Speech Enhancement with Weighted Denoising Auto-Encoder

Bing-yin Xia, Chang-chun Bao

Speech and Audio Signal Processing Lab, School of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100124, China

xby-abc@emails.bjut.edu.cn, baochch@bjut.edu.cn

Abstract

A novel speech enhancement method with Weighted Denoising Auto-encoder (WDA) is proposed in this paper. A weighted reconstruction loss function is introduced to the conventional Denoising Auto-encoder (DA), and makes it suitable for the task of speech enhancement. First, the proposed WDA is used to model the relationship between the noisy and clean power spectrums of speech signal. Then, the estimated clean power spectrum is used in the *a Posteriori* SNR Controlled Recursive Averaging (PCRA) approach for the estimation of the *a priori* SNR. Finally, the enhanced speech is obtained by Wiener filter operating in the frequency domain. From the test results under ITU-T G.160, in comparison with the reference method, the proposed method could achieve similar amount of noise reduction in both white and colored noise, and the distortion on the level of speech signal is smaller. Also, the objective speech quality is improved in all the test conditions.

Index Terms: speech enhancement, weighted denoising auto-encoder, SNR estimation, Wiener filter

1. Introduction

For the mobile communication system which is operated in the complex environments, background noise is one of the main impairments to the speech quality. As a result, it is necessary to adopt the speech enhancement techniques to remove the noise and improve the quality of speech communication.

Single channel speech enhancement is the most challenging task and research focus in the recent years. Since the 1970s, several kinds of speech enhancement algorithms have been proposed. In 1979, S. F. Boll proposed the spectral subtraction algorithm [1] to remove the additive background noise. In the same year, speech enhancement method with frequency domain Wiener filter [2] was proposed by J. Lim and A. V. Oppenheim. In 1984, based on the statistical models of speech and noise signals, Y. Ephraim developed a Minimum Mean-Square Error (MMSE) Short-Time Spectral Amplitude (STSA) estimator [3] for speech enhancement. In 1995, wavelet thresholding method [4] was proposed by D. L. Donoho.

The state-of-art speech enhancement methods are based on either the additive nature of background noise (like the spectral subtraction method or wavelet thresholding method), or the statistical properties of speech and noise signal (like the statistical model based methods). However, it is a complex process that the noise corrupts the speech signal. An adaptive and non-linear model, like the neural networks, should be more suitable for modeling the relationship between the clean and noisy speech signals in the time and frequency domain.

In this paper, Denoising Auto-encoder (DA) [5], which is a variation of artificial neural networks, is adopted to model the relationship between clean and noisy power spectrums of speech signal. DA is a two-layer neural network in which the input is a distorted version of the target output. The training process of DA is to minimize the reconstruction loss between the restored version and the target output. In the recent years, a Deep Neural Network (DNN) called Stacked Denoising Auto-encoders (SDA), which is formed by a series of DAs, is used successfully in unsupervised pre-training and feature learning [6]. In 2012, an image denoising and inpainting method based on SDA [7] was proposed by J. Xie. In this paper, DA and SDA are proved to be capable for the signal denoising task.

In the proposed method, considering the fact that the same distortion in different frequency bands has different effect on speech quality, a weighted reconstruction loss function is introduced to the traditional DA. And the new model, which is referred to as Weighted Denoising Auto-encoder (WDA), is adopted to estimate the power spectrum of clean speech. Then the *a Posteriori* SNR Controlled Recursive Averaging (PCRA) approach is used for the estimation of the *a priori* SNR. Finally, the noisy speech is enhanced by the frequency domain Wiener filter. Experimental results show that, comparing with the Wiener filter using Decision-Directed method [3] for SNR estimation, the proposed method can provide similar amount of noise reduction with smaller speech distortion, and the objective speech quality is improved at the same time.

The rest of this paper is organized as follows: Section 2 provides the details of the proposed method. The results and discussion of performance evaluation is presented in Section 3. Section 4 shows the conclusion.

2. The Proposed Method

The block diagram of the proposed speech enhancement method with WDA is illustrated in Figure 1.

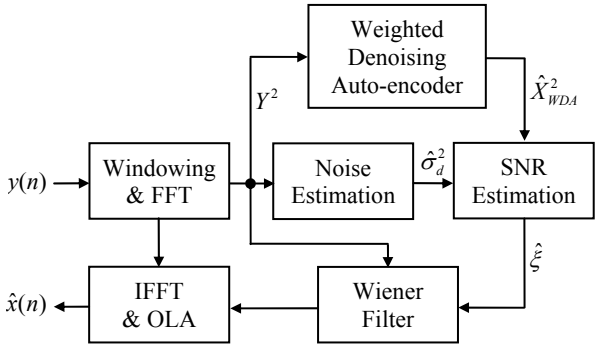

Figure 1: *Block diagram of the proposed method*

First, Fast Fourier Transform (FFT) is performed on the windowed noisy speech to get the amplitude and phase spectrums. Then, the noise power spectrum is estimated by Minima Controlled Recursive Averaging (MCRA) method [8]. Next, WDA is adopted to obtain the power spectrum estimate of clean speech, and the *a priori* SNR is estimated using the PCRA approach. Finally, the power spectrum of enhanced

speech is obtained by the frequency domain Wiener filtering, and IFFT and Overlap-and-Add (OLA) are performed to get the enhanced speech.

The details of the proposed method will be described in the following sub-sections.

2.1. Weighted Denoising Auto-Encoder

Assuming that $\mathbf{X^2}$ and $\mathbf{Y^2}$ are the power spectrums of the clean speech and noisy speech, respectively, we can formulate the noise corruption process of speech signal as:

$$\mathbf{Y}^2 = \eta(\mathbf{X}^2) \tag{1}$$

where $\eta : \mathbb{R}^n \to \mathbb{R}^n$ is an stochastic corrupting process that corrupts the input power spectrum $\mathbf{X^2}$. Then, the learning objective of speech enhancement task can be expressed as:

$$\varphi = \arg\min_{\varphi} E_{\mathbf{X}^2}\left[\left\|\varphi(\mathbf{Y}^2) - \mathbf{X}^2\right\|_2^2\right] \tag{2}$$

From this formulation, we can see that the task here is to find a function φ that is the best approximation of η^{-1}.

The task of speech enhancement can be realized using DA, whose architecture is shown in Figure 2 [7].

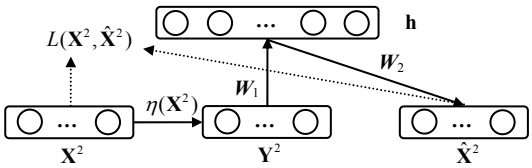

Figure 2: *Denoising auto-encoder (DA) architecture*

DA is defined by the following formulas:

$$\mathbf{h} = \sigma\left(\mathbf{W}_1\mathbf{Y}^2 + \mathbf{b}_1\right) \tag{3}$$

$$\hat{\mathbf{X}}^2 = \mathbf{W}_2\mathbf{h} + \mathbf{b}_2 \tag{4}$$

where $\sigma(x)=(1+exp(-x))^{-1}$ is the sigmoid activation function which is applied element-wise to the input vectors, $\mathbf{h}$ is the activation of hidden layer, $\hat{\mathbf{X}}^2$ is an approximation of the clean power spectrum $\mathbf{X^2}$. $\mathbf{W}_1$ and $\mathbf{W}_2$ are the weight matrices of hidden and output layers, respectively. $\mathbf{b}_1$ and $\mathbf{b}_2$ are the bias parameters of hidden and output layers, respectively.

When it is adopted for unsupervised pre-training, feature learning [6], and image denoising [7], DA is usually trained with various optimization methods to minimize the following reconstruction loss function:

$$L_{rec}(\mathbf{X}^2, \hat{\mathbf{X}}^2) = \left\|\mathbf{X}^2 - \hat{\mathbf{X}}^2\right\|_2^2 \tag{5}$$

In speech signal processing, the same reconstruction errors in different frequency bands usually have different impacts on speech quality. In consideration of this fact, a novel denoising auto-encoder model with weighted reconstruction loss function, which is referred to as Weighted Denoising Auto-encoder (WDA), is proposed in this paper.

The weighted reconstruction loss function is expressed as:

$$L_w(\mathbf{X}^2, \hat{\mathbf{X}}^2) = \left\|\mathbf{F}_w \circ \left(\mathbf{X}^2 - \hat{\mathbf{X}}^2\right)\right\|_2^2 \tag{6}$$

where $\mathbf{F}_w$ is the weighting function, " $\circ$ " stands for the element-wise product of two vectors.

The Stochastic Back-Propagation (SBP) algorithm [9] is applied to train WDA, so it is necessary to derive the relevant gradients of (6).

For the output layer of WDA, the gradient of L_w with respect to $\mathbf{u}_2 = \mathbf{W}_2\mathbf{h} + \mathbf{b}_2$ can be expressed as (for the sake of brevity, the loss function in (6) is simply denoted as L_w):

$$\frac{\partial L_w}{\partial \mathbf{u}_2} = 2\mathbf{F}_w \circ \left(\hat{\mathbf{X}}^2 - \mathbf{X}^2\right) \tag{7}$$

Then, the update rules for the weight and bias parameters in the output layer of WDA are expressed as:

$$\begin{aligned}\Delta\mathbf{W}_2 &= \varepsilon\frac{\partial L_w}{\partial \mathbf{W}_2} = \varepsilon\frac{\partial L_w}{\partial \mathbf{u}_2}\frac{\partial \mathbf{u}_2}{\partial \mathbf{W}_2} \\ &= 2\varepsilon\mathbf{F}_w \circ (\hat{\mathbf{X}}^2 - \mathbf{X}^2)\mathbf{h}^T \\ &= 2\boldsymbol{\varepsilon}_w \circ (\hat{\mathbf{X}}^2 - \mathbf{X}^2)\mathbf{h}^T\end{aligned} \tag{8}$$

$$\begin{aligned}\Delta\mathbf{b}_2 &= \varepsilon\frac{\partial L_w}{\partial \mathbf{b}_2} = \varepsilon\frac{\partial L_w}{\partial \mathbf{u}_2}\frac{\partial \mathbf{u}_2}{\partial \mathbf{b}_2} \\ &= 2\varepsilon\mathbf{F}_w \circ (\hat{\mathbf{X}}^2 - \mathbf{X}^2) \\ &= 2\boldsymbol{\varepsilon}_w \circ (\hat{\mathbf{X}}^2 - \mathbf{X}^2)\end{aligned} \tag{9}$$

where ε is the fixed learning rate, $\boldsymbol{\varepsilon}_w = \varepsilon\mathbf{F}_w$ is the equivalent learning rate in WDA model.

The update rules for the parameters of hidden layer can be derived similarly by the chain rule.

From (8) and (9) we can see that, when the same learning rate ε is adopted for all the dimensions of speech feature, the introduction of weighting function $\mathbf{F}_w$ makes it possible to choose a proper learning rate for each dimension of the feature in the specified application.

In this paper, we use a linear weighting function defined as:

$$F_w(i) = \frac{N_{band} - i}{N_{band}},\ i = 0,1,...,N_{band} - 1 \tag{10}$$

where N_{band} is the number of sub-bands, i is the sub-band index.

By applying the weighting function which decreases with the increase of frequency, the training of WDA is focused on reducing the reconstruction loss in the low frequency band which is more important for the perceptual quality, and the errors in high frequency bands are deemphasized to make the training process more stable.

The spectrum of speech signal is divided into N_{band} = 63 non-uniform sub-bands as described in [10], and the sub-band logarithmic power spectrum is used as the feature of WDA.

In our experiment, WDA with 300 hidden units is trained using the database that consist of noisy speech in white noise with the SNR of 6 dB, 12 dB and 18 dB, and the clean speech materials are selected from NTT database of 8 languages. The weighted reconstruction loss for the training and test sets are illustrated in Figure 3.

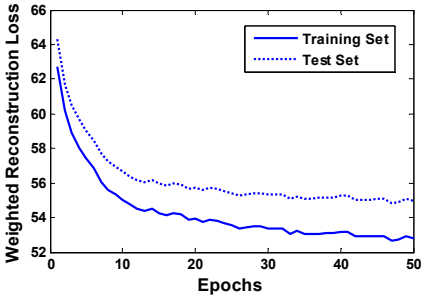

Figure 3: *Weighted reconstruction loss for training and test sets*

From Figure 3, we can see that, the weighted reconstruction errors of training and testing sets are reduced evidently during the 50 epochs of training procedure, and no over-fitting phenomenon has occurred.

An example of power spectrum estimation is illustrated in Figure 4. The speech material is corrupted by white noise with the SNR of 6 dB, and it is selected from the test data set.

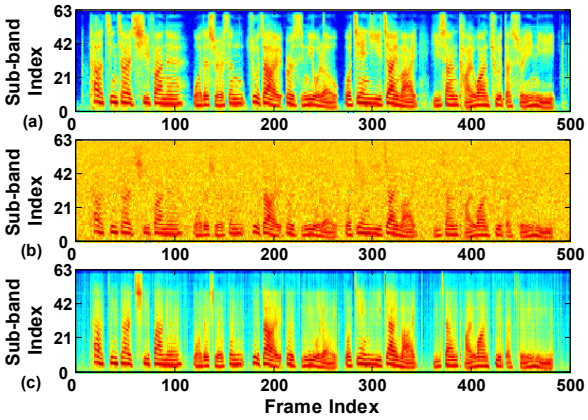

Figure 4: *Power spectrum comparison, (a) clean speech, (b) noisy speech (white noise, SNR = 6dB), (c) enhanced speech.*

From Figure 4, we can see that, the speech components are well preserved by the clean power spectrum estimation method based on WDA, and some of the weak speech components are reconstructed at the same time.

In real-time speech enhancement, it is important to note that, the method of input data normalization is crucial to the performance of WDA. In the applications of pattern recognition like speech recognition, several speech features are used to form a data batch. And the batch of features is normalized and processed at one time. To solve this problem, an adaptive normalization method is proposed in this paper. We use a data buffer for speech segments with the length of 70 frames. If the current frame is classified as speech by the Log Likelihood Ratio (LLR) based VAD method [11], the feature of this frame will be added to this buffer, and the last feature in the buffer will be removed. The mean and variance calculated in the buffer are used to normalize the feature in current frame. This method makes it possible for WDA to be used in real-time speech enhancement applications.

2.2. SNR Estimation

In order to make full use of the power spectrum of noisy speech, the estimated power spectrum of clean speech from WDA, and the noise power estimate from MCRA algorithm, *a posteriori* SNR controlled recursive averaging (PCRA) method for the estimation of the *a priori* SNR is proposed in this paper. The *a posteriori* SNR is used to estimate the speech presence probability in each sub-band, which is utilized to control the updating rate of the *a priori* SNR estimation.

The *a posteriori* SNR is first calculated as:

$$\gamma(\lambda,i) = \max\left(Y^2(\lambda,i)/\hat{\sigma}_d^2(\lambda,i),1\right) \tag{11}$$

where λ is the frame index, i is the sub-band index, $Y^2(\lambda,i)$ is the noisy power spectrum in the i^{th} sub-band of frame λ, $\hat{\sigma}_d^2(\lambda,i)$ is the corresponding noise power spectrum estimate from MCRA algorithm.

The *a posteriori* SNR $\gamma(\lambda, i)$ is smoothed along the time using the following relationship:

$$\overline{\gamma}(\lambda,i) = \alpha_\gamma \overline{\gamma}(\lambda-1,i) + (1-\alpha_\gamma)\gamma(\lambda,i) \tag{12}$$

where $\alpha_\gamma = 0.8$ is the smoothing factor of the *a posteriori* SNR. The fast fluctuations in $\gamma(\lambda,i)$ can be removed by the first order recursive averaging in (12).

Comparing $\overline{\gamma}(\lambda,i)$ with the predefined threshold T_γ, if it is larger than the threshold, the speech presence indicator $I(i)$ is set to one, otherwise, $I(i) = 0$.

Then the speech presence probability is calculated as:

$$p(\lambda,i) = \alpha_p p(\lambda-1,i) + (1-\alpha_p)I(i) \tag{13}$$

where $\alpha_p = 0.95$ is the smoothing factor.

Next, the smoothing factor α_ξ for the *a priori* SNR estimation is determined by the speech presence probability:

$$\alpha_\xi(i) = \alpha_{\xi\min} + \left(1-p(\lambda,i)\right)\left(\alpha_{\xi\max} - \alpha_{\xi\min}\right) \tag{14}$$

where $\alpha_{\xi\max}$ and $\alpha_{\xi\min}$ are the maximum and minimum values of the smoothing factor α_ξ.

Finally, the *a priori* SNR is estimated as:

$$\hat{\xi}(\lambda,i) = \alpha_\xi(i)\hat{\xi}(\lambda-1,i) + \left(1-\alpha_\xi(i)\right)\left[\beta\frac{\hat{X}_{WDA}^2(\lambda,i)}{\hat{\sigma}_d^2(\lambda,i)} + (1-\beta)\left(\gamma(\lambda,i)-1\right)\right] \tag{15}$$

where β is the weighting factor, $\hat{X}_{WDA}^2(\lambda,i)$ is the power spectrum estimate of clean speech from WDA model in the i^{th} sub-band of frame λ. The *a priori* SNR estimation in (15) is composed of three parts. The first part $\hat{\xi}(\lambda-1,i)$ is the estimate in the previous frame. The second part $\hat{X}_{WDA}^2(\lambda,i)/\hat{\sigma}_d^2(\lambda,i)$ is defined by the estimate of clean power spectrum from WDA and the noise estimation by MCRA method. And the third part is $\gamma(\lambda,i)-1$, which is the Maximum Likelihood estimate of the *a priori* SNR.

The estimation of clean power spectrum using WDA can reconstruct some weak speech components which are over-attenuated by the traditional method, but due to the poor smoothness between the adjacent frames, the direct use of $\hat{X}_{WDA}^2(\lambda,i)$ may result in annoying musical noise in noise period. In the proposed method, through the proper weighting between the WDA based estimate and ML estimate of the *a priori* SNR, and the adaptive smoothing between adjacent frames, a compromise between the preservation of speech components and the reduction of musical noise can be reached.

2.3. Wiener Filtering

In the proposed method, the frequency domain Wiener filter [2] is adopted as the STSA estimator, which can be expressed as:

$$G(\lambda,i) = \frac{\hat{\xi}(\lambda,i)}{1+\hat{\xi}(\lambda,i)} \tag{16}$$

where $G(\lambda, i)$ is the spectral gain function for the i^{th} sub-band in frame λ, $\hat{\xi}(\lambda,i)$ is the estimate of *a priori* SNR.

3. Performance Evaluation

The performance evaluation is carried out under the standard of ITU-T G. 160 [12]. The purpose of this test is to evaluate the performance of speech enhancement in terms of the amount of noise reduction, the SNR improvement and the objective speech quality.

In this test, the clean speech samples are selected from NTT database. The noise signals are chosen from ITU noise database. The sampling rate of speech and noise is 16 kHz.

In this paper, the reference algorithm is the frequency domain Wiener filter with Decision-Directed (DD) approach [3] for SNR estimation (referred to as Wiener + DD). The proposed method is referred to as Wiener + WDA.

Four types of noise are used in this test, including white, street, car interior and babble. They can be classified into 3 categories. The first one is the full-band distributed noise like white noise. The second one is the low-frequency distributed noise like street and car interior noise. And the third one is the speech-like noise, which is babble in this test. One WDA model is trained for each noise category. Clean and noisy speech materials with the length of about 1 hour are used as the training data. Three SNR conditions, 6 dB, 12 dB and 18 dB, are included in the training process.

3.1. Noise Reduction Test in White Noise

This test is used to ensure that the speech enhancement method under test could provide specified amount of noise reduction, and the change of speech level is constrained in the acceptable range.

Q_m is the specified level of noise reduction. Three parameters, including Q_{n1}, Q_{n2} and Q_s are calculated in this test. Q_{n1} and Q_{n2} are the noise reduction factors in the noise periods. Q_s is the difference of speech level before and after noise reduction. The larger values of Q_m, Q_{n1} and Q_{n2}, and the value of Q_s that is close to zero correspond to better noise reduction performance. The test results are summarized in Table 1.

Table 1. *The results of noise reduction test in white noise*

Enhancement Method	Q_m (dB)	Q_{n1} (dB)	Q_{n2} (dB)	Q_s (dB)
Wiener + DD	26.02	26.12	26.13	0.27
Wiener + WDA	26.03	26.13	26.00	0.19

From the test results, we can see that the proposed method could provide similar amount of noise reduction to the reference method, while the attenuation of speech components (reflected in Q_s) is relatively smaller.

3.2. Noise Reduction Test in Colored Noise

This test is designed to measure the ability of noise reduction and SNR improvement, and the effect on the level of speech signal in colored noise. There are three parameters, including SNR Improvement (SNRI), Total Noise Level Reduction (TNLR), and SNRI to NPLR Difference (DSN). Here, NPLR stands for Noise Power Level Reduction. The larger SNRI, the smaller TNLR and the value of DSN that is close to zero correspond to the better performance.

The test is performed in street noise and car interior noise, with the SNR of 6 dB, 12 dB and 18 dB, respectively. The test results are averaged over all the test conditions, and are shown in Table 2.

Table 2. *The results of noise reduction test in colored noise*

Enhancement Method	SNRI (dB)	TNLR (dB)	DSN (dB)
Wiener + DD	12.30	-20.50	-0.81
Wiener + WDA	11.95	-19.99	-0.65

From Table 2, the SNRI and TNLR of the proposed method are slightly smaller than the reference method, but the distortion of speech level is lower at the same time.

3.3. Objective Speech Quality Test

This test is used to measure the objective quality of enhanced speech. The test method is not specified in ITU-T G.160. In this paper, Perceptual Evaluation Speech Quality (PESQ) [13] scores are adopted. The test is carried out in four noise types, including babble, street, car interior and white noises. The SNR conditions of 6 dB, 12 dB and 18 dB are used in this test. The test results are illustrated in Figure 5.

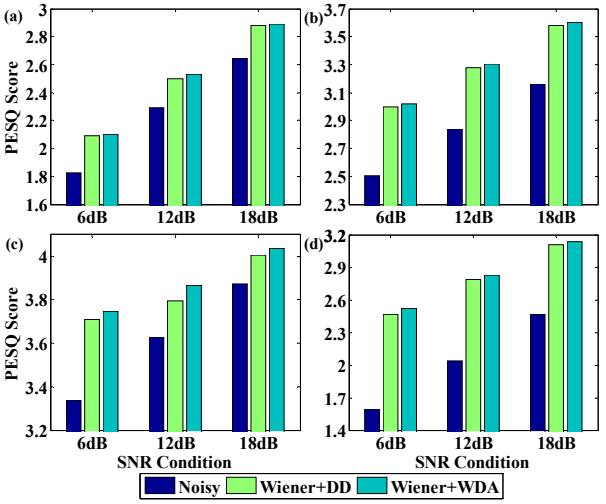

Figure 5: *The results of objective speech quality test (a) Babble, (b) Street, (c) Car interior, (d) White*

From the test results in Figure 5, we can see that, in comparison with the reference method, the proposed method could achieve better objective speech quality in all the noise types and SNR conditions. The quality improvement of the proposed method is mainly due to the better preservation of speech components and less musical noise.

4. Conclusions

Through the introduction of weighted reconstruction loss function, a Weighted Denoising Auto-encoder (WDA) is proposed in this paper. WDA is used to model the relationship between the clean and noisy power spectrums of speech signal. The estimated clean power spectrum from WDA is used in the *a Posteriori* SNR Controlled Recursive Averaging (PCRA) approach for the estimation of the *a priori* SNR. And the enhanced speech is obtained by the frequency domain Wiener filter. In comparison with the Wiener filter with decision-directed approach for SNR estimation, the proposed method can provide similar amount of noise reduction with smaller distortion on speech level, and the objective speech quality is improved at the same time.

5. Acknowledgements

This work was supported by the Beijing Natural Science Foundation Program and Scientific Research Key Program of Beijing Municipal Commission of Education (No. KZ201110005005), the National Natural Science Foundation of China (Grant No. 61072089).

6. References

[1] Boll, S. F., “Suppression of Acoustic Noise in Speech using Spectral Subtraction”, IEEE Trans. Acoust. Speech Signal Process., 27(2): 113-120, 1979.

[2] Lim, J. and Oppenheim, A. V., “Enhancement and Bandwidth Compression of Noisy Speech”, IEEE Proc., 67(12): 1586-1604, 1979.

[3] Ephraim, Y. and Malah, D., “Speech Enhancement using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator”, IEEE Trans. Acoust. Speech Signal Process., 32(6): 1109-1121, 1984.

[4] Donoho, D. L., “De-noising by Soft-Thresholding”, IEEE Trans. Inf. Theory, 41(3): 613-627, 1995.

[5] Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P., “Extracting and Composing Robust Features with Denoising Autoencoders”, in Proc. Art. Intell. Statis., pp. 1-8, 2007.

[6] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P. A., “Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion”, J. Math. Learn. Res. 11: 3371-3408, 2010.

[7] Xie, J., Xu, L., and Chen, E., “Image Denoising and Inpainting with Deep Neural Networks”, in Proc. Neural Information Processing Systems, pp. 1-9, 2012.

[8] Cohen, I. and Berdugo, B., “Noise Esitmation by Minima Controlled Recursive Averaging for Robust Speech Enhancement”, IEEE Signal Process. Lett., 9(1): 12-15, 2002.

[9] Lecun, Y., Bottou, L., Bengio, Y. and Haffner, P., “Gradient based Learning Applied to Document Recognition”, IEEE Proc., 86(11): 2278-2324, 1998.

[10] 3GPP2, “Enhanced Variable Rate Codec, Speech Service Options 3, 68, 70, 73 for Wideband Spread Spectrum Digital Systems”, 2010.

[11] Sohn, J., Kim, N. S. and Sung, W., “A Statistical Model-based Voice Activity Detection”, IEEE Signal Process. Lett., 6(1): 1-3, 1999.

[12] ITU-T Rec. G.160, “Voice Enhancement Devices for Mobile Networks”, 2005.

[13] ITU-T Rec. P.862, “Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment for Narrow-Band Telephone Networks and Speech Coders”, 1996.