Monaural Source Separation Based on Sequentially Trained LSTMs in Real Room Environments

Yi Li, Yang Sun, Syed Mohsen Naqvi
Intelligent Sensing and Communications Group
Newcastle University
Newcastle upon Tyne, UK
{y.li140, y.sun29, mohsen.naqvi}@newcastle.ac.uk

***Abstract*—In recent studies on Monaural Source Separation (MSS), the long short-term memory (LSTM) network has been introduced to solve this problem, however, its performance is still limited particularly in real room environments. According to the training objectives, the LSTM-based MSS is categorized into three aspects, namely mapping, masking and signal approximation (SA) based methods. In this paper, we introduce dereverberation mask (DM) and establish a system to train two SA-LSTMs sequentially, which dereverberate speech mixture and improve the separation performance. The DM is exploited as the training target of the first LSTM. Then, the enhanced ratio mask (ERM) is proposed and set as the training target of the second LSTM. We evaluate the proposed method with the IEEE and the TIMIT datasets with real room impulse responses and noise interferences from the NOISEX dataset. The detailed evaluations confirm that the proposed method outperforms the state-of-the-art.**

***Index Terms*—Monaural source separation, long short-term memory, signal approximation, dereverberation mask, enhanced ratio mask.**

I. Introduction

In speech source separation, the desired speech signals are required to be recovered from the mixture. Although the problem is challenging due to the unknown mixing process, it is essential for real-world applications such as automatic speech recognition (ASR), hearing aids and robotics [1]. According to the number of microphones, the separation problem is categorized into three cases, namely monaural, binaural and multichannel source separation. However, MSS is more challenging because the solution is not unique with only one single channel of information available [2]. Furthermore, in real room environments, the room reflections can be challenging because smearing is caused across time and frequency, which affects and reduces the separation performance [3].

Recently, deep neural networks (DNN) have been introduced to solve MSS problem and their separation performance has significant improvement. The DNN-based MSS is categorized into three aspects, namely mapping, masking and signal approximation (SA) based methods, according to the training objectives [2]. Mapping-based targets correspond to the spectral representations of the desired signal, while masking-based targets concentrate the time-frequency relationships of the desired signal to background interference [3] [4]. However, SA-based targets combine the advantages of the other two sorts of targets and improve separation performance [5].

Several approaches have been introduced to address the MSS problem with the above mentioned three types of training targets. For example, Jin and Wang applied DNN to estimate the ideal binary mask (IBM) for the speech separation [6]. But, IBM is a binary mask, and the associated hard decision causes a loss in the separation performance [7]. Then, a DNN that estimates the ideal ratio mask (IRM) has been confirmed to improve the objective speech quality in addition to predicted speech intelligibility [8]. Besides, each Time-Frequency (T-F) unit is assigned as the ratio of desired speech signal energy to mixture energy [9]. The IRM-based method outperforms the IBM-based method but cannot efficiently reduce the reflections in real room environments. In [4], Sun et al. exploited an ideal enhanced mask (IEM) with two trained DNNs to dereverberate and separate speech signals.

Recurrent neural networks (RNN) treat input samples as a sequence and model the changes over time [3]. The RNN plays an important role in learning the temporal dynamic of speech but are limited to the vanishing or exploding gradient problem [10]. Hence, long short-term memory block was introduced to further improve the performance [11]. For example, Chen et al. introduced long short-term memory block in RNN and the generalization ability of the neural network model was refined.

In this paper, we introduce a new sequentially trained LSTMs method to further improve the separation performance in real room environments. The organization of this paper is as follows: the proposed method is introduced in Section II. In Section III, the experimental settings and results are provided. The conclusions and future work are given in Section IV.

II. Proposed Method

In the MSS problem, the convolutive mixture is generated by the clean speech signal, background interference, and the real room impulse response:

$$y\left(m\right)=s\left(m\right)*h_s\left(m\right)+i\left(m\right)*h_i\left(m\right) \tag{1}$$

where the $s(m)$, $i(m)$ and $y(m)$ denote the clean speech signal, the background interference and the mixture at discrete time m, respectively. The h_s and h_i are the impulse responses of speech signal and interference, respectively. And '$*$' is the convolution operator. Different from DNNs, the LSTM block in RNN uses memory cells and three gates with long-term speech contexts. The forget gate decides how much previous

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

information is erased from the cell and the input gate decides how much information is added to the cell [12].

$$i_t = \sigma\left(W_{ix}x_t + W_{ih}h_{t-1} + b_i\right) \tag{2}$$

$$f_t = \sigma\left(W_{fx}x_t + W_{fh}h_{t-1} + b_f\right) \tag{3}$$

The equations (2) and (3) describe the input and forget gates which are represented by i_t and f_t, respectively. x_t and h_t are input and hidden activations at time t. W' s, and b' s are weights and biases. Because the gates are bounded to [0, 1] by the function $\sigma(s)$, the output of the LSTM block is bounded to [-1, 1] [12].

By using the Short Time Fourier Transform (STFT), the speech mixture can be represented as:

$$Y\left(t,f\right) = S\left(t,f\right)H_s\left(t,f\right) + I\left(t,f\right)H_i\left(t,f\right) \tag{4}$$

where $S(t,f)$, $I(t,f)$ and $Y(t,f)$ are the spectrum of speech, interference and speech mixture, respectively. $H_s(t,f)$ and $H_i(t,f)$ are the room impulse response of speech signal and interference in time-frequency domain, respectively. The spectrum of the clean speech is estimated with the ideal T-F mask $M(t,f)$ as:

$$S\left(t,f\right) = Y\left(t,f\right)M\left(t,f\right) \tag{5}$$

One of the training targets, IRM is defined as a soft decision as [3]:

$$IRM\left(t,f\right) = \left(\frac{\left|S\left(t,f\right)\right|^2}{\left|S\left(t,f\right)\right|^2 + \left|I\left(t,f\right)\right|^2}\right)^{\beta} \tag{6}$$

where β represents a tunable scaling parameter and is selected 0.5 in most of situations for the best separation performance. When the environment is reverberant, the early reflections are considered. Therefore, the IRM with reverberant environment is expressed as [3]:

$$IRM_r\left(t,f\right) = \left(\frac{\left|D\left(t,f\right)\right|^2}{\left|S\left(t,f\right)\right|^2 + \left|I\left(t,f\right)\right|^2}\right)^{\beta} \tag{7}$$

where $D\left(t,f\right)$ is the direct sound. In practice, reverberant speech consists of three components: the direct sound, early and late reflections. However, by using this method, the direct sound is obtained, which is still different from the clean speech signal, because the value of direct path is not always equal to 1.

To address this problem, the DM was proposed to eliminate most of the reflections [4]:

$$DM\left(t,f\right) = \left|S\left(t,f\right) + I\left(t,f\right)\right| \left|Y\left(t,f\right)\right|^{-1} \tag{8}$$

Even though, in practice, to obtain the dereverberated mixture is very challenging. Besides, the original SA-LSTM method performance is limited to highly reverberated environments. Therefore, in this paper, we propose a sequentially trained two-LSTMs with DM and new enhanced ratio mask (ERM) for speech separation in highly reverberant room environments.

The block diagram of the proposed method is shown in Fig. 1. In the training stage, DM, the training target of the first LSTM is obtained from the targets calculation module after the mixture is generated by the clean speech signals and the background interferences. After the first LSTM is trained, the estimated dereverberated mixture is obtained from the estimated mask $\widehat{DM}\left(t,f\right)$:

$$\hat{Y}_{dere}\left(t,f\right) = Y\left(t,f\right)\widehat{DM}\left(t,f\right) \tag{9}$$

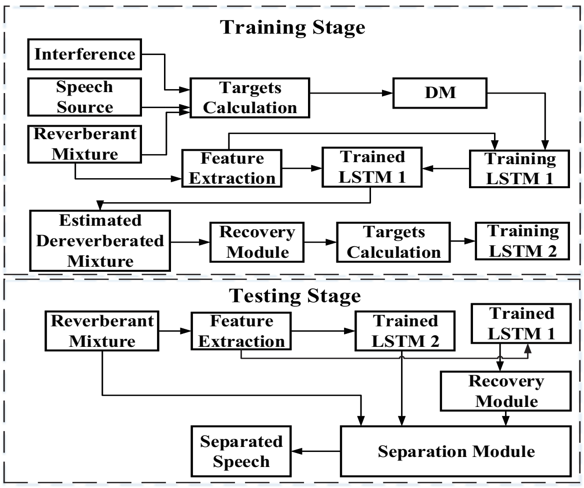

Fig. 1. The block diagram of the proposed method. In the training stage, two LSTMs are trained sequentially with the DM and the ERM as training targets, respectively. In the testing stage, the features of the mixture are used to estimate DM and ERM. Finally, the desired speech signal is obtained in the separation module.

Hence, we can generate a new proposed ERM to separate the desired speech signal, which can be expressed as:

$$\widehat{ERM}\left(t,f\right) = \frac{|S(t,f)|}{|\hat{Y}_{dere}\left(t,f\right)|} \tag{10}$$

In the testing stage, two sequentially trained LSTMs are used. The final separated speech signal can be obtained from the separation module as:

$$\hat{S}(t,f) = \widehat{ERM}(t,f)\widehat{DM}(t,f)Y(t,f) \tag{11}$$

The feature combination, similar to [4] and [13], is used in our proposed method, which contains the amplitude modulation spectrogram (AMS), relative spectral transform and perceptual linear prediction (RASTA-PLP) [14], mel-frequency cepstral coefficients (MFCC), cochleagram response and their deltas are extracted by the 64-channel gammatone filterbank to generate the compound feature [15].

Compared with single SA-LSTM, by using sequentially trained LSTMs, the estimated dereverberated mixture is obtained by using the estimated DM from trained LSTM1. Then, the new ratio mask, ERM, is calculated by using the desired speech signal and the estimated dereverberated mixture. The proposed ERM can better model the relationship between the clean speech signal and the estimated dereverberated mixture. Therefore, the separation performance is further improved and can be confirmed by the detailed evaluations in the following section.

III. Experimental Results and Discussion

A. Experimental Settings

The clean speech source signals are randomly selected from the TIMIT [16] and the IEEE [17] corpora which are 720 clean utterances from a male speaker in the IEEE corpus and 6300 utterances from 630 speakers in the TIMIT database. By using both datasets, it confirms that the proposed method is speaker-independent. The noise interferences are selected from NOISEX database [18]. The $factory$, $babble$, and $cafe$ noise interferences from NOISEX database are used in our evaluations. Each noise interference signal has four minutes long and it is divided into two clips with the same length. One is used to generate training data and another is used to generate testing data. Hence, in total, there are 5400 mixtures (600×3×3) in training data, 1080 mixtures (120×3×3) in testing data. All of the DNNs of the comparison group and the proposed method have three hidden layers and each hidden layer has 512 units.

The speech mixtures are generated by the convolution of speech signals and interferences with the room impulse responses (RIRs) [4] which are recorded in four different types of room environments i.e. different RT60s. The parameters are illustrated in Table 1:

TABLE I
ROOM SETTINGS FOR REAL RIRS [4]

Room	Size	Dimension (m^3)	$RT60$ (s)
A	Medium	5.7 × 6.6 × 2.3	0.32
B	Small	4.7 × 4.7 × 2.7	0.47
C	Large	23.5 × 18.8 × 4.6	0.68
D	Medium	8.0 × 8.7 × 4.3	0.89

Besides, to evaluate the generalization ability of the proposed method, in training and testing stages, we use different RIRs, the RIRs in training data are unseen in the testing data. For comparison, these clean speech signals are mixed with various interferences at three different SNR levels (-3 dB, 0 dB and 3 dB).

B. Experimental Results and Analysis

As the outputs of the different stages of different state-of-the-art methods, the spectrograms are plotted in Fig. 2. From the spectrograms, the separated signal with the proposed method is more similar to the clean speech signal because the ERM is introduced and the reflections are further eliminated by the proposed method. Therefore, the separation performance is further improved.

There are two evaluation measures, the short-time objective intelligibility (STOI) and improved source to distortion ratio (ΔSDR). The values of the STOI are bounded in the range of [0, 1], which indicates the human speech intelligibility scores.

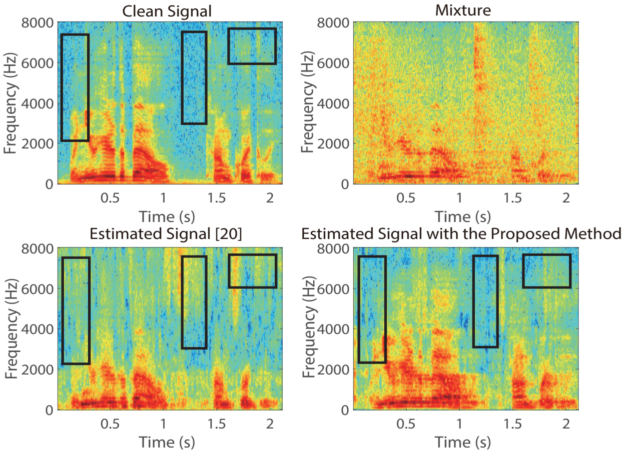

Fig. 2. The spectrograms of clean speech signal, reverberant mixture, estimated speech signals with the SA-LSTM method and the proposed method, respectively. The result is selected from 120 experiments due to the best separation performance and the inference used to generate mixture is the $factory$ noise with -3 dB SNR level.

The SDR is exploited to evaluate the overall separation performance [21]. The ΔSDR is calculated by using unprocessed speech mixture and the estimated speech signal. The higher values of these measurements mean that the desired speech signal is better reconstructed. The experimental results are presented in Table II and Figs. 3-6.

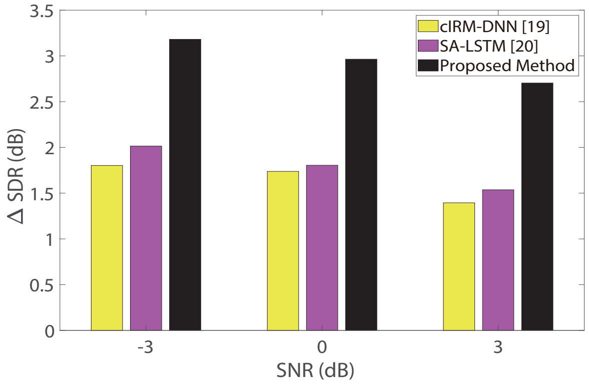

Fig. 3. The improved SDR (dB) in terms of different SNR levels and proposed and state-of-the-art methods. The X-axis is the SNR level and the Y-axis is the improved SDR. Each result is the average value of 120 experiments and the inference used to generate mixture is the $factory$ noise.

The first comparison is among the complex ideal ratio mask-DNN (cIRM-DNN) [19], the original SA-LSTM [20] and the proposed methods in terms of STOI performance. As the increase of the mixing SNR levels, the separation performance is improved. Because, in the mixture, the energy level of the desired speech signal is larger. Besides, compared with the performance with different types of noise interferences, it can be observed from Table II that when the noise type is $babble$, the separation performance is relatively low. Due to the

TABLE II
SEPARATION PERFORMANCE COMPARISON IN TERMS OF STOI WITH DIFFERENT TRAINING TARGETS, SNR LEVELS AND NOISE. EACH RESULT IS THE AVERAGE VALUE OF 120 EXPERIMENTS. **BOLD** INDICATES THE BEST RESULT.

Noise	factory			babble			cafe		
SNR level	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB
Unprocessed	0.53	0.53	0.53	0.52	0.52	0.52	0.58	0.58	0.58
cIRM-DNN [19]	0.56	0.57	0.58	0.54	0.56	0.58	0.52	0.55	0.59
SA-LSTM [20]	0.57	0.58	0.59	0.56	0.57	0.59	0.60	0.60	0.61
Proposed Method	**0.64**	**0.65**	**0.67**	**0.61**	**0.63**	**0.65**	**0.69**	**0.69**	**0.70**

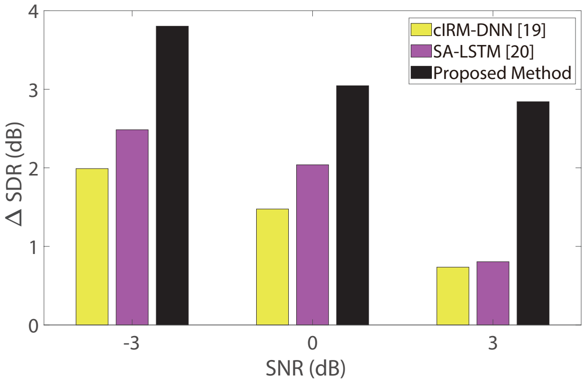

Fig. 4. The improved SDR (dB) in terms of different SNR levels and proposed and state-of-the-art methods. The X-axis is the SNR level and the Y-axis is the improved SDR. Each result is the average value of 120 experiments and the inference used to generate mixture is the *babble* noise.

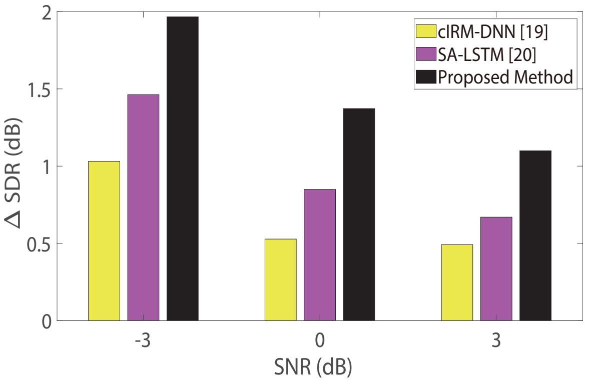

Fig. 5. The improved SDR (dB) in terms of different SNR levels and proposed and state-of-the-art methods. The X-axis is the SNR level and the Y-axis is the improved SDR. Each result is the average value of 120 experiments and the inference used to generate mixture is the $cafe$ noise.

unseen speakers in the *babble* noise interference, it is difficult for networks to distinguish and separate the desired speech signal. From Table II, it is clear that the proposed method outperforms the cIRM-DNN and SA-LSTM methods in all SNR levels and scenarios. For instance, in 3 dB SNR level, for the $factory$ noise, the proposed method can achieve 0.67 over STOI although the original SA-LSTM method only achieves 0.59 and the cIRM-DNN method only achieves 0.58.

Improved SDR in terms of different SNR levels and methods for the $factory$ noise can be observed in Fig. 3. Three SNR levels are used (-3, 0 and 3 dB) to evaluate the separation performance of proposed methods. Unlike STOI, as the increase of the mixing SNR levels, the separation performance falls. The difference in terms of ΔSDR with the unseen RIRs between the original SA-LSTM method and the proposed method is significant at three SNR levels. The proposed sequentially trained LSTMs method outperforms the state-of-the-art method. For example, in -3 dB SNR levels, the proposed method can achieve 57.9 % more improvements compared with the original SA-LSTM method over ΔSDR.

Similarly, the other two noise interferences are introduced and the performance of the proposed method is still outstanding. The results with *babble* noise in terms of ΔSDR are shown in Fig. 4. The difference is clear except at 3 dB SNR. As for the $cafe$ noise result in Fig. 5, ΔSDR with the unseen RIRs between the original SA-LSTM method and the proposed method reaches 1.46 dB and 1.97 dB at -3 dB SNR. Furthermore, for the same SNR level, the values in the improved SDR with the unseen RIRs under the original SA-LSTM method and the proposed method is not consistent with a variety of noise interferences.

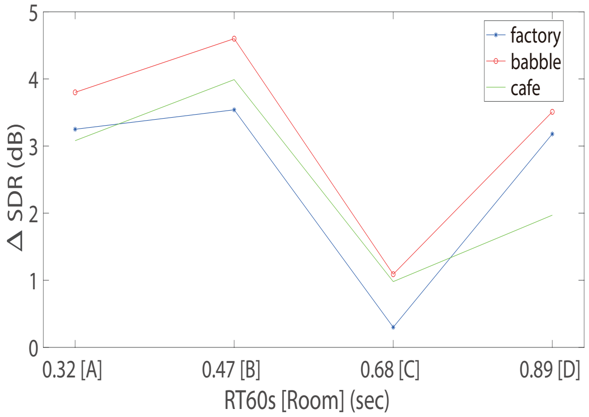

Fig. 6. The improved SDR (dB) in terms of different real rooms with various RT60s and proposed method under -3 dB SNR. Room settings for real RIRs are listed in Table I. The X-axis is the rooms and the Y-axis is the improved SDR. Each result is the average value of 120 experiments and the inference used to generate mixture is the $cafe$ noise.

It can be observed from Fig. 6 that as the increase of RT60s, more reflections will exist in the mixture and the ERM can lead to a higher SDR improvement. For instance, compared with the performance with $cafe$ noise, the proposed method can obtain 56.3 $\%$ more improvement in Room A than in Room D. If the RIRs are seen, the better separation performance will be obtained. For example, the proposed method achieves 21.1 $\%$ more improvements in Room B compared with Room A over ΔSDR with the $babble$ noise. Besides, due to the influence of the direct-to-reverberant energy ratio (DRR), although the RT60s in Room C is higher than Room D, the separation performance is much lower over ΔSDR [22].

In summary, the experimental results confirm that the proposed method can further improve the separation performance compared with the cIRM-DNN and the original SA-LSTM methods in both SDR and STOI. By introducing the DM and the ERM, the majority of the reflections are removed and the desired speech signal can be better estimated from the dereverberated speech mixture. Furthermore, the required information in the dereverberated mixture is considered by our proposed method where the original SA-LSTM method ignores.

IV. Conclusions and Future Work

In this paper, we explored sequentially trained LSTMs architectures for MSS problems. Overall, the separation performance in terms of ΔSDR and STOI was enhanced by optimizing soft T-F masks, DM and ERM, and obtaining the dereverberated mixture. In the proposed method the sequentially trained LSTMs better utilized the estimation of the first trained LSTM and improved the separation performance. By using the proposed ERM to separate the desired speech signal from the dereverberated speech mixture, the separation performance was further improved.

To further improve the performance, the first direction is utilizing the phase information of the desired signal to operate the separation in the complex domain. The phase spectrum plays an important part in increasing perceptual quality [19]. The second direction is stacking LSTM cells to model longer temporal information [23].

References

[1] S. M. Naqvi, M. Yu, and J. A. Chambers, “A multimodal approach to blind source separation of moving sources,” IEEE Journal of Selected Topics in Signal Processing, vol. 4, pp. 895–910, 2010.

[2] P. S. Huang, M. Kim, M. H. Johnson and P. Smaragdis, “Joint optimization of masks and deep recurrent neural networks of monaural source separation,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, pp. 1–12, 2015.

[3] D. L. Wang, and J. T. Chen, “Supervised speech separation sased on deep learning: An Overview,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, pp. 1702–1726, 2018.

[4] Y. Sun, W. Wang, J. A. Chambers, and S. M. Naqvi, “Two-Stage monaural source separation in reverberant room environments using deep neural networks,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 27, no. 1, pp. 125–138, 2019.

[5] M. Yu, A. Rhuma, S. M. Naqvi, L. Wang, and J. A. Chambers, “A posture recognition-based fall detection system for monitoring an elderly person in a smart home environment,” IEEE Transactions on Information Technology in Biomedicine, vol. 16, no. 6, pp. 1274–1286, 2012.

[6] Z. Jin and D. L. Wang, “A supervised learning approach to monaural segregation in reverberant conditions,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 17, no. 4, pp. 625–638, 2009.

[7] M. Yu, Y. Yu, A. Rhuma, S. M. Naqvi, L. Wang, and J. A. Chambers, “An online one class support vector machine-based person-specific fall detection system for monitoring an elderly individual in a room environment,” IEEE Journal of Biomedical and health informatics, vol. 17, no. 6, pp. 1002–1014, 2013.

[8] Y. Sun, Y. Xian, P. Feng, J. Chambers, and S. M. Naqvi, “Estimation of the number of sources in measured speech mixtures with collapsed Gibbs sampling,” Sensor Signal Processing for Defence Conference, vol. 22, no. 12, pp. 849–858, 2017.

[9] Y. Wang, A. Narayanan, and D. L. Wang, “On training targets for supervised speech separation,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 22, no. 12, pp. 1849–1858, 2014.

[10] Y. Sun, W. Wang, J. A. Chambers, and S. M. Naqvi, “Enhanced time-frequency masking by using neural networks for monaural source separation in reverberant room environments,” European Signal Processing Conference, vol. 26, pp. 1647–1651, 2018.

[11] J. Chen and D. L. Wang, “Long short-term memory for speaker generalization in supervised speech separation,” The Journal of the Acoustical Sociry of America, vol. 141, no. 6, pp. 4705–4714, 2017.

[12] E. M. Grais, and H. Erdogan, “Single channel speech music separation using nonnegative matrix factorization and spectral masks,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 22, no. 12, pp. 1849–1858, 2014.

[13] D. S. Williamson and D. L. Wang, “Time-frequency masking in the complex domain for speech dereverberation and denoising,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 25, no. 7, pp. 1492–1501, 2017.

[14] H. Hermansky, and N. Morgan, “RASTA processing of speech,” Processon IEEE International Conference Digital Signal Process., pp. 1–6, 2011.

[15] M. Delfarah, and D. L. Wang, “Features for masking-based monaural speech separation in reverberant conditions,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 25, no. 5, pp. 1085–1094, 2017.

[16] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, “TIMIT acoustic phonetic continuous speech corpus [CD-ROM],” Linguistic Data Consortium, 1993.

[17] IEEE Audio and Electroacoustics Group, “IEEE recommended practice for speech quality measurements,” IEEE Transactions on Audio, Speech, and Language Processing, vol. AE-17, no. 3, pp. 225–246, 1969.

[18] 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,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 12, no. 3, pp. 247–251, 1993.

[19] D. S. Williamson, Y. Wang, and D. L. Wang, “Complex ratio masking for monaural speech separation,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 24, no. 3, pp. 483–492, 2016.

[20] F. Weninger, J. R. Hershey, J. Le. Roux, and B Schuller, “Discriminatively trained recurrent neural networks for single-channel speech separation,” Proc. of IEEE Global Conference on Signal and Information Processing (GlobalSIP), 2014.

[21] Y. Sun, W. Rafique, J. A. Chambers, and S. M. Naqvi, “Underdetermined source separation using time-frequency masks and an adaptive combined Gaussian-Student’s T probabllistic model,” IEEE International Conference on Acoustics, Speech and Signal Processing, 2017.

[22] Y. Sun, W. Wang, J. A. Chambers, and S. M. Naqvi, “Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 7, pp. 1793–1805, 2010.

[23] K. Han, Y. Wang, D. L. 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, vol. 23, no. 6, pp. 982–992, 2015.