Robust Speech Dereverberation Based on WPE and Deep Learning

Hao Li, Xueliang Zhang, Guanglai Gao
College of Computer Science, Inner Mongolia University, China
E-mail: lihao@mail.imu.edu.cn cszxl@imu.edu.cn csggl@imu.edu.cn

***Abstract*—Reverberation has a considerable impacts on speech quality and intelligibility. Weighted prediction error (WPE) employs a linear regression model to estimate late reverberation and then cancel it. The key point of the WPE is to estimate the power spectrum of the early speech. However, its estimation relies on an iterative procedure with high computational complexity. Another problem is that the WPE has a noise-free assumption. So, the performance degrades in noisy conditions. To address these problems, we propose an algorithm for speech dereverberation in the presence of background noise, in which deep learning is integrated into the WPE method. Specifically, we employ a neural network to predict the power spectral density (PSD) of early speech and a binary mask which distinguishes target speech from background noise. To alleviate the noise impact on estimation of echo path, a dual-filter strategy is adopted to model the echo paths of target speech and background noise individually. Experimental results show that the proposed method significantly improves speech quality in noisy environments.**

Index Terms: long short-term memory, weighted prediction error, dereverberation.

I. Introduction

In an enclosed space, such as a living room, speech signal is often corrupted by its reflections from walls and other surfaces. Reverberation degrades speech quality and intelligibility. It presents a severe problem for a wide range of speech applications, e.g. hearing aids, hands-free telephony and automatic speech recognition [1] [2] [3]. One way to solve the problem is to use dereverberation technique to recover the target speech from the observed reverberant signals.

To solve the reverberation problem, Nakatani *et. al.* [4] proposed a method called WPE. In WPE, reverberations were modeled as an autoregressive (AR) process. Early reverberations of speech were recovered using maximum likelihood (ML) estimation. WPE relied on an iterative procedure to optimize AR weights and PSD of the desired speech alternately. Although WPE performs well, it has several limitations. First, WPE does not consider the noise. Hence the performance of WPE is frequently influenced by background noise. Second, original WPE employs an iterative procedure, which is time consuming. Third, for a short-duration recording, it is hard to estimate the PSD accurately. Fourth, the performance may be degraded when more iterations are applied [4]. In other words, the iterative procedure does not guarantee to converge. Therefore, several studies (e.g. in [5]) focus on replacing the iterative procedure.

Recently, deep neural networks (DNNs) are introduced to signal processing problems and have achieved substantial improvements over the traditional methods [6] [7] [8] [9] [10] [11]. Common usage of DNNs in signal processing is to predict the magnitude spectrogram of the target signal or a time-frequency mask. Compared with a feedforward DNN, long short-term memory (LSTM) accounts for temporal dynamics of speech more naturally. In [12], Zhao*et al.* used an LSTM to remove the late reverberation. Experimental results also showed that the LSTM-based model substantially outperforms DNN-based model on unseen speakers and noises in terms of objective speech intelligibility. Although DNN or LSTM have great potential in building the non-linear relationship between reverberant and desired speech, they are totally data-driven model without considering domain knowledge of speech signal processing which may improve the generalization of supervised algorithms (for example in [13]).

Kinoshita *et al*. [14] proposed a method that combines DNNs and WPE (DNN-WPE) for dereverberation. In the method, a DNN was employed to map reverberant noisy speech to its corresponding anechoic noisy speech (without reverberations). Then the outputs of the DNN were used as the PSD in WPE. Therefore, it leads to a WPE-based method without the iterative procedure. DNN-WPE implies that noise and speech correspond to the same RTF by sharing the inverse filter. However, their RTFs are different unless the sound sources are in the same location. This is also why noise-free assumption should be held in the original WPE.

The echo path of speech and noise are different and noise is unavoidable. In this paper, we propose a dual-filter strategy that combine long short-term memory (LSTM) and WPE together for noisy speech dereverberation. This study assumes a point-source noise field. Actually, no matter what kind of noise field, as long as the echo path of target speech estimation is accurate, the proposed method can work well.

The rest of the paper is organized as follows. We will describe our proposed algorithm in Section 2 and Section 3. The experimental setup and evaluation results are presented in Section 4. We conclude this paper in Section 5.

II. PROBLEM FORMULATION

A. Conventional WPE

Considering a scenario where a single speech source is captured by microphones. In the STFT domain, $s_{n,k}$ denotes the clean speech signal with time frame index $n \in \{1, ..., N\}$, and frequency bin index $k \in \{1, ..., K\}$. The speech observed

at the m-th ($m \in \{1, ..., M\}$) microphone, $x^m_{n,k}$ can be modeled as,

$$x^m_{n,k} = \sum_{l=0}^{L_h-1} \left(h^m_{l,k}\right)^* s_{n-l,k} + e^m_{n,k}, \quad (1)$$

where $h^m_{l,k}$ is an approximation of the ATF between the speech source and the m-th microphone with length of L_h. $(\cdot)^*$ denotes the complex conjugate operator. The additive term $e^m_{n,k}$ represents the background noise. The signal observed at the first microphone ($m = 1$) can be rewritten in the well-known multi-channel linear prediction (MCLP) form,

$$x^1_{n,k} = d^1_{n,k} + \sum_{m=0}^{M} \left(\mathbf{g}^m_k\right)^H \left(\mathbf{x}^m_{n-D,k}\text{-}\mathbf{e}^m_{n-D,k}\right) + e^1_{n,k}, \quad (2)$$

where $d^1_{n,k}$ is the desired signal consisting of direct speech component and early reflections determined by the prediction delay D, and $(\cdot)^H$ denotes the conjugate transposition operator. The vector $\mathbf{g}^m_k \in \mathbf{R}^{L_k}$ is the regression vector of order L_k for the m-th channel. The desired signal can be written as,

$$d^1_{n,k} = \left(x^1_{n,k}\text{-}e^1_{n,k}\right) \text{-} \sum_{m=0}^{M} \left(\mathbf{g}^m_k\right)^H \left(\mathbf{x}^m_{n-D,k}\text{-}\mathbf{e}^m_{n-D,k}\right). \quad (3)$$

Conventional WPE assumed $e^m_{n,k}\,(\forall n, k, m)$ to be 0. And (3) can be written in a compact form using the multi-channel regression vector $\mathbf{g}_k \in \mathbf{R}^{ML_k}$ as,

$$d_{n,k} = x^1_{n,k}\text{-}\mathbf{g}^H_k \mathbf{x}_{n-D,k}. \quad (4)$$

The desired signal in each frequency bin can be modeled as a circular complex Gaussian distribution with zero-mean and frequency-dependent variance. Assuming independence across time frames, by using the maximum likelihood (ML) estimation of the desired speech at each frequency, the joint distribution of the desired speech coefficients at frequency bin, k, is given by,

$$p\left(\mathbf{d}_k\right) = \prod_{n=1}^{N} \frac{1}{\pi\sigma^2_{d_{n,k}}} \exp\left(-\frac{|d_{n,k}|^2}{\sigma^2_{d_{n,k}}}\right), \quad (5)$$

where $\sigma^2_{d_{n,k}}$ is the time-varying PSD of the desired speech. By inserting $d_{n,k}$ from (4) into (5) and taking the negative of logarithm of $p\left(\mathbf{d}_k\right)$, the objective function can be written as,

$$\begin{aligned} \ell(\Theta_k) &= -\log p(d_k|\Theta_k) \\ &= \sum_{n=1}^{N} (\log \sigma^2_{d_{n,k}} + \frac{|x^1_{n,k} - \mathbf{g}^H_k \mathbf{x}_{n-D,k}|^2}{\sigma^2_{d_{n,k}}}), \end{aligned} \quad (6)$$

where $\Theta_k = \{\mathbf{g}_k, \sigma^2_{d_{1,k}}, \sigma^2_{d_{2,k}}, ..., \sigma^2_{d_{N,k}}\}$ is the unknown parameter for the k-th frequency bin and constant terms are ignored. A two-step algorithm to minimize the objective function is adopted by optimizing the AR weights and the PSD, alternatively. Repeating the (7) and (8) until some convergence criterions are satisfied or a maximum number of iteration is exceeded. The flow chat of WPE is shown in Fig. 1.

$$\hat{\mathbf{g}}_k = \left(\sum_{n=1}^{N} \frac{\mathbf{x}_{n-D,k}\mathbf{x}^H_{n-D,k}}{\sigma^2_{d_{n,k}}}\right)^{-1} \sum_{n=1}^{N} \frac{\mathbf{x}_{n-D,k}(x^1_{n,k})^*}{\sigma^2_{d_{n,k}}}. \quad (7)$$

$$\hat{\sigma}^2_{d_{n,k}} = |d_{n,k}|^2, \quad n = 1, 2, ..., N. \quad (8)$$

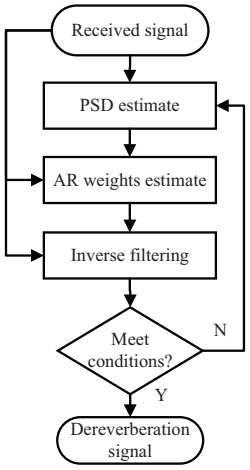

Fig. 1. *Conventional WPE.*

From (8), it is noted that the ideal value of the $\hat{\sigma}^2_{d_{n,k}}$ is the power spectrum of desired speech. The DNN-WPE and proposed method are take advantage of this and introduce supervised learning technology to improve WPE.

III. PROPOSED METHOD

As described in the above section, the inverse filter obtained by WPE and DNN-WPE is optimal for mixture, but not for speech and noise respectively. In this paper, we take a dual-filter strategy: a speech inverse filter and a noise inverse filter, and integrate LSTM and WPE for dereverberation. First, the ideal binary mask (IBM) is estimated by LSTM to distinguish the speech and noise at one frequency bin. In addition, to remove iteration procedure, the desired PSD of speech is also predicted by LSTM. Second, after obtaining IBM and PSD, the two inverse filters strategy is adopted by using WPE. In this paper, we mainly focus on the dereverberation of speech. For the process of the noise, an approximate method is adopted where the power spectrum of mixture is used as the desired PSD of noise.

We will describe the proposed method in three steps. In the first step, the IBM and the desired PSD are estimated by a recurrent neural network. In second step, the output of WPE is obtained with the estimation of the speech AR weights and noise AR weights. Finally, the enhanced signal is converted from the frequency domain to the time domain using the inverse short time Fourier transform (ISTFT).

A. IBM and PSD Estimation

IBM is used to distinguish whether speech dominates or noise dominates at one frequency bin, which is defined by comparing the signal-to-noise ratio (SNR) within each T-F unit against a local criterion (LC) or threshold measure in units of decibels. Only the T-F units with local SNR exceeding LC are

assigned 1 in the binary mask. Let $T(n)$ and $M(n)$ denote target and masker signal power measured in decibels at time n. The IBM is defined as,

$$\mathbf{IBM}(n) = \begin{cases} 1 & if\ T(n) - M(n) > \text{LC} \\ 0 & otherwise. \end{cases} \tag{9}$$

In this paper, T and M indicate reverberate speech and noise speech, respectively. The LC is set to 0. In addition, from (8) we can find that the PSD is the power spectrum of the desired signal. In the LSTM pre-process stage, there have two targets: 1) PSD of desired speech, 2) IBM. The model parameters can be optimized to minimize the Mean Squared Error (MSE) between the estimate and target, defined as,

$$\begin{aligned} J = &\frac{1}{N \times F} \sum_{n=1}^{N} \sum_{f=1}^{F} \left(\left\| \hat{X}_f^{psd}(n) - X_f^{psd}(n)) \right\|_{FN}^2 \right) \\ &+ \frac{1}{N \times F} \sum_{n=1}^{N} \sum_{f=1}^{F} \left(\left\| \hat{M}_f^{nf}(n) - M_f^{nf}(n) \right\|_{FN}^2 \right), \end{aligned} \tag{10}$$

where X^{psd} indicates the PSD of desired signal, M^{nf} indicates the IBM, and FN is the Frobenius Norm.

B. Inverse filter estimate

The speech dereverberation filter and noise dereverberation filter can be calculated by (11) and (12).

$$\begin{aligned} \widehat{\mathbf{g}}_{s_k} = &\left(\sum_{n=1}^{N} \left(\frac{\mathbf{x}_{n-D,k} \mathbf{x}_{n-D,k}^H}{{X^{psd}}^2} \cdot \hat{M}^{nf}(n,k) \right) \right)^{-1} \\ &\cdot \left(\sum_{n=1}^{N} \left(\frac{\mathbf{x}_{n-D,k} (x_{n,k}^1)^*}{{X^{psd}}^2} \cdot \hat{M}^{nf}(n,k) \right) \right) \end{aligned} \tag{11}$$

$$\begin{aligned} \widehat{\mathbf{g}}_{n_k} = &\left(\sum_{n=1}^{N} \left(\frac{\mathbf{x}_{n-D,k} \mathbf{x}_{n-D,k}^H}{(x_{n,k}^1)^2} \cdot \left(1 - \hat{M}^{nf}(n,k) \right) \right) \right)^{-1} \\ &\cdot \left(\sum_{n=1}^{N} \left(\frac{\mathbf{x}_{n-D,k} (x_{n,k}^1)^*}{(x_{n,k}^1)^2} \cdot \left(1 - \hat{M}^{nf}(n,k) \right) \right) \right) \end{aligned} \tag{12}$$

C. Dual-filter dereverberation

After getting inverse filter of speech and noise, using the dual-filter inverse filter strategy, the final dereverberation signal can be obtained by (13).

$$\begin{aligned} d_{n,k} = &x_{n,k}^1 - \hat{M}^{nf}(n,k) \cdot \hat{\mathbf{g}}_{s_k}^H \mathbf{x}_{n-D,k} \\ &- \left(1 - \hat{M}^{nf}(n,k) \right) \cdot \hat{\mathbf{g}}_{n_k}^H \mathbf{x}_{n-D,k} \end{aligned} \tag{13}$$

IV. EXPERIMENTAL SETUPENTS

A. Dataset and Metrics

The proposed system is evaluated by using the IEEE corpus [15] spoken by a female speaker. There are 72 phonetically balanced lists of sentences in the corpus, with 10 sentences in each list. In our experiments, we select sentences from List 1-50, List 67-72 and List 51-60 to construct training data, validation data and test data, respectively. The room impulse response (RIRs) and noises come form 2014 REVERB Challenge2 [16]. The training and development sets are convolved with 24 RIRs and corrupted by several types of noises. Each RIR consists of 8 channels. The reverberation time (T60) of 24 RIRs ranges roughly from 0.2 to 0.8 second. The SNRs are at 0, 3, 6, 10, 15, 21, 24 dB. The test set contains a set of reverberant noisy speech signals, generated by convolving clean speech signals with recorded RIRs and subsequently adding recorded noise signals. There are 6 different reverberation conditions: 3 rooms with different volumes (small, medium and large size) and 2 types of distances between the speaker and the microphones (near=50cm and far=200cm). RIRs are recorded by an 8-ch circular array with diameter of 20 cm. The recorded noise is added with SNR at 0, 5, 10, 15, 20 dB. T60s of small, medium, and large-size rooms are about 0.25s, 0.68s, and 0.73s, respectively. We randomly generate 100k utterances as training set to train the LSTM weights. 3000 sentences are generated as test set to evaluate the proposed method.

We quantitatively evaluate the performance of the proposed method by perceptual evaluation of speech quality (PESQ) [17] and cepstral distance (CD) [18], both of which are widely used to evaluate speech dereverberation task. The higher the PESQ score, the better the speech quality is. For the CD, the lower number indicates the better performance.

B. Comparison Method

Since the algorithm is independent on the number of microphones, we evaluate and compare the performances in single- and multi-channel scenarios, separately. We evaluate and compare the performances in different SNR and room conditions, separately. WPE [4] and DNN-WPE [14] are involved for comparison.

C. Algorithm settings

The regression order and the prediction delay in the WPE are set to 10 and 3 for multi-channel conditions. For single-channel conditions, 37 and 3 are used. The DNN-WPE and the proposed method employed the same configuration for fair comparison. The length of the STFT analysis window is 20 ms, and the window shift is 10 ms. Thus, the number of FFT points is 320 for 16 kHz sampling rate. The number of iterations in original WPE is set to 3, which got the best results according to our experiments.

We use two unidirectional LSTM layers followed by one fully-connected layers. The activation functions of IBM output layer and PSD output layer are sigmoid and ReLU, respectively. The number of memory cells in each LSTM is 512, and the number of nodes in the fully-connected layer is 1024. The cost function is mean square error (MSE). Weights of the networks are randomly initialized. The ADAM optimizer [19] is utilized for back propagation. The models are trained using Pytorch.

In addition, for fairness, we replace DNN in DNN-WPE by LSTM which has the same configuration as in the proposed method, except for the output layer.

D. EVALUATION AND COMPARISON

In order to compare the performance of the dereverberation methods, we analyze the average PESQ and CD on test set in terms of different RIRs, SNRs and input channels in following subsection, respectively.

1) Different RIR conditions: Table I and II show the average PESQ and CD in terms of different rooms, respectively. It can be seen that the proposed method obtains the highest PESQ scores and lowest CD in most of the room conditions. In room 1 with little reverberation, WPE is better than DNN-WPE, and comparable with the proposed method. When the reverberation becomes severe, DNN-WPE performs better than WPE especially in terms of CD. In fact, the performance of all three methods depend on the estimation of the early reverberant spectrum. WPE employs mixture spectrum as the initial early reverberation spectrum. In a little reverberation room, mixture spectrum is close to the true early reverberation spectrum. Therefore, WPE is more likely to converge to the target from a good start point by its iterative process. In contrast, it is more difficult for WPE to get a good estimation in a room with large reverberation where the advantage of WPE over DNN-WPE is not obvious even worse using PESQ as evaluation metric. It can also be seen that WPE outperforms DNN-WPE in small room (room 1) with less reverberation but vice versa in large room (room 3).

When using CD as the evaluation metric, we can see in Table II that the proposed algorithm is still the best in room 2 and room 3. In room 1, unprocessed speech has the best result for near case, and slightly worse than WPE for far case. It means that all three methods introduce speech distortion during dereverberation. We can see that the DNN-WPE brings more distortion than the WPE, and the proposed method provides a rational balance between speech distortion and dereverberation.

TABLE I
AVERAGE PESQ SCORE IN TERMS OF DIFFERENT T60.

Method \ Room	Room1 (T60=250ms)		Room2 (T60=680ms)		Room3 (T60=730ms)	
	Near	Far	Near	Far	Near	Far
Unproposed	3.01	2.55	1.91	1.70	1.97	1.76
WPE	**3.02**	2.63	1.97	1.79	2.07	1.89
DNN-WPE	2.93	2.61	1.99	**1.80**	**2.11**	**1.95**
Proposed	**3.02**	**2.65**	**2.01**	**1.80**	2.10	1.89

TABLE II
AVERAGE CD IN TERMS OF DIFFERENT T60.

Method \ Room	Room1 (T60=250ms)		Room2 (T60=680ms)		Room3 (T60=730ms)	
	Near	Far	Near	Far	Near	Far
Unproposed	**2.09**	2.55	4.56	4.83	4.20	4.47
WPE	2.15	**2.52**	4.47	4.73	4.04	4.25
DNN-WPE	2.63	2.91	4.53	4.80	4.14	4.35
Proposed	2.23	2.57	**4.44**	**4.61**	**4.01**	**4.17**

2) Different SNR conditions: In order to analyze the effect of noise on dereverberation, we calculate the average PESQ and CD in term of the SNR. Table III and IV show the PESQ and CD results of methods on different SNRs, respectively. From Table III, it can be seen that the proposed method can get the highest PESQ scores in all SNR conditions except for the 20 dB. The higher SNR, the larger PESQ gain is obtained by all three methods. In table IV, we can see that the proposed method obtains the lowest CD in all SNR conditions. In low SNR conditions, with the effect of additive noise, the PESQ and CD of DNN-WPE and WPE are similar to the mixture sometimes. The reason is that in low SNR conditions, the echo path of noise dominates the optimal solution.

TABLE III
AVERAGE PESQ SCORE IN TERMS OF DIFFERENT SNR.

Method \ SNR(dB)	0	5	10	15	20
Unproposed	1.76	1.98	2.18	2.35	2.48
WPE	1.75	2.00	2.25	2.47	**2.67**
DNN-WPE	1.78	2.02	2.26	2.47	2.63
Proposed	**1.79**	**2.03**	**2.48**	**2.64**	2.60

TABLE IV
AVERAGE CD SCORE IN TERMS OF DIFFERENT SNR.

Method \ SNR(dB)	0	5	10	15	20
Unproposed	4.64	4.23	3.78	3.33	2.95
WPE	4.65	4.20	3.70	3.18	**2.73**
DNN-WPE	4.70	4.30	3.88	3.46	3.11
Proposed	**4.63**	**4.17**	**3.66**	**3.16**	**2.73**

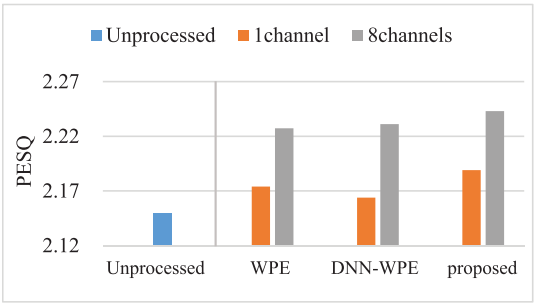

Fig. 2. Average PESQ score of different methods with 1- and 8-channel inputs.

3) Different channel conditions: The proposed method is independent of the number of microphones. We also compare the performance of our method with WPE and DNN-WPE in single-channel conditions. Fig. 2 and 3 show the average PESQ and CD in 1 and 8-channel conditions. It can be seen that the proposed method also outperforms the other two methods in 1-channel conditions, and the more microphones, the better the performance. The proposed method estimates the PSD and IBM directly and inherits the advantage of the WPE which can be applied to any topology of microphone array without changing the algorithm.

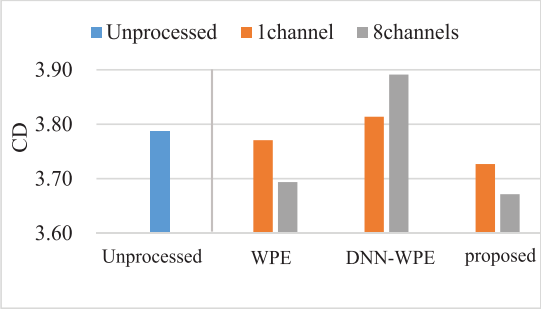

Fig. 3. Average CD of different methods with 1- and 8-channel inputs.

V. CONCLUSION

Deep learning has been introduced into speech signal processing in recent years and exhibits great potential for the classical problems in this area. One way to use deep learning is to map the input signal to the output target directly. The other way is to combine the deep learning with the traditional methods, in which deep learning is used to estimate the parameters needed in traditional methods. In this paper, we propose a dereverberation algorithm which integrates deep learning and the WPE. The deep neural network is used to predict the power spectrum of the early speech, which is the key parameter for WPE algorithm. At the same time, the deep neural network is used to meet the noise-free assumption of the WPE to some extent. According to the experiments results, the proposed method achieves the best performance in most of the conditions.

REFERENCES

[1] G. Kim, Y. Lu, Y. Hu, and P.C. Loizou. An algorithm that improves speech intelligibility in noise for normal-hearing listeners. *The Journal of the Acoustical Society of America*, 126(3):1486–1494, 2009.

[2] C. Demir, M. Saraclar, and A.T. Cemgil. Single-channel speech-music separation for robust ASR with mixture models. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(4):725–736, 2013.

[3] F. Weninger, J. Geiger, M. Wollmer, B. Schuller, and G. Rigoll. Feature enhancement by deep lstm networks for asr in reverberant multisource environments. *Computer Speech & Language*, 28(4):888–902, 2014.

[4] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang. Speech dereverberation based on variance-normalized delayed linear prediction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 18(7):1717–1731, 2010.

[5] M. Parchami, W.-P. Zhu, and B. Champagne. Speech dereverberation using linear prediction with estimation of early speech spectral variance. In *ICASSP*, pages 504–508. IEEE, 2016.

[6] X. Zhang, H. Zhang, S. Nie, G. Gao, and W. Liu. A pairwise algorithm using the deep stacking network for speech separation and pitch estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(6):1066–1078, 2016.

[7] X. Zhang and D. Wang. Deep learning based binaural speech separation in reverberant environments. *IEEE/ACM transactions on audio, speech, and language processing*, 25(5):1075–1084, 2017.

[8] K. Han, Y. Wang, and D. Wang. Learning spectral mapping for speech dereverberation. In *ICASSP*, pages 4628–4632. IEEE, 2014.

[9] D.S. Williamson and D. Wang. Time-frequency masking in the complex domain for speech dereverberation and denoising. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(7):1492–1501, 2017.

[10] K. Tan, J. Chen, and D. Wang. Gated residual networks with dilated convolutions for supervised speech separation. In *ICASSP*, pages 21–25, 2018.

[11] K. Han, Y. Wang, D. Wang, W.S. Woods, I. Merks, and T. Zhang. Learning spectral mapping for speech dereverberation and denoising. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 23(6):982–992, 2015.

[12] Y. Zhao and D. Wang. Late reverberation suppression using recurrent neural networks with long short-term memory. In *ICASSP*, pages 5434–5438, 2018.

[13] H. Li, S. Nie, X. Zhang, and H. Zhang. Jointly optimizing activation coefficients of convolutive NMF using DNN for speech separation. In *INTERSPEECH*, pages 550–554, 2016.

[14] K. Kinoshita, M. Delcroix, H. Kwon, T. Mori, and T. Nakatani. Neural network-based spectrum estimation for online WPE dereverbertion. In *INTERSPEECH*, pages 384–388, 2017.

[15] E. Rothauser. IEEE recommended practice for speech quality measurements. *IEEE Trans. on Audio and Electroacoustics*, 17:225–246, 1969.

[16] K. Kinoshita, M. Delcroix, S. Gannot, E.A. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, and B. Raj. A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research. *EURASIP Journal on Advances in Signal Processing*, 2016(1):7, 2016.

[17] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *ICASSP*, volume 2, pages 749–752. IEEE, 2001.

[18] C.H. Taal, R.C. Hendriks, R. Heusdens, and J. Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136, 2011.

[19] D.P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.