNR is small, which leads to decline in SPP. In this way, the noise between harmonics is removed.

3. EXPERIMENTS AND RESULT

In this section, we compare our proposed method with three reference methods including DNN-based amplitude recovering [11], Sparse Hidden Markov Models method [8] and DNN-based ideal ratio mask (IRM) method [12]. For convenience, we name them as Ref. A, Ref. B and Ref. C, respectively. Our two proposed methods are named as Pro. A which is the AR-Wiener filtering without SPP and Pro. B which is AR-Wiener filtering with SPP. In our experiments, eight hours of training speech from TIMIT databases is used to train the DNN. The sampling rate is 8kHz. The input noisy

speech with the frame size 32ms is windowed using Hamming window with 50% overlap between the adjacent frames. The FFT size is 256. The training noises including babble, f16, factory and buccaneer come from Noisex-92 database.

In the training stage, input features (LPS) are extracted from noisy speech with -5dB, 0dB and 5dB signal-to-noise ratio (SNR). In the test stage, the test set contains ten male and ten female utterances from TIMIT database and are corrupted by four kinds of noise just as in training stage. We add noisy utterances in the test set with 10dB input SNR in order to test the generalization ability of network. The performance is evaluated by the segment SNR (SSNR) [20], the perceptual evaluation of speech quality (PESQ) [21], and the short-time objective intelligibility (STOI) [22]. Moreover, we also give the spectrum comparison of the enhanced speech.

Table.1 Evaluation results of SSNR

Enhancement methods	Average SSNR Improvement			
	-5dB	0dB	5dB	10dB
Ref. A	12.0243	9.1580	6.3286	3.3413
Ref. B	11.5093	9.1606	6.1674	5.8473
Ref. C	10.7430	9.9106	8.7555	7.2213
Pro. A	13.8090	12.3177	10.1973	7.4086
Pro. B	**14.1283**	**13.1526**	**11.6552**	**9.4908**

Table.2 Evaluation results of PESQ

Enhancement methods	Average PESQ			
	-5dB	0dB	5dB	10dB
Noisy	1.4180	1.6824	2.0107	2.3455
Ref. A	1.3209	1.6338	2.0067	2.3140
Ref. B	1.4120	1.8497	2.3050	2.6476
Ref. C	1.5932	1.9532	2.3352	**2.7484**
Pro. A	1.5819	1.9854	2.3414	2.6585
Pro. B	**1.6666**	**2.0452**	**2.3942**	2.7318

Table.3 Evaluation results of STOI

Enhancement methods	Average STOI			
	-5dB	0dB	5dB	10dB
Noisy	0.5148	0.6300	0.7453	0.8433
Ref. A	0.5253	0.6381	0.7456	0.8158
Ref. B	0.4981	0.6118	0.7211	0.8093
Ref. C	0.5989	0.7119	0.8114	0.8876
Pro. A	0.6114	0.7218	0.8298	0. 8829
Pro. B	**0.6312**	**0.7490**	**0.8351**	**0.8945**

From the Table 1-3, we can see that the Pro. A get more satisfactory results in each objective evaluation result comparing with Ref. A. By comparing Fig. 2 (f) with the Fig. 2 (c), the Pro. A can preserve more components of speech's spectrum while remove the background noise, which can get better quality in the enhanced speech.

In the Fig. 2 (d), the Ref. B has less residual noise comparing with other methods, but meanwhile, the components of speech are also destroyed a lot. So the STOI score of the Ref. B is the lowest.

The Ref. C is comparable with our proposed methods. In the -5dB, 0dB and 5dB noisy environments, our proposed methods outperform the Ref. C a little bit. But in the 10dB noisy environments, The Ref. C get the highest score in the PESQ and STOI. That is because the 10dB noisy speech is not in our training set, so the performance is not satisfactory.

By comparing the Fig. 2 (g) and (f), we can see the Pro. B preserve more details than the Pro. A. In each objective test result, the Pro. B gets more satisfactory results than the Pro. A. It is mainly because that the SPP can remove the noise between harmonics to some extent.

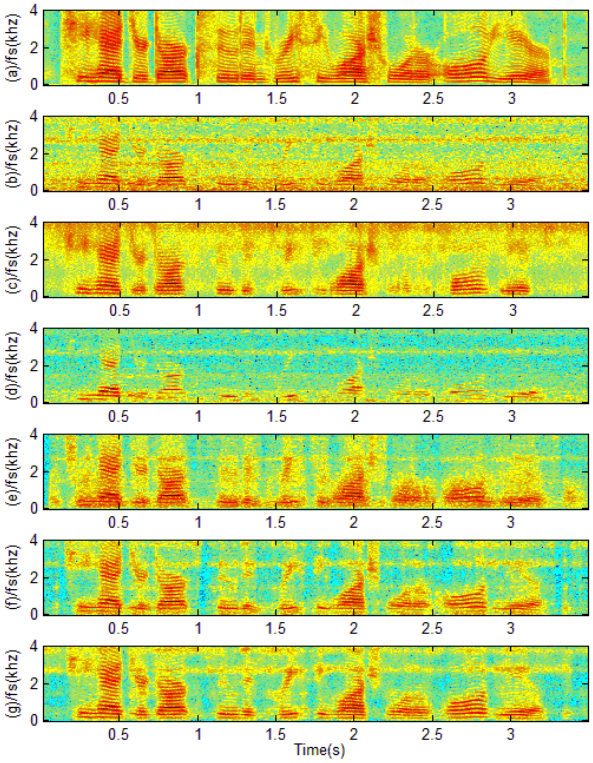

Fig. 2 Spectrum comparison, (a) clean speech; (b) noisy speech (f16 noise, input SNR=0dB); (c) Ref. A; (d) Ref. B; (e) Ref. C; (f)Pro. A (g)Pro. B

4. CONCLUSIONS

In this paper, the DNN is used to estimate the AR model parameters of speech and noise simultaneously, and the AR-Wiener filter is constructed by the estimated AR model parameters. The speech-presence probability (SPP) is adopted to remove the residual noise between harmonics. The test results show that our proposed method gets satisfactory result comparing with other references. In the future work, we can try other input features and other network structure which can take into account the temporal correlations.

5. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 6147014, No. 61231015).

6. REFERENCES

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction", IEEE Transaction on Acoustic, Speech and Signal Processing, vol. 27, no. 2, pp. 113-120, 1979.

[2] R. Martin, "Speech enhancement using MMSE short time spectral estimation with Gamma distributed speech priors", International Conference on Speech, Acoustics and Signal Processing, vol. 32, pp. 253-256, 2002.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral estimator." IEEE Transaction on Acoustic, Speech and Signal Processing, pp. 1109-1121, 1984.

[4] R. Vetter and H. L. Van Trees. "A signal subspace approach for speech enhancement". IEEE Transaction on Acoustic, Speech and Signal Processing, vol. 34, pp. 251-266, 1995.

[5] Schmidt, R, "A signal subspace approach to multiple source location and spectral estimation" IEEE Signal Processing Magazine, 1981

[6] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven shortterm predictor parameter estimation for speech enhancement", IEEE Trans. Audio, Speech, Lang. Process., vol. 14, no. 1, pp. 163–176, Jan.2006.

[7] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based Bayesian speech enhancement for nonstationary environments", IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 2, pp. 441–452,Feb. 2007.

[8] Feng Deng and Changchun Bao, "Sparse Hidden Markov Models for Speech Enhancement in Non-Stationary Noise Environments". IEEE Trans. Audio, Speech, Lang. Process., vol. 23, no. 1, pp. 163–176, Nov.2015.

[9] Qi He, Chang-chun Bao and Feng Bao. "Multiplicative Update of AR Gains in Codebook-driven Speech Enhancement". In Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing. pp. 5230-5234, 2016.

[10] Chin-Chen Chang; Yu-Chen Hu, "A LBG codebook training algorithm for vector quantization", IEEE Transactions on Consumer Electronics, vol. 44, no 1, pp. 1201-1208, 1998.

[11] Yong Xu, Jun Du Li-Rong Dai. "A Regression Approach to Speech Enhancement Based on Deep Neural Networks" IEEE Transaction on Acoustic, Speech and Signal Processing, vol 23, no 1, Jan 2015.

[12] Wang Y., Narayanan A. and Wang D.L. "On training targets for supervised speech separation". IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, pp. 1849-1858, 2014.

[13] F.K. Soong and B. H. Juang. "Line Spectrum Pair (LSP) and Speech Data Compression" Conf. on Acoustics, Speech and Signal Processing, 1984

[14] P. C. Loizou, "Speech enhancement: theory and practice". Boca Raton, FL, USA: CRC Press, 2007.

[15] Wang Y., Han K., and Wang D.L, "Exploring monaural features for classification-based speech segregation". IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, pp. 270-279.

[16] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier networks". Conf. Artif. Intell. Statist. JMLR W&CP Volume, vol. 15, pp. 315–323. 2011.

[17] D. R. Sweet. "Performance of stochastic gradient descent adaptive beamforming using sonar data". Microwaves, Optics and Antennas, IEE Proceedings H. vol. 130. pp 147-151.1983.

[18] Nitish Srivastava, Geoffrey Hiton. "Dropout: A simple way to prevent Neural Networks from overfitting". Journal of Machine Learning Research. pp. 1929-1958. 2014.

[19] F. K . Kang and L. J. Fransen. "Application of Line Spectrum Pairs to Low-Bit-Rate Speech Encoders". IEEE Journal on Selcted Areas in Communications, pp. 432-440, 1988.

[20] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, "Objective measures of speech quality," Englewood Cliffs, NJ: Prentice Hall, 1988.

[21] "Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-End Speech Quality Assessment of Narrow-band Telephone Networks and Speech Codecs," ITU-T Recommendation, P.862, Feb, 2001.

[22] C.H.Taal, R.C.Hendriks, R.Heusdens, J.Jensen "A Short-Time Objective Intelligibility Measure for Time-Frequency Weighted Noisy Speech", ICASSP, pp. 157-162. 2010.