Dynamic Noise Aware Training for Speech Enhancement Based on Deep Neural Networks

Yong Xu[1], *Jun Du*[1], *Li-Rong Dai*[1] *and Chin-Hui Lee*[2]

[1]National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China
[2]School of Electrical and Computer Engineering, Georgia Institute of Technology
xuyong62@mail.ustc.edu.cn, {jundu, lrdai}@ustc.edu.cn, chl@ece.gatech.edu

10.21437/Interspeech.2014-571

Abstract

We propose three algorithms to address the mismatch problem in deep neural network (DNN) based speech enhancement. First, we investigate noise aware training by incorporating noise information in the test utterance with an ideal binary mask based dynamic noise estimation approach to improve DNN's speech separation ability from the noisy signal. Next, a set of more than 100 noise types is adopted to enrich the generalization capabilities of the DNN to unseen and non-stationary noise conditions. Finally, the quality of the enhanced speech can further be improved by global variance equalization. Empirical results show that each of the three proposed techniques contributes to the performance improvement. Compared to the conventional logarithmic minimum mean squared error speech enhancement method, our DNN system achieves 0.32 PESQ (perceptual evaluation of speech quality) improvement across six signal-to-noise ratio levels ranging from -5dB to 20dB on a test set with unknown noise types. We also observe that the combined strategies can well suppress highly non-stationary noise better than all the competing state-of-the-art techniques we have evaluated.
Index Terms: Speech enhancement, deep neural networks, noise aware training, ideal binary mask, non-stationary noise

1. Introduction

Speech enhancement has been widely used in many real-world applications, such as automatic speech recognition (ASR), mobile communication and hearing aids [1]. Considering the process of noise corruption on speech is very complicated, the enhancement performance is still unsatisfactory and many issues should be explored.

Various speech enhancement approaches have been proposed, such as spectral subtraction [2], Wiener filtering [3], minimum mean squared error (MMSE) estimation [4, 5] and optimally-modified log-spectral amplitude (OM-LSA) speech estimator (e.g., [6, 7]). In most of these algorithms, it is assumed that an estimate of the noise spectrum is available [19]. The optimal noise estimate in traditional methods (e.g., [6]) is usually updated by averaging the noisy speech power spectrum using time and frequency dependent smoothing factors, which are adjusted based on the estimated speech presence probability in individual frequency bins (e.g., [6], [19]). Nonetheless, its noise tracking capacity is limited for highly non-stationary noise cases, and it tends to distort the speech component in mixed signals if it is tuned for a better noise reduction. Even so, they are on-line algorithms and the dynamic noise information of the testing utterance is well estimated and utilized.

In developing deep learning techniques (e.g., [9, 10]), a deep architecture was adopted to model the complicated relationship between the noisy speech and the clean speech (e.g., [11, 12, 15, 16]). We have also introduced a speech enhancement framework based on DNNs taking advantage of the abundant acoustic context information and large training data [13], and it was shown to achieve better generalization to new speakers, different SNRs and even other languages, etc. Although these mapping functions can be effective to deal with the seen noisy conditions, the evaluation on the mismatch noise types was not extensively investigated. Yet a large number of different noise environments could be included in the training set to address this mismatch problem. In [17], many different kinds of noise types were used to train DNNs to predict the ideal binary mask (IBM), and robustness to unseen noise types was demonstrated. However, the IBM-based speech separation method might improve the intelligibility but not the speech quality [18]. The stacked denoising autoencoder (SDA) trained in dozens of noise types could also well generalize to new noise conditions [14]. Therefore one advantage of DNN-based speech enhancement method is that the relationship between the noisy speech and the clean speech could be well learned from the multi-condition data off-line. In addition, if the noise information could be estimated and given to DNNs as an additional cue, the mismatch problem could be alleviated to a great extent.

In [20], a static noise aware training (NAT) technique was firstly proposed to help the DNN to suppress the noise interference for noise robust speech recognition. It assumes the noise is stationary and uses a noise estimate that is fixed over the utterance [20], which is not the case in practice considering the fast changing characteristics of the noise. In this paper, we try to estimate the noise signal in a dynamic manner and help the DNN to separate the clean speech from the mixture signal. An IBM-based noise estimation method is proposed and then feed it into the DNN in an informing way.

2. System Overview

A block diagram of the proposed speech enhancement framework is illustrated in Fig. 1. A DNN is adopted as the mapping function from noisy to clean speech features. Our baseline system [13] is constructed in two stages. In the training stage, a DNN-based regression model was trained using the log-power spectral features from pairs of noisy and clean speech data. As for the DNN training, as in [27], we first perform pre-training of a deep generative model with the log-power spectra of noisy speech by a stacking of multiple restricted Boltzmann machines (RBMs) [9]. Then the back-propagation

14 – 18 September 2014, Singapore

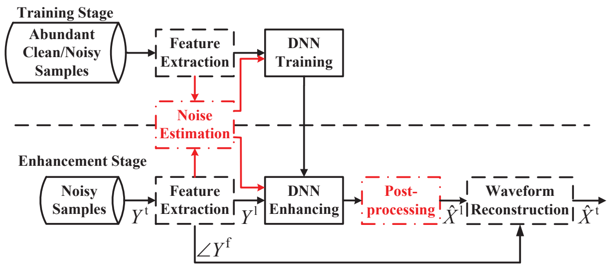

Figure 1: *A block diagram of the proposed DNN-based speech enhancement system.*

algorithm with the MMSE-based object function between the log-power spectral features of the estimated and the reference clean speech is adopted to train the DNN. This corresponds to the perceptually-motivated log-spectral amplitude estimator [5]. A stochastic gradient descent algorithm is performed in mini-batches with multiple epochs to improve learning convergence as follows,

$$Er = \frac{1}{N}\sum_{n=1}^{N}\|\hat{\mathbf{X}}_n(\mathbf{Y}_{n-\tau}^{n+\tau},\mathbf{W},\mathbf{b}) - \mathbf{X}_n\|_2^2 + \kappa\|\mathbf{W}\|_2^2. \quad (1)$$

where Er is the mean squared error with a regularization term, $\hat{\mathbf{X}}_n(\mathbf{Y}_{n-\tau}^{n+\tau},\mathbf{W},\mathbf{b})$ and $\mathbf{X}_n$ denote the estimated and reference log-spectral features at sample index n, respectively, with N representing the mini-batch size, $\mathbf{Y}_{n-\tau}^{n+\tau}$ being the noisy log-spectral feature vector where the window size of context is $2 * \tau + 1$, $(\mathbf{W}, \mathbf{b})$ denoting the weight and bias parameters to be learned. And $\|\mathbf{W}\|_2^2 = \sum_{i,j} w_{i,j}^2$, κ is the regularization weighting coefficient to avoid over-fitting. In the enhancement stage, the noisy speech features are processed by the well-trained DNN model to predict the clean speech features. After we obtain the estimated log-power spectral features of clean speech, $\hat{X}^{\mathrm{l}}(d)$, the reconstructed spectrum $\hat{X}^{\mathrm{f}}(d)$ could be obtained using IDFT with the noisy phase. Finally an overlap-add method is used to synthesize the waveform of the estimated clean speech [26].

Compared with the baseline system, we proposed the improved system in Fig. 1. Firstly, we trained the DNNs by a large training data containing more than 100 noise types. This can improve the generalization capacity to unseen noise types. Then the noise estimation module will be discussed in the following sections. Finally to alleviate the over-smoothing problem of the DNN-based speech enhancement, the global variance (GV) equalization is proposed to post-process the DNN output, and it could improve the overall listening quality. Hence, a dimension-independent global equalization factor β can be defined as:

$$\beta = \sqrt{\frac{GV_{\mathrm{ref}}}{GV_{\mathrm{est}}}} \quad (2)$$

Where GV_{ref} and GV_{est} represented the dimension-independent global variance of the reference features and the estimation features, respectively.

$$\hat{X}'(d) = \hat{X}(d) * \beta * v(d) + m(d) \quad (3)$$

where $m(d)$ and $v(d)$ are the d-th component of the mean and variance of the input noisy speech features, respectively. Since the DNN output $\hat{X}(d)$ was in the normalized log-power spectrum domain, the multiplicative factor β was just operated as a exponential factor in the linear spectrum domain. And this exponential factor could effectively sharpen the formant peaks of the recovered speech and suppress the residual noise.

3. Dynamic Noise Aware Training

As the DNN for speech enhancement presented in [13] is off-line trained, the noise information of each utterance was not specifically utilized. To enable this noise awareness, the DNN is fed with the noisy speech samples augmented with an estimate of the noise. In this way, the DNN can use additional on-line noise information to better predict the clean speech. Also the estimated noise could be regarded as a specific code for adaptation, like a speaker code in speaker adaptation [21]. Here the input vector of the DNN is similar to what was adopted in [20] with a noise estimate appended:

$$\mathbf{V}_n = [\mathbf{Y}_{n-\tau}, ..., \mathbf{Y}_{n-1}, \mathbf{Y}_n, \mathbf{Y}_{n+1}, ..., \mathbf{Y}_{n+\tau}, \hat{\mathbf{Z}}_n] \quad (4)$$

In [20], the noise $\hat{\mathbf{Z}}_n$ was fixed over the utterance as:

$$\hat{\mathbf{Z}}_n = \frac{1}{T}\sum_{t=1}^{T}\mathbf{Y}_t \quad (5)$$

where the noise $\hat{\mathbf{Z}}_n$ is estimated using the first T frames. We call this method as the static noise aware training, denoted as **SNAT**. While investigating dynamic noise aware training in DNNs, an obvious method is to replace the noise variable $\hat{\mathbf{Z}}_n$ in Eqs. (4)-(5) by the dynamic noise information using the conventional MMSE-based noise estimation method [22] at each frame, which is represented as **DNAT1**. However, some non-linear distortion exists in the estimated noise spectrum in DNAT1, which might lead to much more difficult for the DNN learning. Inspired by the work in [24] where the traditional MMSE-based estimator was improved by incorporating masking properties of the auditory system, two kinds of real noise estimation via IBM will be presented below and the related framework is illustrated in Fig. 2.

3.1. Direct IBM-based noise estimation

The estimation of the IBM is suggested as a primary goal of computational auditory scene analysis (CASA) [18]. The IBM is a time-frequency (T-F) binary mask with value 0 standing for the noise dominance and value 1 representing the speech dominance. The direct IBM estimator is obtained by training a DNN, denoted as IBM-DNN. Similar to [28], the IBM-DNN is trained on the noisy log-power spectra to predict the desired outputs across all frequency bands, and the mean squared error (MSE) is used as the cost function. The sigmoid activation functions are used in the output layer considering the IBM range $\{0, 1\}$. The label information could be constructed under the definition of the IBM [28] for training the IBM-DNN.

Note that although the label information is binary in training, the DNN will give the posterior probabilities in testing to indicate the possibility of being noise-dominant or speech-dominant at the certain T-F unit. Since the real noise information is expected to be estimated from the noisy spectra, a threshold γ is set to make a decision of the binary value as follow,

$$\widehat{\mathrm{IBM}}_n(d) = \begin{cases} 0 & \text{posterior}_n(d) < \gamma \\ 1 & \text{otherwise} \end{cases} \quad (6)$$

where γ belongs to [0, 1]. Then the estimated log-power noise

Figure 2: *A block diagram of the proposed two stages of DNAT based on the dynamic noise estimation via IBM.*

spectrum $\hat{N}_n(d)$ could be calculated as follows,

$$\hat{N}_n(d) = \begin{cases} Y_n(d) & \widehat{\mathrm{IBM}}_n(d) = 0 \\ -50 & \widehat{\mathrm{IBM}}_n(d) = 1 \end{cases} \tag{7}$$

where value -50 is the minima log-power spectrum in the overlap-add waveform reconstruction algorithm [26]. This method is denoted as **DNAT2**.

3.2. IBM-based noise estimation through post-processing

Different from DNAT2 with the IBM is directly predicted by a well trained IBM-DNN, the IBM here is estimated by comparing the baseline DNN with linear output enhanced speech $\hat{X}_n(d)$ with the noisy speech $Y_n(d)$ as follows,

$$\alpha = \frac{\exp(\hat{X}_n(d))}{\exp(Y_n(d))} \tag{8}$$

$$\widehat{\mathrm{IBM}}_n(d) = \begin{cases} 0 & \alpha < \lambda \\ 1 & \text{otherwise} \end{cases} \tag{9}$$

where λ is the threshold to exclude speech from the estimated log-power noise spectrum $\hat{N}_n(d)$ as in Eq. (7). This method is denoted as **DNAT3**. It should be noted that IBM is commonly estimated to completely eliminate the influence of the noise to improve the speech intelligibility in CASA [18], on the contrary, here it was adopted to obtain the real noise information for helping the DNN fine-tuning.

4. Experimental Results and Analysis

4.1. Experimental configurations

In [13], only four noise types, namely *AWGN*, *Babble*, *Restaurant* and *Street*, from the Aurora2 database [30] were used as the noise signals for synthesizing the noisy speech. In this study we increased the number of noise types to 104 with another 100 environmental noises [33][1]. The clean speech data was still derived from the TIMIT database [31]. All 4620 utterances from the training set of the TIMIT database were corrupted with the abovementioned 104 noise types at six levels of SNR, i.e., 20dB, 15dB, 10dB, 5dB, 0dB, and -5dB, to build a multi-condition training set, consisting of pairs of clean and noisy speech.

We randomly selected part of them to construct a 10-hour training subset with 11550 utterances. Another 200 randomly selected utterances from the TIMIT test set were used to construct the test set for each combination of noise types and SNR levels. As we only conducted the evaluation of mismatched noise types in this study, 3 other unseen noise types[2], from the Aurora2 database [30] and the NOISEX-92 corpus [25], were used for testing. An improved version of OM-LSA [6, 7], denoted as **LogMMSE**, were used for performance comparison. All of the waveforms were down-sampled to 8KHz. Perceptual evaluation of speech quality (PESQ) [32] was used as a compressive objective measure. The regularization weighting coefficient κ in Eq. (1) was 0.00001. The first $T = 6$ frames of each utterance were used for noise estimation in SNAT. Mean and variance normalization was applied to the input and target feature vectors of the DNN. All DNN configurations were fixed at $L = 3$ hidden layers, 2048 units at each hidden layer, and 11-frame input. γ in Eq. (6) was set to 0.4 and λ in Eq. (9) was set to 0.1. Other detail of the setup can be found in [13].

4.2. Evaluations of different NAT methods

In Table 1, we compare the average PESQ results among noisy, LogMMSE, DNN baseline, SNAT, DNAT1, DNAT2, DNAT3, DNAT3 improved by GV equalization (denoted as DNAT3-GV) and two oracle DNAT experiments assuming the real noise spectrum or real IBM information was known on the test set at different SNRs of the three unseen noise environments, namely *Exhibition*, *Destroyer engine* and *HF channel*.

4.2.1. SNAT and DNAT1 experiments

The DNN baseline trained with only 10-hour data of 104 noise types is better than the LogMMSE method at low SNRs (less than 10dB), especially at SNR=-5dB with PESQ going from 1.38 to 1.71. Here the abundant noise types in training set were crucial to improve the generalization capacity to unseen noise conditions. After improved by SNAT, the DNN model outperformed the LogMMSE method at all SNR levels. And SNAT was more beneficial at high SNRs. This is reasonable because SNAT assumed that the noise was unchanging across all frames of the utterance, which might be not the case if the noise was non-stationary and at high level. DNAT1 achieved almost the same performance with SNAT or even a little worse at high SNRs, e.g., PESQ going down from 3.38 to 3.35 at SNR=20dB. We conjecture that conventional MMSE-based noise estimation would introduce some non-linear distortion which might be even more difficult to learn for DNNs.

4.2.2. Oracle experiments

Then we examine in detail the last two columns in Table 1 the two oracle DNAT experiments assuming the real noise spectrum or the real IBM information was known. It is interesting to find that the oracle IBM DNAT is much better than the oracle noise DNAT, especially at low SNRs, e.g., PESQ jumping from 2.28 to 2.74 at SNR=-5dB. Fig. 3 shows an utterance example corrupted by *Exhibition* noise at SNR=0dB. It was enhanced by oracle noise DNAT (upper right) and oracle IBM DNAT (upper left), respectively. The oracle IBM DNAT (upper left) achieved the best PESQ and its spectrogram was much better with clearer harmonic components and less residual noise. This indicates that it is unnecessary to reduce noise at the speech-dominant T-F units and human may not perceive the noise component when the speech energy is higher than the noise energy [18].

[1]The 104 noise types are N1-N17: Crowd noise; N18-N29: Machine noise; N30-N43: Alarm and siren; N44-N46: Traffic and car noise; N47-N55: Animal sound; N56-N69: Water sound; N70-N78: Wind; N79-N82: Bell; N83-N85: Cough; N86: Clap; N87: Snore; N88: Click; N88-N90: Laugh; N91-N92: Yawn; N93: Cry; N94: Shower; N95: Tooth brushing; N96-N97: Footsteps; N98: Door moving; N99-N100: Phone dialing. To compare with the results of [13], N101: AWGN, N102: Babble, N103: Restaurant, N104: Street, were used.

[2]The 3 unseen environment noises for evaluation are Exhibition, Destroyer engine and HF channel. The first one noise is from the Aurora2 database and the others are collected from the NOISEX-92 corpus.

Table 1: Average PESQ comparison on the test set at different SNRs of the three unseen noise environments, among: Noisy, LogMMSE approach, DNN baseline, SNAT, DNAT1, DNAT2, DNAT3, DNAT3 improved by GV equalization (denoted as DNAT3-GV) and two oracle DNAT experiments supposed the real noise spectrum information or the real IBM information was known.

	Noisy	LogMMSE	Baseline	SNAT	DNAT1	DNAT2	DNAT3	DNAT3-GV	Oracle noise DNAT	Oracle IBM DNAT
SNR20	2.88	3.37	3.33	3.38	3.35	3.43	3.46	**3.60**	3.80	3.95
SNR15	2.55	3.07	3.05	3.09	3.08	3.17	3.19	**3.31**	3.58	3.78
SNR10	2.22	2.73	2.75	2.78	2.76	2.87	2.90	**2.99**	3.31	3.59
SNR5	1.90	2.32	2.42	2.45	2.45	2.53	2.57	**2.65**	3.00	3.35
SNR0	1.61	1.87	2.07	2.09	2.08	2.18	2.22	**2.27**	2.66	3.07
SNR-5	1.37	1.38	1.71	1.73	1.73	1.77	1.82	**1.86**	2.28	2.74
Ave	2.09	2.46	2.55	2.59	2.58	2.66	2.69	**2.78**	3.11	3.41

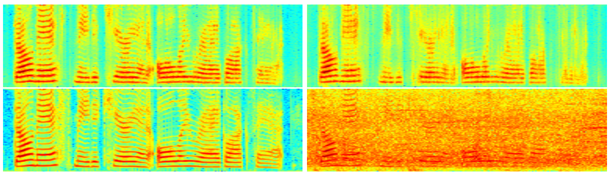

Figure 3: *Spectrograms of an utterance tested on* Exhibition *noise at SNR = 0dB: oracle IBM DNAT (upper left, PESQ=2.76), oracle noise DNAT (upper right, PESQ=2.24), clean speech (bottom left, PESQ=4.50), and noisy speech (bottom right, PESQ=1.24).*

4.2.3. *Proposed IBM-based DNAT experiments*

With this analysis, IBM-based DNAT2 and DNAT3 were proposed in Sec. 3. In Table 1, DNAT3 achieved a better performance, e.g., PESQ jumping from 2.75 to 2.90 at SNR=10dB when compared with the DNN baseline. Since we only learned the IBM with the log-power spectra feature as the input, DNAT2 was a little worse than DNAT3. While in [28, 29], a set of complementary features, such as, amplitude modulation spectrogram (AMS), pitch-based features, etc., were adopted to learn the IBM in DNNs. Even so, the DNAT2 is much better than SNAT and the DNN baseline. The best DNAT3 system can be further improved at all SNRs by GV equalization which was more helpful for high SNRs, e.g., PESQ going up from 3.46 to 3.60 at SNR=20dB. The final system outperformed LogMMSE by 0.32 in PESQ on average.

Fig. 4 shows an utterance example corrupted in succession by different unseen noise types at several speech segments. These noise types were *Machine Gun*, *F16*, *Destroyer engine* and *Exhibition*. The DNN-enhanced spectrograms shown in Fig. 4(a)-(b) successfully removed most of the noises while the LogMMSE-enhanced spectrogram shown in Fig. 4(c) failed to remove most of them. This was reasonable as the LogMMSE method predicted the noise in a recursive averaging mode according to previous frames and it was hard to track the potentially dramatic changes in non-stationary noises. However, the DNN model processed the noisy spectrum in a frame-by-frame manner, and the relationship between the clean speech and noise had been learned off-line. The non-stationary noise components were shown in Fig. 4(d) in dashed rectangular boxes. Compared to the SNAT enhanced spectrogram shown in Fig. 4(b), especially for that in the dashed ovals, the DNAT3 enhanced spectrogram shown in Fig. 4(a) could suppress more non-stationary noise and highlight the speech spectrum. Hence DNAT3 could obtain higher PESQ score jumping from 2.57 to 2.76, which indicates that the IBM-based dynamic noise estimation scheme can accurately track the non-stationary noise in the noisy speech.

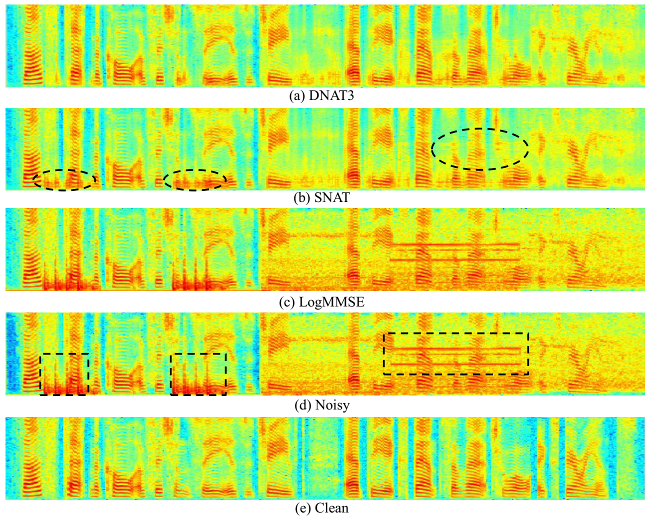

Figure 4: *Spectrograms of an utterance corrupted in succession by different noise types tested on changing noise environments at SNR = 0dB: (a) DNAT2 enhanced (PESQ=2.76), (b) SNAT enhanced (PESQ=2.57), (c) LogMMSE enhanced (PESQ=2.06) (d) noisy (PESQ=2.05), and (e) clean speech (PESQ=4.50).*

5. Conclusion and Future Work

In this paper, different noise aware training schemes were compared in DNN-based speech enhancement. We found that IBM-based real noise estimation and informing strategy were effective to track the change of the noise. Two dynamic noise aware training methods, namely, DNAT2 and DNAT3, were proposed, and the latter achieved a better performance. DNAT3 could be further improved by GV equalization. Moreover, multi-condition training with many kinds of noise types could achieve a good generalization capability to unseen noise environments. Finally the proposed DNN-based system is also powerful to cope with non-stationary noises with quickly changing characteristics. For future work, we will concentrate on accurately estimating the IBM target using some complementary features like in [28, 29]. Therefore the performance of the dynamic noise aware training in DNNs could reach the oracle performance upper bound indicated in Table 1.

6. Acknowledgment

This work was partially supported by the National Nature Science Foundation of China (Grant No. 61273264 and No. 61305002) and the Programs for Science and Technology Development of Anhui Province (Grants No. 13Z02008-4).

7. References

[1] J. Benesty, S. Makino, and J. D. Chen, *Speech Enhancement*, Springer, 2005.

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustic, Speech and Signal Processing*, Vol. 27, No. 2, pp. 113-120, 1979.

[3] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," in *Proc. IEEE*, Vol. 67, No. 12, pp. 1586-1604, 1979.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 32, No.6, pp. 1109-1121, 1984.

[5] Y. Ephraim and D. Malah, "Speech enhancement using minimu mean square log spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 33, No. 2, pp. 443-445, 1985.

[6] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, Vol. 81, No. 11, pp. 2403-2418, 2001.

[7] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. on Speech and Audio Processing*, Vol. 11, No. 5, pp. 466-475, 2003.

[8] E. A. Wan and A. T. Nelson, "Networks for speech enhancement," in *Handbook of Neural Networks for Speech Processing*, Edited by Shigeru Katagiri, Artech House, Boston, 1998.

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

[10] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, Vol. 313, No. 5786, pp. 504-507, 2006.

[11] A. L. Maas, T. M. O'Neil, A. Y. Hannun and A. Y. Ng, "Recurrent neural network feature enhancement: The 2nd CHiME challenge," *In Proceedings The 2nd CHiME Workshop on Machine Listening in Multisource Environments held in conjunction with ICASSP*, pp. 79-80, 2013.

[12] A. L. Maas, Q. V. Le, T. M. O'Neil, O. Vinyals, P. Nguyen and A. Y. Ng, "Recurrent Neural Networks for Noise Reduction in Robust ASR," *Proc. Interspeech*, pp. 22-25, 2012.

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

[14] B.-Y. Xia and C.-C. Bao, "Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification," *Speech Communication*, Vol. 60, pp. 13-29, 2014.

[15] B.-Y. Xia and C.-C. Bao, "Speech enhancement with weighted denoising Auto-Encoder," *Proc. Interspeech*, pp. 3444-3448, 2013.

[16] X.-G. Lu and Y. Tsao and S. Matsuda and C. Hori, "Speech enhancement based on deep denoising Auto-Encoder," *Proc. Interspeech*, pp. 436-440, 2013.

[17] Y. X. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. on Audio, Speech and Language Processing*, Vol. 21, No. 7, pp. 1381-1390, 2013.

[18] D. L. Wang and G. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Hoboken, NJ, USA: Wiley-IEEE Press, 2006.

[19] S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech Communication*, Vol. 48, No. 2, pp. 220-231, 2006.

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

[21] O. Abdel-Hamid and H. Jiang, "Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code," *Proc. ICASSP*, pp. 7942-7946, 2013.

[22] T. Gerkmann, R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. on Audio, Speech, and Language Processing*, Vol. 20, No. 4, pp. 1383-1393, 2012.

[23] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," *Proc. ICASSP*, pp. 4266-4269, 2010.

[24] J. H. Hansen, V. Radhakrishnan and K. H. Arehart, "Speech enhancement based on generalized minimum mean square error estimators and masking properties of the auditory system," *IEEE trans. on Audio, Speech and Language Processing*, Vol. 14, No. 6, pp. 2049-2063, 2006.

[25] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, Vol. 12, No. 3, pp. 247-251, 1993.

[26] J. Du and Q. Huo, "A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions," *Proc. Interspeech*, pp. 569-572, 2008.

[27] L. Deng, M. L. Seltzer, D. Yu, A. Acero, A. R. Mohamed and G. E. Hinton. "Binary coding of speech spectrograms using a deep auto-encoder," *Proc. Interspeech*, pp. 1692-1695, 2010.

[28] Y. X. Wang, A. Narayanan and D. L. Wang, "On training targets for supervised speech separation," *Technical Report OSU-CISRC-2/14-TR05, Department of Computer Science and Engineering, The Ohio State University*, 2014.

[29] Y. X. Wang, K. Han and D. L. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Trans. on Audio, Speech, and Language Processing*, Vo. 21, No. 2, pp. 270-279, 2013.

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

[31] J. S. Garofolo, *Getting started with the DARPA TIMIT CD-ROM: an acoustic phonetic continuous speech database*, NIST Tech Report, 1988.

[32] ITU-T, Recommendation P.862, "Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *International Telecommunication Union-Telecommunication Standardisation Sector*, 2001.

[33] G. Hu, 100 nonspeech environmental sounds, 2004. http://www.cse.ohio-state.edu/pnl/corpus/HuCorpus.html.