2017 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 25–28, 2017, TOKYO, JAPAN

ON GENERATING MIXING NOISE SIGNALS WITH BASIS FUNCTIONS FOR SIMULATING NOISY SPEECH AND LEARNING DNN-BASED SPEECH ENHANCEMENT MODELS

Shi-Xue Wen, Jun Du

University of Science and Technology of China
shixue@mail.ustc.edu.cn, jundu@ustc.edu.cn

Chin-Hui Lee

Georgia Institute of Technology
chl@ece.gatech.edu

ABSTRACT

We first examine the generalization issue with the noise samples used in training nonlinear mapping functions between noisy and clean speech features for deep neural network (DNN) based speech enhancement. Then an empirical proof is established to explain why the DNN-based approach has a good noise generalization capability provided that a large collection of noise types are included in generating diverse noisy speech samples for training. It is shown that an arbitrary noise signal segment can be well represented by a linear combination of microstructure noise bases. Accordingly, we propose to generate these mixing noise signals by designing a set of compact and analytic noise bases without using any realistic noise types. The experiments demonstrate that this noise generation scheme can yield comparable performance to that using 50 real noise types. Furthermore, by supplementing the collected noise types with the synthesized noise bases, we observe remarkable performance improvements implying that not only a large collection of real-world noise signals can be alleviated, but also a good noise generalization capability can be achieved.

Index Terms— speech enhancement, deep neural network, noise generalization, noise basis, objective performance measures

1. INTRODUCTION

Speech enhancement has been a long standing research problem in speech communication for the past several decades (e.g., [1], [2]). The main research efforts in early literature include spectral enhancement [3], [4], and [5] and model-based techniques [6], [7], and [8], all based on an *explicit* distortion model in the time domain as follows:

$$y(l) = x(l) + g \cdot n(l) \tag{1}$$

where g is a noise gain factor, and signals, $y(l)$, $x(l)$, and $n(l)$, represent the l^{th} samples of corrupted noisy speech, clean speech, and additive noise, respectively. Model assumptions of the three signals involved and their relationship in a certain representation domain (e.g., frequency or log-power spectral domain) are then made for any subsequent inferences. They might lead to performance limitations, such as the musical artifact of enhanced speech and a failure to deal with non-stationary noises in real-world noisy speech situations.

To address this problem, data-driven approaches learning speech or noise information as priors, were proposed. Nonnegative matrix factorization (NMF) based speech enhancement [9], [10], [11], and [12] was one widely used method by factorizing noisy speech with the learned speech and noise dictionaries, but NMF needs a large collection of realistic noise. Another broad class was neural network based speech enhancement, which was also the focus of this study. In the 1980s and 1990s, speech enhancement using shallow neural networks (SNNs) as nonlinear filters to predict the clean speech signal in the time or frequency domain has been investigated [13], [14], and [15]. Nevertheless, the performance was not satisfactory due to the limited modeling capability with small-size SNNs. Recently, deep learning with a large training data set has shown its superiority over the conventional approaches and become increasingly popular. Most of them focused on the design of neural network architecture, including DNN [16], [17], [18], and [19], denoising auto-encoder (DAE) (e.g., [20], [21]), recurrent neural network (e.g., [22], [23]), and generative stochastic network (GSN) [24]. One critical issue of deep learning based speech enhancement was the model generalization capability to unseen speech or noise signals. Several strategies have been examined, e.g., augmenting the variety of noise types for the DNN training [16], [18], noise perturbation [25], and unseen noise estimation [26].

The work of [18] trained regression DNNs with more than 100 noise types and demonstrated good generalization capabilities to unseen noise types, but a theoretical analysis on noise generalization was not elaborated. Moreover, a passive collection of noise signals from realistic environments cannot guarantee diversity and completeness needed to have a good set of mixing noises for the DNN training. Therefore, a key motivation here is to illustrate noise generalization in regression-based speech enhancement. First, an empirical proof is introduced to show that the DNN-based approach can achieve a good generalization capability provided that plenty of noise types are involved in training. The main principle is that an arbitrary noise signal segment can be well represented by a linear combination of microstructure noise bases.

This implies that a good enhancement performance can be obtained if the noisy test speech is generated with the combination of noise types seen in the training stage. Accordingly, we propose a novel approach to fundamentally solving the noise generalization problem through generating the noise signals for the DNN training by designing a set of noise bases without using any available real noise types. Our active construction in this manner can well control the diversity and compactness of noise signals. The preliminary experiments demonstrate that the proposed noise generation approach can yield a comparable performances to that using 50 real noise types. Furthermore, by supplementing the already-collected noises with the basis-constructed noises, we observe remarkable performance improvements even without adding any new noise types or samples to the existing and yet small noise set.

978-1-5090-6341-3/17/$31.00

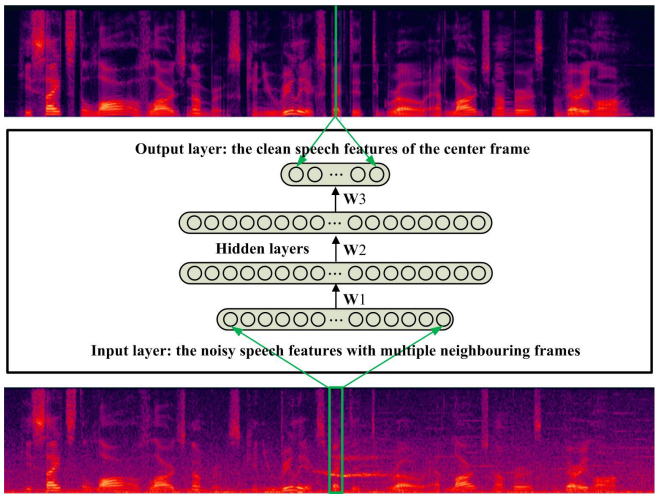


Fig. 1. An illustration of the DNN-based speech enhancement.

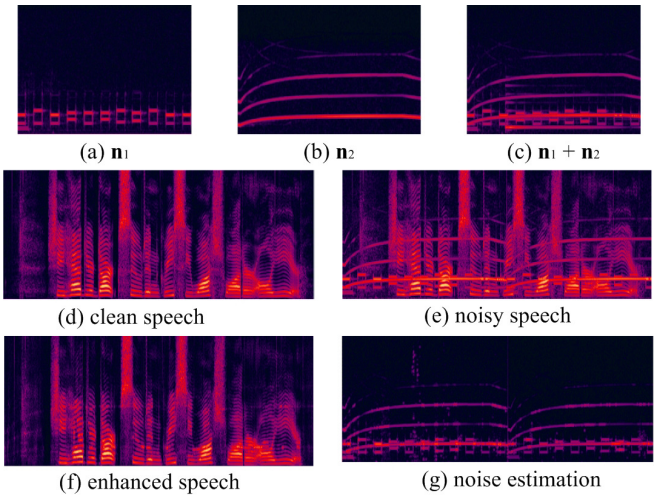


Fig. 2. Spectrograms of an example for DNN-based speech enhancement: (a) $\mathbf{n}_1$ phone dialing noise used for training, (b) $\mathbf{n}_2$ alarm noise used for training, (c) $\mathbf{n}_1$+$\mathbf{n}_2$ as a linear combination of $\mathbf{n}_1$ and $\mathbf{n}_2$ for testing, (d) reference clean speech, (e) noisy speech with $\mathbf{n}_1$+$\mathbf{n}_2$ at 5dB SNR, (f) DNN enhanced speech, (g) noise estimation via [28].

2. ANALYSIS OF NOISE GENERALIZATION

The DNN-based speech enhancement [18] algorithm used in this study is illustrated in Figure 1. The DNN is adopted as a regression model to map the log-power spectra (LPS) of noisy speech to the clean speech LPS features. A collection of time-synchronized clean and noisy utterance pairs based on Eq. (1) are used for the DNN training by minimizing the mean squared error (MSE) between the estimated DNN output and the reference clean LPS features as follows:

$$E = \frac{1}{N}\sum_{n=1}^{N}\sum_{d=1}^{D}(\hat{X}_n^d(\mathbf{W}^l, \mathbf{b}^l) - X_n^d)^2 \tag{2}$$

where E is the mean squared error, $\hat{X}_n^d(\mathbf{W}^l, \mathbf{b}^l)$ and X_n^d denote the enhanced and target LPS features at sample index n and frequency bin d, respectively, with N representing the mini-batch size, D being the size of the LPS feature vector, $(\mathbf{W}^l, \mathbf{b}^l)$ denoting the weights and bias parameters to be learned at the l-th layer of the DNN.

One key point to determine the speech and noise generalization capability in enhancing real-world noisy speech is that the DNN-based regression operation is conducted on a per-frame basis. Such frame-level learning makes a full use of speech and noise information with high-resolution microstructures. The concept "microstructure" denotes a quite small time-frequency region (subband in the frequency axis and few frames in the time axis) of spectrogram which often can not be distinguished by human ears. This implies that unseen noise signals, sounding quite differently in testing, might share similar microstructures to those noises used in the DNN training stage. Furthermore, the acoustic context for the input noisy speech features, namely multiple neighborhood with M frames, not only improves the continuity of enhanced speech, but also distinguishes the speech and noise signals with different statistical properties. In the following subsections, we attempt to provide a theoretical analysis on noise generalization.

2.1. Representation of unseen noise signals

Suppose the frame and shift lengths are L and S samples in the time domain, respectively. Then the sample size of input noisy speech with a M-frame microstructure fed to DNN is $L_{\text{in}} = L + (M-1)\cdot S$. Accordingly, a set of K noise signals $\{\mathbf{n}_k|k = 1, 2, ..., K\}$ are defined, where $\mathbf{n}_k$ is the k^{th} microstructure segment with the same L_{in} dimension as the input segment of the DNN in the time domain. We assume these noise signals are heterogeneous and used for synthesizing the noisy speech in the training stage.

First, we prove that in the testing stage, the arbitrary unseen noise segment with $\mathbf{n}^{\text{us}}$ with dimension L_{in} in the noisy speech can be well represented as a linear combination of the seen noise segments under certain conditions:

$$\mathbf{n}^{\text{us}} = \sum_{k=1}^{K} g_k \mathbf{n}_k \tag{3}$$

where g_k is the gain factor for the noise component $\mathbf{n}_k$. This problem is equivalent to that whether a solution of finding $\{g_k|k = 1, 2, ..., K\}$ to Equation (3) in a vector form exists given $\{\mathbf{n}_k|k = 1, 2, ..., K\}$. The necessary and sufficient conditions for solving this set of linear equations are [27]:

$$R(A) = R(B) \tag{4}$$

where $R(\cdot)$ is the rank operator [27] of a matrix and here

$$A = [\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_K]_{L_{\text{in}}\cdot K} \tag{5}$$

$$B = [\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_K, \mathbf{n}^{\text{us}}]_{L_{\text{in}}\cdot(K+1)} \tag{6}$$

where B is an augmented matrix [27] of A. Obviously, the condition in Equation (4) is not difficult to be satisfied if $K \gg L_{\text{in}}$ and all the noise segments in $\{\mathbf{n}_k|k = 1, 2, ..., K\}$ are heterogeneous, i.e., as many noise types as possible [18].

Second, we experimentally verify that the linear combination of the seen noise segments as defined in Equation (3) could be well eliminated by the DNN when present in the noisy speech, provided that sufficient training data pairs, by mixing each noise base $\mathbf{n}_k$ with

Table 1. Four types of noise bases.

Type	Randomness	Frequency Response
NB_1	Deterministic signal	Single-frequency, subband
NB_2	White Gaussian noise	Subband
NB_3	Color Gaussian noise	Subband
NB_4	White Non-Gaussian noise	Subband

a full coverage of the clean speech signals, could be provided for learning. In Figure 2, an example is shown to demonstrate that a linear combination of two seen noise signals is almost perfectly removed from the noisy speech even such a noise signal is not directly used in the DNN training, which implies that clean speech can be well estimated from noisy speech by the DNN if the noise signal is a linear combination of seen noise types in the training stage. Figure 2(g) also gives the corresponding noise estimation based on a post-processing of the DNN output in [28], which demonstrates that DNN can also accurately estimate the newly constructed unseen noises.

2.2. Practical issues

Based on the above analysis in Sections 2.1, the DNN model can achieve a quite good generalization capability to unseen noise signals. However, in real implementations, several issues might lead to performance limitations. First, clean speech signals, may not be exhaustively searched due to the variabilities of speakers, languages, etc. Second, the amount of the training data pairs is determined by the product space of the clean speech data, noise data, and the mixing factors. This "oracle" size is usually not realizable in an experimental design. Finally, even with a large-scale data collection, containing diversified training data with a variety of noise samples, the issue with local optima in the DNN learning is still inevitable.

3. IMPROVING GENERALIZATION BY NOISE BASES

For conventional deep learning based speech enhancement [16]-[26], the noise signals, passively collected from real environments, are usually required for model training. However, the diversity and compactness of those noise signals cannot be guaranteed. In this study, with the conclusion drawn from Section 2, we aim at constructing a set of noise bases, $\{\mathbf{n}_k|k = 1, 2, ..., K\}$, without incorporating any realistic noise types. The design of noise bases should fundamentally solve the problems of both diversity and compactness. Meanwhile, the noise signals generally have two key features, namely the randomness and frequency response. Therefore four types of noise bases are adopted, as described in Table 1.

First, deterministic signals, denoted as NB_1, are used to simulate the acoustic environmental sounds with special spectral structures, e.g., alarm and phone dialing noises in Figure 2. Two broad classes of frequency responses, namely the single-frequency and frequency subband, are carefully designed. For the single-frequency signal, as shown in Figure 3(a), the basic sinusoidal waveform is adopted:

$$n_{m_1}^{\text{single}}(l) = \sin\left(\frac{\pi m_1 l}{L_1}\right), \ \ l > 0, \ m_1 = 0, 1, \ldots, L_1 \tag{7}$$

where l is the sample index and m_1 is the single frequency index. L_1 represents the number of frequency points uniformly drawing from the speech signal bandwidth (one half of the sampling frequency

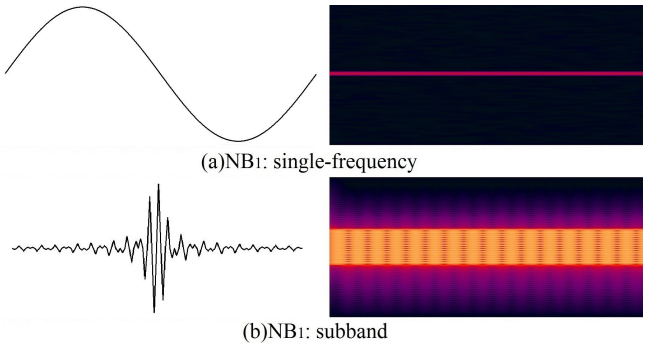


Fig. 3. The waveforms and spectrograms of two examples in NB_1: (a) deterministic single-frequency signal, frequency is 4000 Hz, (b) deterministic subband signal, center frequency is 4000 Hz and bandwidth is 2000Hz.

L_{sf}). Meanwhile, as shown in Figure 3(b), the frequency subband signals are expressed as:

$$n_{m_2,m_3}^{\text{sub}}(l) = \frac{1}{l}\sin\left(\frac{\pi m_2 l}{L_2}\right)\sin\left(\frac{m_3 l}{4\cdot L_3}\right), \ \ l > 0$$

$$m_2 = 1, 2, \ldots, \left\lfloor \frac{L_2}{m_3} \right\rfloor - 1,$$
$$m_3 = \left\lfloor \frac{L_3}{2^m} \right\rfloor, m = 0, 1, \ldots, \lfloor \log_2 L_3 \rfloor \tag{8}$$

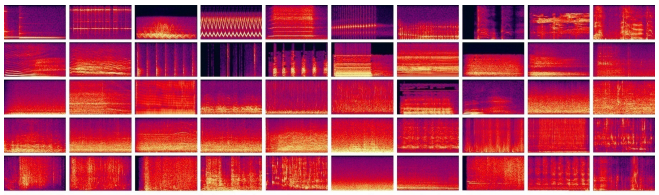
where m_2 and m_3 are the indices of the center frequency and bandwidth for $n_{m_2,m_3}^{\text{sub}}(l)$, respectively. L_2 and L_3 are the maximum numbers of center frequencies and frequency subbands uniformly drawing from the speech signal bandwidth.

Second, artificially generated random noises with different probability distributions are explored. NB_2 represents the most commonly used white Gaussian noise, which follows a Gaussian distribution in the time domain and has a uniform power spectrum density across all frequency bands. In comparison to NB_2, NB_3 employs the "color" Gaussian noises with non-flat spectral profiles, including pink and brown noises. The main difference of NB_4 from NB_2 is non-Gaussian distributions, e.g., the uniform distribution and t-distribution, are complemented. For all the three types of noise bases, not only the original noises with full frequency bands are used, but also D subband noise signals with bandpass filters corresponding to D frequency bins of the LPS feature vector are adopted for each type of noise bases. Finally, the number of noise bases, K, is determined by L_1, L_2, L_3 and D.

In terms of the noise basis concept, our proposed idea is similar to the NMF approach. The noise bases in NMF are learned from the realistic noise training data, which are only one type of compact representations. Nonetheless, the noise bases in this study are originally designed for improving the noise generalization of the DNN model and real noise samples are not necessarily needed for the construction.

4. EXPERIMENTS AND RESULT ANALYSIS

As in [18], the experiments were conducted on the TIMIT corpus [29] and OSU-100 environmental noise database [30]. The noisy utterances were synthesized by adding a noise signal with a specified



(a) real noise types for training

(b) real noise types for testing

Fig. 4. The spectrograms of real noise types.

Table 2. PESQ comparison for unseen noises and seen targets.

System Description	Narrowband Noises		Wideband Noises	
	0dB	5dB	0dB	5dB
Noisy	1.70	2.08	1.85	2.25
RN	3.68	3.86	3.58	3.82
NB_1	3.47	3.78	3.04	3.53
NB_1+NB_2	3.55	3.81	3.32	3.68
NB_1+NB_2+NB_3	3.59	3.82	3.32	3.70
NB_1+NB_2+NB_3+NB_4	3.61	3.86	3.35	3.69

SNR level to a clean speech waveform. The sampling frequency was 16KHz (i.e., L_{sf}=16000). The frame length L was set to 512 while the frame shift S was set to 256. The parameters for noise bases were set as L_1=4096, L_2=160, L_3=80, and D=257. The DNN architecture was 2056-2048-2048-2048-257, where the input layer consisted of 7-frame (M=7) noisy LPS features plus 1-frame noise LPS features for noise-aware training [18], 3 hidden layers were used with 2048 nodes for each layer, and the output layer was 1-frame estimated clean LPS features. The mini-batch size N is 128. More detailed configurations for training and testing can be found in [18].

For the DNN system using real noise samples, 50 noise types were selected for training, as shown in Figure 4(a). The input SNR of training noisy speech ranged from 0dB-10dB uniformly, which was applied for the training of both DNNs with real noises or/and synthesized noises. The generalization to noise levels could be well handled according to [17], which was not the focus of this study. For the test set, 10 narrowband and 10 wideband noises were chosen from the remaining 50 noise types. The narrowband/wideband noises means distribution in spectrum band of noises is narrow/wide, as shown in Figure 4(b). By default, all DNN systems were built with 100 hours of training utterance pairs.

4.1. Experiments on unseen noises and seen targets

First, a series of proof-of-concept experiments, as shown in Table 2, was designed to focus on the noise generalization issue, where only one target clean utterance was adopted for both training and testing. For each test subset (e.g., 10 narrowband noises under 0dB SNR), 500 noisy speech utterances were generated. As target clean speech was seen, all the DNN systems using real noises (RN) and noise bases (NB) yielded very significant PESQ [31] gains over the unprocessed baseline system (Noisy) across all testing cases. By only using the deterministic signals in NB_1, the PESQ scores could be quite close to those of the RN system using 50 real noises for training, especially for the narrowband noises. This confirmed with the assertion we made in Section 2 that a large number of microstructure noise bases led to a good generalization to unseen noises and also justified the motivation for the design of NB_1 in Section 3. The other three types of noise bases (NB_2 to NB_4) using random distributions were good supplements to NB_1, especially for wideband noises. The final NB system (NB_1+NB_2+NB_3+NB_4) achieved comparable PESQ performances with the RN system, demonstrating the effectiveness of the designed noise bases. Please note that the operator "+" here denoted that multiple types of noise bases were adopted and separately mixed with the clean speech, not indicating that all the noise bases were summed up in the time domain.

4.2. Experiments on unseen noises and unseen targets

Second, more realistic experiments for both unseen target speech and unseen noises [18] were designed. In Table 3, all the 4620 utterances in the TIMIT training set were adopted for the DNN training while 500 utterances selected from the TIMIT test set were used as unseen targets for testing. Compared with the results in Table 2, the improvements of all DNN systems over the unprocessed system were less significant due to the more challenging cases when both noise and speech generalization problems should be considered. Moreover, for the unseen targets, only the use of NB_1 could not bring comparable PESQ gains, especially for wideband noises. And the NB_2 to NB_4 played more important roles in improving the noise generalization capability of the NB system for all testing cases. By measuring the PESQ gains over the unprocessed system, the NB system was still comparable to the RN system. More results and demos can be found at this website[1].

Based on the above experiments, it is clear that our proposed noise bases produced reasonable results without leveraging upon any realistic noises and showed a great potential to fundamentally mitigate the noise generalization problem. As a further demonstration, the experimental results by mixing real noises and noise bases were given in Table 4. The overall PESQ, segmental SNR (SegSNR, in dB) [32], and SDR/SIR/SAR (in dB) [33] results were calculated across all noise types (narrowband and wideband noises) and at two input SNR levels (0dB and 5dB). For the 100-hour (100h) training data case, the mixed RN+NB system using 100-hour original speech data with 50-hour real-noise based data and 50-hour noise-basis based data improved PESQ by 0.07, SegSNR by 0.42dB, and SDR/SIR/SAR by 0.1dB/1.5dB/0.34dB over the RN system. By increasing the amount of training data to 200 hours (200h) using 200-hour original speech data with 100-hour real-noise based data and 100-hour noise-basis based data, similar gains could also be observed. More interestingly, the RN+NB system using 100-hour training data even outperformed the RN system using 200-hour training data. All those results indicate that using noise bases in mixing noise generation consistently improves the generalization capability of the RN system.

5. CONCLUSION

In this work we have conducted a comprehensive study on noise generalization issues in DNN-based speech enhancement. The use of artificially generated noise bases alone gives comparable performances with systems using 50 types of realistic noise signals in the DNN training. By supplementing the real noises with synthetic noise

[1] http://home.ustc.edu.cn/~shixue/demo/SE_NB_MLSP.html

Table 3. PESQ comparison for unseen noises and targets.

System Description	Narrowband Noises		Wideband Noises	
	0dB	5dB	0dB	5dB
Noisy	2.01	2.27	2.07	2.39
RN	2.79	3.03	2.60	2.94
NB_1	2.39	2.60	2.18	2.49
NB_1+NB_2	2.67	2.91	2.38	2.70
NB_1+NB_2+NB_3	2.66	2.89	2.41	2.74
NB_1+NB_2+NB_3+NB_4	2.69	2.93	2.44	2.76

Table 4. Overall comparison for unseen noises and targets.

Objective Measure	RN		NB	RN+NB	
	100h	200h	100h	100h	200h
PESQ	2.84	2.89	2.71	2.91	2.94
SegSNR	3.54	3.77	3.31	3.96	4.26
SDR	8.20	8.28	8.21	8.30	8.43
SIR	16.09	16.10	15.32	17.59	17.69
SAR	9.80	9.88	9.73	10.14	10.18

bases, better performances could be achieved, indicating the improvement of the noise generalization capability. In future studies, we intend to explore more mechanisms to form noise bases. Furthermore, the speech generalization issue will also be investigated in a similar manner.

6. AKNOLEGEMENT

This work was supported in part by the National Natural Science Foundation of China under Grants 61671422 and U1613211, in part by the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant XDB02070006, and in part by the National Key Research and Development Program of China under Grant 2016YFB1001300.

7. REFERENCES

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

[2] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd Ed, Boca Raton, FL, USA: CRC Press, 2013.

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

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

[5] I. Cohen and S. Gannot, "Spectral enhancement methods," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, Y. Huang (Eds.), Springer 2008, pp.873-901.

[6] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proceedings of the IEEE*, Vol.67, pp.1526-1555, 1992.

[7] D. Burshtein and S. Gannot, "Speech enhancement using a mixture-maximum model," *IEEE Transactions on Speech and Audio Processing*, Vol. 10, No. 6, pp.341-351, 2002.

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

[9] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature* Vol. 401, No. 6755, pp.788-791, 1999.

[10] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Proc. INTERSPEECH*, 2006, pp.2614-2617.

[11] M. N. Schmidt, J. Larsen, and F. T. Hsiao, "Wind noise reduction using non-negative sparse coding," in *IEEE Workshop on Machine Learning for Signal Processing*, 2007, pp.431-436.

[12] J. Le Roux, J. R. Hershey, and F. Weninger, "Deep NMF for speech separation," in *Proc. ICASSP*, 2015, pp.66-70.

[13] S. I. Tamura, "An analysis of a noise reduction neural network," in *Proc. ICASSP*, 1989, pp.2001-2004.

[14] F. Xie and D. V. Compernolle, "A family of MLP based nonlinear spectral estimators for noise reduction," in *Proc. ICASSP*, 1994, pp.53-56.

[15] E. A. Wan and A. T. Nelson, "Networks for speech enhancement," in *Handbook of Neural Networks for Speech Processing*, S. Katagiri, Ed. Norwell, MA, USA: Artech House, 1998.

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

[17] 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.

[18] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 23, No. 1, pp.7-19, 2015.

[19] J. Du, Y.-H. Tu, L.-R. Dai, and C.-H. Lee, "A regression approach to single-channel speech separation via high-resolution deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 24, No. 8, pp.1424-1437, 2016.

[20] X.-G. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising auto-encoder," in *Proc. INTERSPEECH*, 2013, pp.436-440.

[21] 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.

[22] F. Weninger, F. Eyben, and B. Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," in *Proc. ICASSP*, 2014, pp.3737-3741.

[23] P.-S. Huang, M. Kim, M. H. Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 23, No. 12, pp.2136-2147, 2015.

[24] M. Zöhrer, R. Peharz, and F. Pernkopf, "Representation learning for single-channel source separation and bandwidth extension," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 23, No. 12, pp.2398-2409, 2015.

[25] J. Chen, Y. Wang, and D.-L. Wang, "Noise perturbation for supervised speech separation," *Speech Communication*, Vol. 78, pp.1-10, 2016.

[26] M. Sun, X. Zhang, H. Van Hamme, and T.-F. Zheng, "Unseen noise estimation using separable deep auto encoder for speech enhancement," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, Vol. 24, No. 1, pp.93-104, 2015.

[27] R. A. Horn and C. R. Johnson, *Matrix Analysis*, NY, USA: Cambridge University Press, 1985.

[28] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "Dynamic noise aware training for speech enhancement based on deep neural networks," in *Proc. INTERSPEECH*, 2014, pp.2670-2674.

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

[30] G. Hu, 100 nonspeech environmental sounds, 2004 [Online]. Available: http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html

[31] ITU-T, Recommendation 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 Standardisation Sector*, 2001.

[32] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*, Prentice-Hall, 1988.

[33] E. Vincent, R. Gribonval and C. Fvotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 14, No. 4, pp 1462-1469, 2006.