Two-Stage Monaural Source Separation in Reverberant Room Environments using Deep Neural Networks

Yang Sun, *Student Member, IEEE,* Wenwu Wang, *Senior Member, IEEE,*
Jonathon Chambers, *Fellow, IEEE,* and Syed Mohsen Naqvi, *Senior Member, IEEE*

***Abstract*—Deep neural networks (DNNs) have been used for dereverberation and separation in the monaural source separation problem. However, the performance of current state-of-the-art methods is limited, particularly when applied in highly reverberant room environments. In this paper, we propose a two-stage approach with two DNN-based methods to address this problem. In the first stage, the dereverberation of the speech mixture is achieved with the proposed dereverberation mask (DM). In the second stage, the dereverberant speech mixture is separated with the ideal ratio mask (IRM). To realize this two-stage approach, in the first DNN-based method, the DM is integrated with the IRM to generate the enhanced time-frequency (T-F) mask, namely the ideal enhanced mask (IEM), as the training target for the single DNN. In the second DNN-based method, the DM and the IRM are predicted with two individual DNNs. The IEEE and the TIMIT corpora with real room impulse responses (RIRs) and noise from the NOISEX dataset are used to generate speech mixtures for evaluations. The proposed methods outperform the state-of-the-art specifically in highly reverberant room environments.**

***Index Terms*—Deep neural networks, monaural source separation, dereverberation mask, highly reverberant room environments**

I. Introduction

SOURCE separation aims to separate the desired speech signals from the mixture, which consists of the speech sources, the background interference and their reflections. Nowadays, due to applications such as automatic speech recognition (ASR), assisted living systems and hearing aids [1]–[6], source separation in real-world scenarios has attracted considerable research attention. The source separation problem is categorized into multichannel, stereo-channel (binaural) and single-channel (monaural). In monaural source separation, only one recording is available, and the spatial information cannot generally be extracted. Moreover, in real-world room environments, the reverberations are challenging, which distort the received mixture and degrade the separation performance [7].

Y. Sun, and S. M. Naqvi are with the Intelligent Sensing and Communications Research Group, School of Engineering, Newcastle University, Newcastle upon Tyne NE1 7RU, U.K. (e-mails: Y.Sun29@newcastle.ac.uk; Mohsen.Naqvi@newcastle.ac.uk)

W. Wang is with the Center for Vision Speech and Signal Processing, Department of Electrical and Electronic Engineering, University of Surrey, Surrey GU2 7XH, U.K. (e-mails: W.Wang@surrey.ac.uk)

J. A. Chambers is with the Department of Engineering, University of Leicester, Leicester LE1 7RU, U.K. (e-mails: Jonathon.Chambers@leicester.ac.uk)

E-mail for correspondence: Mohsen.Naqvi@newcastle.ac.uk

Many approaches have been used to solve the monaural source separation problem in reverberant environments. Firstly, Delcroix *et al.* exploit the weighted prediction error (WPE) algorithm to achieve dereverberation in both single and multi-microphone cases [8]. Then, non-negative matrix factorization (NMF) is exploited to separate signals, which is a well established method for single channel speech separation [9]. Grais and Erdogan model the noisy observations based on weighted sums of non-negative sources [10]. However, when these methods are applied in real room environments, their performance and robustness are limited [11].

In the last decade, DNNs have been exploited for the monaural source separation problem and their performance has notable improvements. In the DNN-based techniques, the T-F masks or clean spectra are estimated by using the trained DNN model and applied to reconstruct the desired speech signal. According to the training objectives, DNN-based supervised monaural speech separation methods can be divided into two categories, namely mapping and masking techniques [12].

In the mapping-based DNN technique, the DNN is trained to generate the clean spectrum of the desired speech signal by using the spectrum of the mixture [12]. Han *et al.* train a DNN to learn a spectral mapping function between the reverberant noisy spectrum and the desired clean spectrum [13]. Huang *et al.* refine the mapping-based technique by introducing a deep recurrent neural network (DRNN) and discriminative criterion in the cost function [1]. In [14], Sun *et al.* further improve the mapping-based technique with the adaptive discriminative criterion. Compared with the masking-based technique, the mapping-based technique requires large memory and computational cost [15]. However, in real acoustic environments, it is difficult to obtain the desired speech signal consistently with high quality by using the above mapping-based methods [12]. In addition, in the traditional mapping-based techniques, the DNN is trained to obtain the desired speech signal directly from the mixture. The spectrum of the reverberant mixture is often more noisy than that of the dereverberated one due to the presence of reverberations and as a result, the DNN is much more difficult to train with a reverberant mixture in mapping-based approaches. Therefore, in this study, we focus on the masking-based technique.

In the masking-based DNN technique, the T-F mask is given and the estimated desired speech signal is obtained by using the predicted T-F mask. Jin and Wang exploit the DNN to generate an ideal binary mask (IBM) to separate the speech

mixture. But the IBM is a binary mask, and the associated hard decision causes loss in the separation performance [16]. Then, Wang *et al.* propose a soft mask, also known as the IRM, for which the T-F unit is assigned as the ratio of desired source energy to mixture energy [17] and the IRM-based method outperforms the IBM-based method. However, the above mentioned methods do not utilize the phase information of the desired signal when synthesizing the clean signal. Wang and Lim consider phase information to be unimportant in speech enhancement [18], but Erdogan *et al.* have shown that the phase information is beneficial to predict an accurate mask and the estimated source [19]. Consequently, in [11], [20], Williamson *et al.* employ both the magnitude and phase spectra to estimate the complex IRM (cIRM) by operating in the complex domain.

In the state-of-the-art methods, the ideal T-F mask is computed for dereverberated and reverberant mixtures in a slightly different way. In the dereverberant case, the ideal T-F mask is calculated by using the clean speech signal and the dereverberated mixture, while in reverberant environments, the T-F mask is calculated by using the direct sound and the reverberant mixture [11], [17]. Because the direct sound is a delayed and attenuated version of the original speech, it has negative influence on the accuracy of the corresponding T-F mask. Hence, the separation performance of these methods is degraded due to the influence of reverberations and the direct sound impulse response.

To address these issues, we propose a two-stage approach where one stage is exploited to attenuate the reflections, followed by another stage to separate the processed mixture.

In summary, the contributions of this paper are:

(1) A novel DM is proposed for dereverberation of the reverberant speech mixture. Different from the previous T-F masking-based method in reverberant environments, the DM we propose is used to eliminate the room reflections in the reverberant mixture, which allows a separation mask to be used for estimating the original speech sources from the dereverberated mixture.

(2) Two DNN-based methods are proposed with different training targets. The single training target in the first method is an enhanced T-F mask i.e. the IEM. In the second method, the DM and the IRM are trained separately.

The rest of the paper is organized as follows. In Section II, the background knowledge related to the proposed two-stage approach is described. Section III introduces the proposed DM and the two-stage approach. Section IV presents the experimental settings and results with the IEEE [21] and the TIMIT [22] corpora. The conclusions and future work are given in Section V.

II. Masking-based DNN for Monaural source separation

Recently, neural networks have been adopted as a regression model to solve the source separation problem, including the monaural case. In this section, the existing state-of-the-art masking-based methods will be described.

In the masking-based DNN, the training target is an ideal T-F mask, which is calculated by using the desired signal and the mixture. Assume that $s(m)$, $i(m)$ and $y(m)$ are the desired speech signal, the interference and the acquired mixture at discrete time m, respectively. The terms $h_s(m)$ and $h_i(m)$ are the RIRs for reverberant speech and interference, respectively. The convolutive mixture is expressed as:

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

where '$*$' indicates the convolution operator. By using the short time Fourier transform (STFT), the mixture is written as:

$$Y(t,f) = S(t,f)H_s(t,f) + I(t,f)H_i(t,f) \tag{2}$$

where $S(t,f)$, $I(t,f)$ and $Y(t,f)$ are the spectra of speech, interference and mixture, respectively. The qualities $H_s(t,f)$ and $H_i(t,f)$ are the RIRs for speech and interference at time frame t and frequency f, respectively.

By employing the ideal T-F mask $M(t,f)$, the spectrum of the clean speech can be reconstructed as:

$$S(t,f) = Y(t,f)M(t,f) \tag{3}$$

Because the IRM and the cIRM are the two targets often chosen in state-of-the-art masking-based DNN methods, in the next subsections, the IRM and the cIRM are briefly described.

A. Ideal Ratio Mask

If there is no RIR, the IRM for time frame t and frequency f can be expressed as [17]:

$$IRM(t,f) = \left(\frac{|S(t,f)|^2}{|S(t,f)|^2 + |I(t,f)|^2}\right)^{\beta} \tag{4}$$

where β is a tunable parameter to scale the mask, $|S(t,f)|$ and $|I(t,f)|$ denote the target speech signal and the noise interference magnitude spectra, respectively. Typically, the tunable parameter is selected as 0.5.

When the environment is reverberant, the direct sound at discrete time m is expressed as [11]:

$$d(m) = h_d(m) * s(m) \tag{5}$$

where $h_d(m)$ is the impulse response for the direct sound. Hence, the IRM for a reverberant environment in the time-frequency domain is expressed as [11]:

$$IRM_{rev}(t,f) = \left(\frac{|D(t,f)|^2}{|Y(t,f)|^2}\right)^{\beta} \tag{6}$$

where $|D(t,f)|$ and $|Y(t,f)|$ denote the direct sound and noisy reverberant mixture magnitude spectra, respectively.

The IRM is the soft mask, and it preserves the speech-dominant parts and suppresses the interference-dominant parts with soft decisions, which decreases the performance loss in speech separation. However, the limitation of the IRM is that the phase information of the clean speech signal is not used in speech reconstruction. To overcome this drawback, the cIRM is proposed, where the phase information of the speech mixture is considered [11], [20].

B. Complex Ideal Ratio Mask

The cIRM is a complex T-F mask which is obtained by using the real and imaginary components of the STFTs of the

desired speech signal and mixture [20].

To calculate the cIRM, the STFTs of the reverberant mixture, direct sound and cIRM are written as:

$$Y(t,f) = Y_r(t,f) + jY_c(t,f) \tag{7}$$

$$D(t,f) = D_r(t,f) + jD_c(t,f) \tag{8}$$

$$cIRM(t,f) = cIRM_r(t,f) + j{\cdot}cIRM_c(t,f) \tag{9}$$

where $j \triangleq \sqrt{-1}$ and the subscripts 'r' and 'c' indicate the real and the imaginary components in the STFTs, respectively.

By using the ideal cIRM, the desired speech signal can be separated from the mixture. The T-F unit of the cIRM is defined as:

$$cIRM(t,f) = \frac{Y_r(t,f)D_r(t,f) + Y_c(t,f)D_c(t,f)}{Y_r^2(t,f) + Y_c^2(t,f)} + j\frac{Y_r(t,f)D_c(t,f) - Y_c(t,f)D_r(t,f)}{Y_r^2(t,f) + Y_c^2(t,f)} \tag{10}$$

In highly reverberant room environments, the separation performance of the above mentioned methods is limited and also not robust [23]. There are two possible reasons: (1) Both IRM_{rev} and $cIRM$ are calculated based on the direct sound [11], which is the delayed and attenuated version of the clean speech signal, and the corresponding T-F mask is used to reconstruct the direct sound instead of the clean speech signal. (2) The presence of reverberation in the mixture degrades the estimation of the IRM_{rev} and $cIRM$, however, no explicit operation is considered to reduce the adverse effect of acoustic reflections on the estimation of the IRM_{rev} and $cIRM$.

Therefore, the DM and the two-stage approach are proposed to address the limitation and refine the separation performance.

III. Proposed Method

In this section, we present a new dereverberation mask and also develop two schemes for joint training of dereverberation and separation masks for improving the separation results for reverberant mixtures. Since the proposed DM is a real valued mask, for the convenience of fusion with the separation mask, we choose the IRM, which is also real-valued, instead of the cIRM, despite the fact that using cIRM may further improve the separation performance.

A. Dereverberation Mask

Estimating the separation mask directly from the reverberant mixture is challenging and the mask obtained is often noisy due to the presence of acoustic reflections. To address this issue, a DM is used to eliminate reverberation, and then the IRM is applied to separate the desired speech signal. According to (13), we rewrite the reverberant mixture as:

$$Y(t,f) = [S(t,f) + I(t,f)] \left(\frac{H_s(t,f)}{1 + \frac{I(t,f)}{S(t,f)}} + \frac{H_n(t,f)}{1 + \frac{S(t,f)}{I(t,f)}} \right) \tag{11}$$

Therefore, by using $Y(t,f)$ and $[S(t,f)+I(t,f)]$, the relationship between the reverberant and dereverberated mixtures is obtained. In our proposed method, we defined the DM as:

$$DM(t,f) = \left(\frac{H_s(t,f)}{1 + \frac{I(t,f)}{S(t,f)}} + \frac{H_n(t,f)}{1 + \frac{S(t,f)}{I(t,f)}} \right)^{-1} \tag{12}$$

In the training stage, the spectra of speech, noise and mixture with reverberations are available, therefore, the DM can be learned as:

$$DM(t,f) = \big[S(t,f)+I(t,f)\big] Y(t,f)^{-1} \tag{13}$$

From (13), it is clear that in the training stage, the training target $DM(t,f)$ can be calculated by using $S(t,f)$, $I(t,f)$ and $Y(t,f)$. Therefore, before the target signal is separated from the mixture, the DM is applied to the reverberant mixture to eliminate most of the reflections. In the training stage, the DM is compressed, and its value range is limited to be consistent with that of IRM, and thereby facilitate the fusion with IRM. According to (13), when there are no RIRs, the elements of the DM will all be ones and the proposed two-stage approach will be reduced to one-stage using only the estimated IRM.

According to (11) and (13), we see that the DM is a dereverberation operation. Thus, we have

$$S(t,f) + I(t,f) = Y(t,f)DM(t,f) \tag{14}$$

Because the DM can only dereverberate the speech mixture, further processing is required for separating the mixture. Compared with the cIRM, the IRM requires less computational cost and both the DM and the IRM are soft masks which are applied in the T-F domain, while the cIRM is applied in the complex domain. In this work, the IRM is applied to separate the desired signal from the mixture. The desired speech signal is extracted from the dereverberant mixture by using the IRM:

$$S(t,f) = \Big(S(t,f) + I(t,f) \Big) IRM(t,f) \tag{15}$$

In the proposed methods, according to the training targets and number of DNNs, the methods are categorized in two aspects, namely integrated training target and separate training targets methods.

B. Integrated Training Target

In the proposed DNN-based method with the integrated training target, only one DNN is trained and its training target is the IEM, which is generated by integrating the DM and the IRM as:

$$IEM(t,f) = DM(t,f)IRM(t,f) \tag{16}$$

Comparing the proposed IEM with the IRM_{rev}, the proposed single DNN method is essentially different from the one in [11]: the $IRMrev$ is calculated based on the direct sound, which is a delayed and attenuated version of the clean speech signal. Hence, after using the T-F mask, the STFT of the direct sound is obtained. However, in real scenarios, $h_d(m)$ in (5) is not equal to 1 and as a result, $IRMrev$ is not

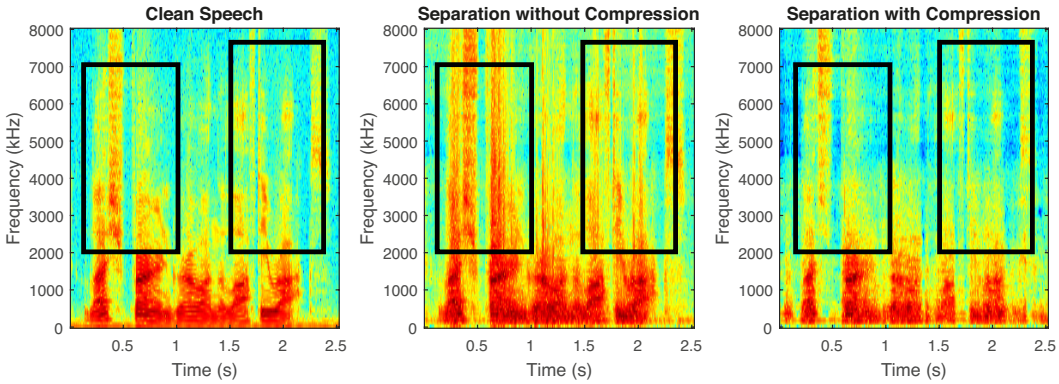


Fig. 1. Spectrogram plots of the clean speech signal (left), separated speech signal without compression module (middle) and separated speech signal with compression module (right). The reverberant mixture is generated with $factory$ noise and $0dB$ SNR level in the $unseen$ RIR case for $RT60 = 470ms$. The hyperparameters $C = 1$ and $V = 10$.

always effective in mitigating the reverberation effect. While in our proposed IEM, the IRM is calculated by using the clean speech signal and the dereverberant mixture, after using the T-F mask, the STFT of the clean speech signal can be obtained. Therefore, compared with the $IRMrev$, the IEM achieves better separation performance. In addition, the compression module is added to restrict the range of the values within the IEM, which is conducive for training the DNN.

According to (14) and (15), we see that the DM is a dereverberation operator and the IRM is the separation operator. Thus, the separated speech signal is obtained as:

$$S(t,f) = Y(t,f)IEM(t,f) \tag{17}$$

The value range of the proposed DM is (0, +∞), when the DM is integrated with the IRM as the training target, the value range of the DM is not consistent with IRM, and hence the mapping relationship is difficult to find. To address this issue, we use (18) to compress the DM to restrict its value range in order to make it consistent with the IRM and convert it back to the original value range in the testing stage by using (19). Empirically, in the training stage, the compressed IEM is written as:

$$IEM_c(t,f) = V\frac{1 - e^{-C\cdot IEM(t,f)}}{1 + e^{-C\cdot IEM(t,f)}} \tag{18}$$

where C is the steepness constraint and the value of $IEM_c(t,f)$ is limited in the range $[-V, V]$. Because the magnitude information is used to calculate the IEM, the value of $IEM_c(t,f)$ is restricted in the range $(0, V]$. After the validation tests in our experiments, the values of C and V are chosen as 1 and 10, respectively. These values were found based on the datasets described in the experimental section. For other datasets, C and V could be choosen in a similar way.

In the testing stage, the estimation of the compressed IEM is recovered and the final predicted IEM is expressed as:

$$\hat{IEM}(t,f) = -\frac{1}{C}\log\Big(\frac{V - O(t,f)}{V + O(t,f)}\Big) \tag{19}$$

where $O(t,f)$ is the estimation of the compressed IEM.

As an example, the spectrograms of the clean speech signal, the separated speech signal without compression module and the separated speech signal with compression module are shown in Figure 1. It can be seen that the compression module is important for the DM, which can eliminate noise in the high frequency component of the separated speech signal.

In the proposed two-stage approach, inspired by [11], [24], the feature combination is given to train the DNNs to refine the performance. The amplitude modulation spectrogram (AMS) [25], relative spectral transform and perceptual linear prediction (RASTA-PLP) [26], mel-frequency cepstral coefficients (MFCC), cochleagram response and their deltas are extracted by a 64-channel gammatone filterbank to obtain the compound feature [15]. The feature combination is extracted in the feature extraction module. To update the DNN weights, the backward propagation algorithm is exploited and the mean-square error (MSE) function is used in the cost function.

The cost function of the proposed single DNN-based method is expressed as:

$$J_1 = \frac{1}{2N}\sum_t\sum_f [O(t,f) - IEM_c(t,f)]^2 \tag{20}$$

where N represents the number of time frames for the inputs, $O(t,f)$ is the estimation of the compressed IEM and $IEM_c(t,f)$ is the compressed IEM at a T-F unit.

Figure 2 is the flow diagram of the proposed single DNN-based method with integrated training target, where (18) and (19) are achieved in the compression module and the recovery module, respectively. In the training stage, the DM and the corresponding IRM are calculated by using the target calculation module and integrated as the IEM. The IEM is compressed in the compression module to generate the training target of the single DNN. In the training stage, (18) is used to update the weights of the DNN. In the testing stage, once the trained DNN is obtained, the feature combination of the mixture is extracted and input to the trained DNN. The output of the DNN is obtained in the recovery module and used to separate the desired signal. Finally, the desired speech signal is separated from the convolutive mixture with the predicted IEM in the separation module.

It is clear to see the advantages of the proposed single DNN-based method with integrated training target:

(1) Only one DNN is trained, the computational cost and

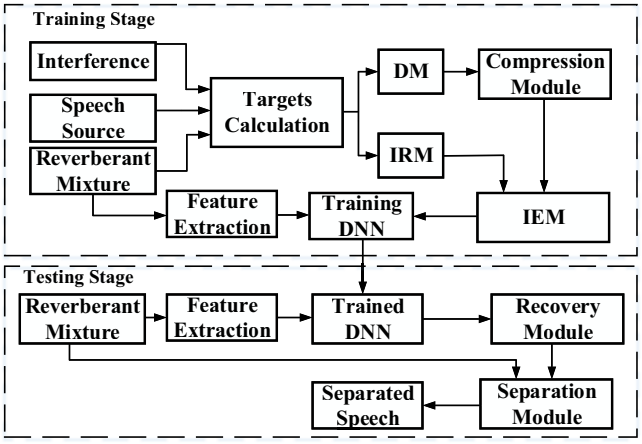


Fig. 2. The block diagram of the proposed single-DNN based method. One DNN is trained with the integrated training target i.e. IEM. The trained DNN is given by the training stage and in the testing stage, the output of the separation module is the desired speech signal.

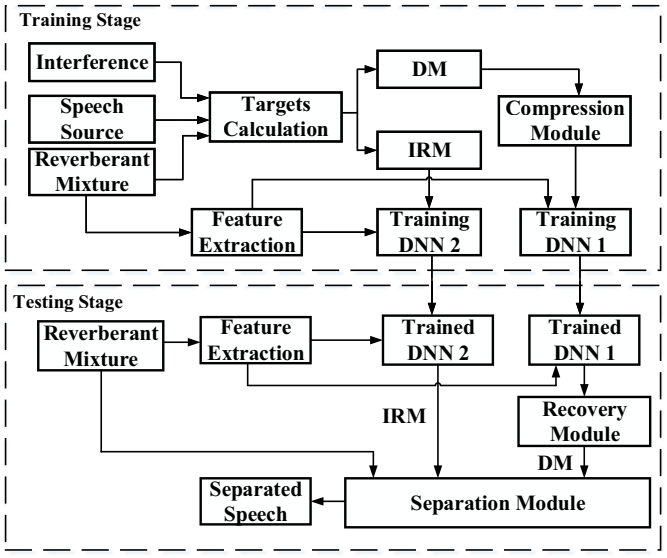


Fig. 3. The block diagram of the proposed two-DNN based method. Two DNNs are trained with the separate training targets. Two trained DNNs are found by the training stage. In the testing stage, the dereverberated speech mixture is obtained by using the predicted DM in the dereverberation module and the desired speech signal is obtained by using the predicted IRM in the separation module, respectively.

the storage space requirement will be lower than the method based on two training targets with two DNNs.

(2) The dereverberation and separation are achieved by the IEM, in the training stage, the estimation error will be decreased by generating the integrated training target. Compared with the traditional IRM, the IEM can achieve better separation performance because the DM is used to eliminate the reflection and the IRM is exploited to estimate the source from the dereverberated mixture.

C. Separate Training Targets

In the proposed second method, two DNNs are trained to model the relationships from the inputs to the DM and the IRM, respectively. In this method, the two T-F masks are predicted, the DM is applied for dereverberation, then the dereverberated mixture is separated by using the IRM. The compression and recovery processes are only applied to the DM, which is similar to the first method.

Assume the predicted dereverberation mask is $\hat{DM}(t,f)$ and the predicted ideal ratio mask is $\hat{IRM}(t,f)$, the separated speech signal is expressed as:

$$\hat{S}(t,f) = Y(t,f)\hat{DM}(t,f)\hat{IRM}(t,f) \tag{21}$$

Figure 3 is the flow diagram of the proposed two DNN-based method with separate training targets. Because the DM is predicted by the trained DNN, the compression module and the recovery module are essential. In the training stage, the compound features (discussed in Subsection III$-$B) extracted from the reverberant mixture are used as input to DNN2, where IRM is used as the the training target. The same compound features are used as input to DNN1, where DM (modified by the compression module) is used as the training target. In the testing stage, the reverberant mixture is used as input to estimate the DM and IRM, respectively. Since the reverberant mixture is used in the training stage for both DNN1 and DNN2, the trained network is able to generalise to reverberant mixtures in the testing stage.

$$J_2 = \frac{1}{2N}\sum_t\sum_f [O_1(t,f) - DM_c(t,f)]^2 \tag{22}$$

where $O_1(t,f)$ is the output of the DNN1 at a T-F unit and $DM_c(t,f)$ is the compressed DM at a T-F unit by using (18). Similarly, for DNN2, its cost function is expressed as:

$$J_3 = \frac{1}{2N}\sum_t\sum_f [O_2(t,f) - IRM(t,f)]^2 \tag{23}$$

where $O_2(t,f)$ is the output of the DNN2 at a T-F unit and $IRM(t,f)$ is the ideal ratio mask at a T-F unit.

In the testing stage, after the trained DNNs are obtained, the feature combination of the mixture is extracted and input to the trained DNNs. The output of the trained DNN1 is the predicted compressed DM and the output of the trained DNN2 is the predicted IRM. Then, the output of the DNN1 is obtained in the recovery module and used to eliminate the reflections. The mixture without reverberation is given by using the dereverberation module and the desired speech source is obtained from the separation module. Finally, the desired speech signal is separated from the convolutive mixture with the predicted DM and the predicted IRM.

As an example, we show some spectrogram plots in Figure 4 for the outputs from the different stages of the proposed method. It can be observed that by using the proposed DM, the reflections in the speech mixture can be eliminated. When the compression module is added (comparing (e) and (f) with (b)), the spectrogram of the separated signal with compression module is more similar to that of the clean speech signal. By adding the compression module, the noise in the high frequency component can be better removed.

In the proposed two-stage approach, before speech separation, the room reflections are better eliminated, therefore,

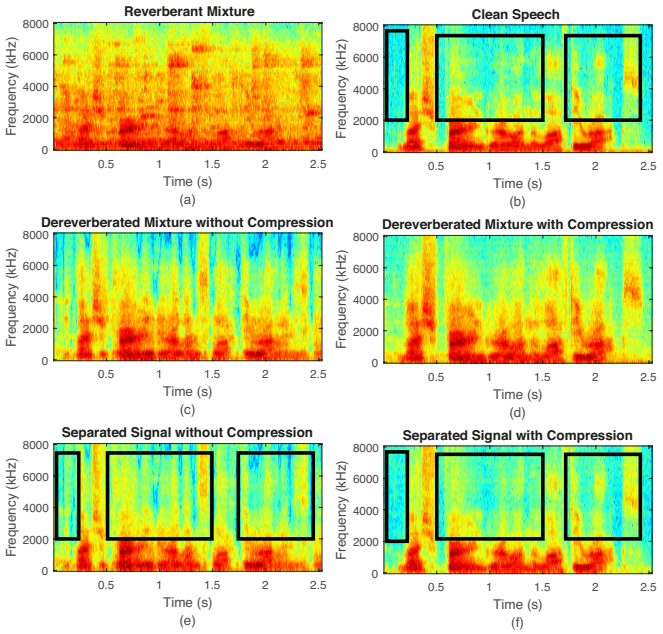


Fig. 4. Spectrograms of different signals: (a) reverberant mixture; (b) clean speech signal; (c) dereverberated mixture without compression; (d) dereverberated mixture with compression; (e) separated speech signal without compression and (f) separated speech signal with compression. The reverberant mixture is generated with $factory$ noise and $0dB$ SNR level in the $unseen$ RIR case for $RT60 = 470ms$. The hyperparameters $C = 1$ and $V = 10$.

the separation performance is improved. Therefore, in both single DNN and two DNNs methods, all factors including the training and testing datasets, the network architectures, hyperparameters and the input feature combination to train the DNNs are the same. It appears that only the training targets and the number of trained DNNs are different between these two proposed methods. Besides, because both the DM and the IRM are estimated, these two masks are more accurate, the performance is further improved with the trade-off of the computational cost.

IV. Experimental Results and Discussions

In this section, we evaluate the proposed two-stage approach with different training objectives, namely the integrated and the separate training targets. The interferences are selected as different types of noise and the undesired speech signals. Various RIRs are applied to generate the reverberant speech mixtures to show the performance in different reverberant room environments. In addition, the generalization ability of the proposed two-stage approach is evaluated with the unseen RIRs.

A. Experimental Settings

The speech sources are selected randomly from the IEEE [21] and the TIMIT corpora [22]. The IEEE corpus has 720 clean utterances spoken by a single male speaker and the TIMIT database has 6300 utterances, 10 utterances spoken by each of 630 speakers. Therefore, using both the IEEE and the TIMIT corpora can demonstrate that the proposed method is not speaker-dependent. The interferences are categorized into two aspects, the noise interference and the speech interference.

For noise interference, the noise signals are selected from the NOISEX database [27], in these noise signals, a speech-shaped noise (SSN) is generated as the stationary noise [28] and all others are the non-stationary noise, namely factory, babble and cafe. The factory noise is a recording of industrial activities and the babble noise is generated by different number of the unseen speakers in an acoustic environment. The cafe noise is more like a combination of babble and factory noise, it contains the speakers and background noise. The SSN is generated based on the clean speech corpus.

In our evaluation studies, in both training and testing stages, the target speech signals are randomly selected from the TIMIT dataset. Then, interfering speech signals are randomly selected from the remaining signals in the dataset to ensure the speakers of the target speech and the interfering speech signals are different. At the testing stage, the desired speech signals are unseen in the training stage, but the interfering speech signals are seen in the training stage. Therefore, the trained neural network is able to differentiate the target and undesirable speech signals.

To generate the speech mixture, the speech utterances and interferences are convolved with the real RIRs [29] which are recorded in four types of room environments i.e. different RT60s. The position of the desired speech signal is fixed and the azimuth of the interfering source is selected from 0° to 75° with 15° increment. Hence, each room has six different RIRs. In the evaluation with the seen RIRs, we use the RIRs from the same room to generate the training and testing datasets. In the evaluation with the unseen RIRs, for each room, four RIRs are randomly selected and used to generate the training data. The testing data are obtained by using the remaining two RIRs. Therefore, in the testing data, the RIRs are unseen and from different room environments. However, direct signals need to be generated for the baseline systems to enable comparisons with our proposed system. Firstly, the impulse response of the direct path is cropped from the whole impulse response. Then, the direct sounds are generated by using the impulse response of the direct path and clean speech signals in order to train the DNN models in [11]. Table I illustrates the parameters in the real RIRs: [29].

TABLE I
The parameters for real RIRs in different rooms [29]

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

In the experiments, we randomly select 1000, 100 and 120 utterances from the IEEE and the TIMIT corpora to generate the training, development and testing datasets. These clean utterances are used to mix with interference at three different signal-to-noise ratio (SNR) levels (-3 dB, 0 dB and 3 dB). In the evaluations with seen RIRs, the numbers of mixtures in

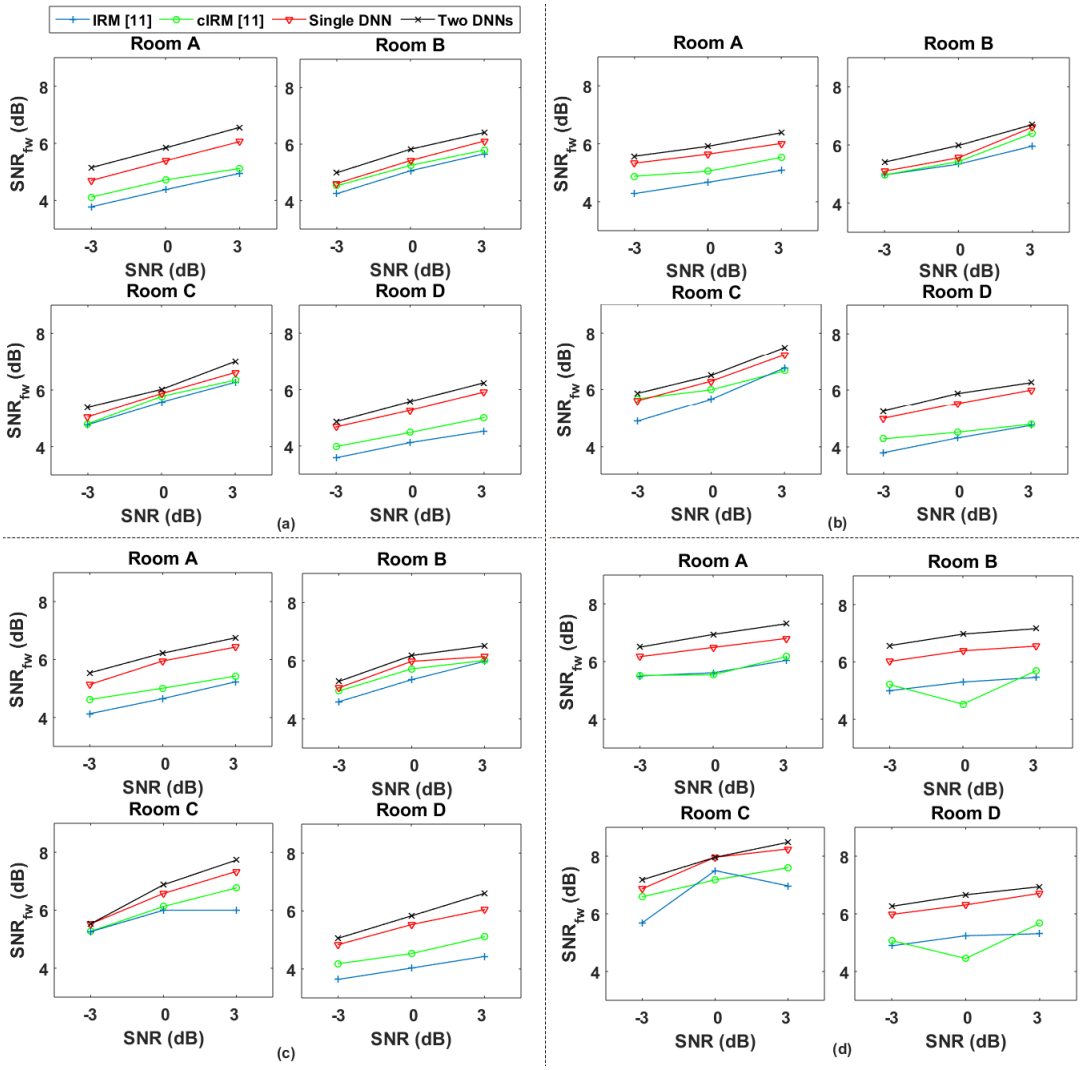


Fig. 5. The SNR$_{fw}$ (dB) in terms of different methods with various rooms. The X-axis is the SNR level, the Y-axis is the SNR$_{fw}$ (dB), each result is the average value of 120 experiments. The noise types in the subfigures (a), (b), (c) and (d) are factory, babble, cafe and SSN, respectively.

training, development and testing data are 72,000, 7,200 and 8,640, respectively. In the evaluation with the unseen RIRs, the numbers of mixtures in training, development and testing data are 192,000, 19,200 and 9,600, respectively.

In our proposed two-stage approach, the DNNs in the integrated training target and the separate training targets methods have the same architecture. All of the DNNs have three hidden layers and each hidden layer has 1024 units. The activation function for each hidden unit is selected as the rectified linear unit (ReLU) to avoid the gradient vanishing problem and the output layer has linear units [11]. The DNNs are trained by using the AdaGrad algorithm [30] with a momentum term for 100 epochs. The learning rate is linearly decreased from 1 to 0.01, while the momentum is fixed as 0.9 in the first ten epochs and changed to 0.5 till the end. Auto-regressive moving average (ARMA) filtering is applied to reduce the interference from the background noise, as in [31].

B. Comparisons and Performance Measures

We compare the proposed method with two state-of-the-art T-F masks: the IRM [17] and the cIRM [11]. Using different types of interferences, SNR levels and the RIRs in simulations show the performance of the proposed method is consistent. Moreover, when the training target is applied in the complex domain (cIRM), the corresponding DNN outputs the estimates of real and imaginary components of the predicted cIRM. The DNN needs to be Y-shaped, which has dual outputs with one input. The performance evaluation measures are the frequency-weighted segmental SNR (SNR$_{fw}$) [32], the source to distortion ratio (SDR) [33] and the short-time objective intelligibility (STOI) [34]. The SNR$_{fw}$ computes a weighted signal-to-noise ratio aggregated across each time frame and critical band, it is highly correlated to human speech intelligibility scores [11]. The SDR is exploited to evaluate the overall separation performance. The values of the STOI are in the range of [0, 1], which indicate the human speech intelligibility scores. The higher values of these metrics means that the desired speech signal is better reconstructed. In terms of the STOI, the t-test is also provided to show the significant difference. If the value of the t-test is smaller than 0.05, it indicates significant difference exists between two result sets. Besides, the IRM_{rev} and $cIRM$ in [11] are trained with

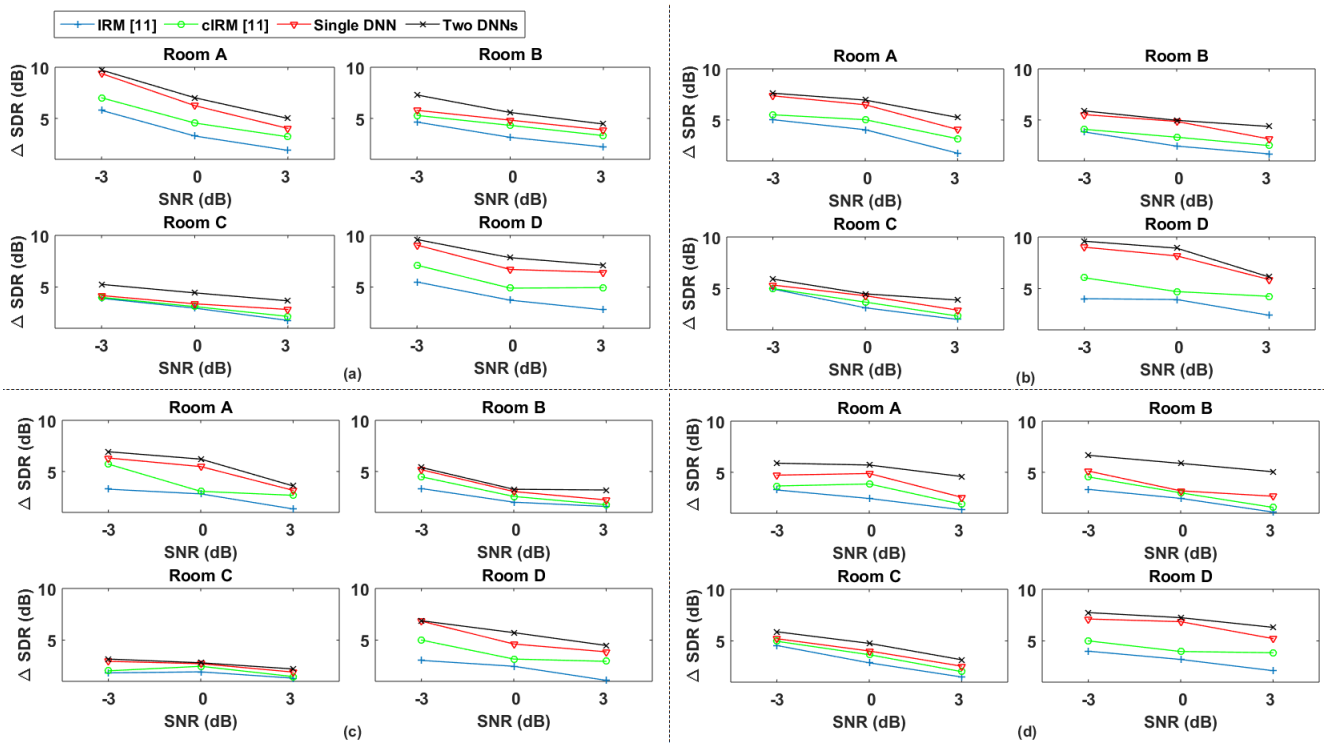


Fig. 6. The SDR improvement (dB) in terms of different methods with various rooms. The X-axis is the SNR level, the Y-axis is the ΔSDR (dB), the improvements of the SDR. Each result is the average value of 120 experiments. The noise types in the subfigures (a), (b), (c) and (d) are factory, babble, cafe and SSN, respectively.

direct sound, however, in real applications, the direct sound is difficult to obtain and the clean speech signal is used as reference in all performance measures.

C. Experimental Results and Analysis

The experimental results are shown in this subsection with noise and speech interferences. The proposed method is evaluated with the seen RIRs and the unseen RIRs under these two different interferences. Because in the first DNN-based method with integrated training target, only one DNN is trained, we use single DNN to represent this method. Similarly, two DNNs represents the second DNN-based method with separate training targets.

1) Experimental Results with Noise Interference: In this subsection, the noise is selected as the interference, and we use seen RIRs and unseen RIRs to generate the testing mixtures to further evaluate the generalization ability of the proposed methods.

a) Evaluations with the Seen RIRs: In these experiments, the proposed methods are evaluated with the seen RIRs in four rooms. The SNR_{fw} and the SDR performance of the proposed methods and the comparison groups are given in Figures 5 & 6, respectively. The STOI performance is shown in Tables II - V.

From Figures 5 & 6, it is clear that when the type of noise interference varies, the performance of the IRM and the cIRM-based methods is not consistent and robust. In the noise interference case, compared with the proposed two-stage approach with single DNN, the proposed two-stage approach with two DNNs produces better results for source separation from the convolutive mixture. In the high SNR level and low RT60, the proposed two-stage approach achieves high separation performance. Compared with the IRM- and the cIRM-based DNN methods, both our proposed methods provide improved performance in terms of the SNR_{fw} and SDR consistently.

To further analyze the proposed two-stage approach, the STOI performance is evaluated. The STOI performance of different methods using the IEEE and the TIMIT corpora with different noise and room environments are shown in Tables II - V.

It can be further confirmed that the proposed two-stage approach outperforms the state-of-the-art masking-based methods in different noise interference and reverberant environments from Tables II - V. With the increase of the RT60, the proposed methods give more STOI improvements. In some cases, the cIRM-based method gives the same STOI performance as or does slightly better than the proposed methods, e.g. SSN is used as interference with 0 SNR level in Room C. In terms of the average result, however, the proposed two-stage approach achieves the highest value. The trend of the STOI is the same as that of the SNR_{fw} and the SDR.

To show the difference of the STOI performance between the cIRM-based method and the proposed method with two DNNs, the t-test is used. *For example, in Room D, the value of the t-test with cafe noise and SSN noise is 0.01 and 0.02, respectively.* It means in Room D, when the noise type is cafe and SSN, the STOI performance of the proposed method with two DNNs and the cIRM-based are significantly different from each other.

TABLE II
SEPARATION PERFORMANCE COMPARISON IN TERMS OF STOI WITH DIFFERENT TRAINING TARGETS, SNR LEVELS AND RT60S. THE NOISE IN THE EXPERIMENTS IS ***factory*** NOISE. EACH RESULT IS THE AVERAGE VALUE OF 120 EXPERIMENTS. **BOLD** INDICATES THE BEST RESULT.

Factory	Room A (0.32 s)			Room B (0.47 s)			Room C (0.68 s)			Room D (0.89 s)		
Noise	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB
Mixture	0.54	0.59	0.64	0.52	0.56	0.61	0.54	0.60	0.64	0.46	0.49	0.51
IRM [11]	0.66	0.71	0.76	0.64	0.69	0.73	0.67	0.71	0.77	0.60	0.63	0.66
cIRM [11]	0.66	0.72	**0.77**	0.65	0.69	0.74	0.67	0.73	0.77	0.61	0.64	0.68
Single DNN	**0.68**	0.72	**0.77**	**0.66**	0.72	0.76	0.67	**0.74**	**0.78**	**0.63**	**0.69**	0.73
Two DNNs	**0.68**	**0.73**	**0.78**	**0.66**	**0.73**	**0.77**	**0.68**	**0.74**	**0.78**	**0.63**	**0.69**	**0.74**

TABLE III
SEPARATION PERFORMANCE COMPARISON IN TERMS OF STOI WITH DIFFERENT TRAINING TARGETS, SNR LEVELS AND RT60S. THE NOISE IN THE EXPERIMENTS IS ***babble*** NOISE. EACH RESULT IS THE AVERAGE VALUE OF 120 EXPERIMENTS. **BOLD** INDICATES THE BEST RESULT.

Babble	Room A (0.32 s)			Room B (0.47 s)			Room C (0.68 s)			Room D (0.89 s)		
Noise	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB
Mixture	0.54	0.59	0.65	0.53	0.58	0.62	0.55	0.61	0.66	0.47	0.49	0.51
IRM [11]	0.69	0.73	0.77	0.68	0.70	0.73	0.71	0.74	0.78	0.63	0.65	0.66
cIRM [11]	0.70	0.73	0.77	0.67	0.72	0.74	0.71	0.74	0.76	0.65	0.66	0.72
Single DNN	0.70	**0.75**	0.77	0.68	**0.74**	0.74	**0.73**	**0.76**	**0.79**	**0.67**	0.70	0.74
Two DNNs	**0.71**	**0.75**	**0.79**	**0.69**	**0.74**	**0.77**	**0.73**	**0.76**	**0.79**	**0.67**	**0.71**	**0.75**

TABLE IV
SEPARATION PERFORMANCE COMPARISON IN TERMS OF STOI WITH DIFFERENT TRAINING TARGETS, SNR LEVELS AND RT60S. THE NOISE IN THE EXPERIMENTS IS ***cafe*** NOISE. EACH RESULT IS THE AVERAGE VALUE OF 120 EXPERIMENTS. **BOLD** INDICATES THE BEST RESULT.

Cafe	Room A (0.32 s)			Room B (0.47 s)			Room C (0.68 s)			Room D (0.89 s)		
Noise	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB
Mixture	0.59	0.65	0.69	0.57	0.62	0.67	0.61	0.66	0.72	0.48	0.51	0.57
IRM [11]	0.67	0.73	0.76	0.65	0.70	0.74	0.68	0.74	0.79	0.58	0.62	0.65
cIRM [11]	**0.68**	0.76	0.79	0.66	0.71	0.75	0.68	0.75	0.80	0.58	0.63	0.65
Single DNN	**0.68**	0.76	0.79	**0.67**	**0.75**	**0.78**	**0.69**	**0.76**	**0.81**	0.60	0.70	0.73
Two DNNs	**0.68**	**0.77**	**0.80**	**0.67**	**0.75**	**0.78**	**0.69**	**0.76**	**0.81**	**0.65**	**0.71**	**0.76**

TABLE V
SEPARATION PERFORMANCE COMPARISON IN TERMS OF STOI WITH DIFFERENT TRAINING TARGETS, SNR LEVELS AND RT60S. THE NOISE IN THE EXPERIMENTS IS ***SSN*** NOISE. EACH RESULT IS THE AVERAGE VALUE OF 120 EXPERIMENTS. **BOLD** INDICATES THE BEST RESULT.

SSN	Room A (0.32 s)			Room B (0.47 s)			Room C (0.68 s)			Room D (0.89 s)		
Noise	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB
Mixture	0.60	0.65	0.70	0.59	0.64	0.68	0.62	0.67	0.73	0.51	0.53	0.56
IRM [11]	0.78	0.80	0.81	0.76	0.78	0.79	0.78	**0.82**	0.84	0.70	0.72	0.73
cIRM [11]	0.72	0.77	0.80	0.76	0.79	0.80	**0.79**	0.81	0.85	0.71	0.74	0.75
Single DNN	0.78	0.81	0.82	0.77	**0.80**	**0.81**	**0.79**	**0.82**	**0.86**	0.74	0.76	0.77
Two DNNs	**0.79**	**0.82**	**0.84**	**0.78**	**0.80**	**0.81**	**0.79**	**0.82**	**0.86**	**0.75**	**0.77**	**0.80**

From Figures 5 & 6 and Tables II - V, it is clear that with the same amount of training data and DNN configurations, the separation performance of the current state-of-the-art is not consistent and robust when the SNR levels and noise types are varied. The two-stage approach, we proposed, can yield effective performance. Thanks to the DM applied to the mixture, when the RT60 is increased, the relative STOI improvements becomes more prominant at higher RT60s. Compared the masking-based techniques with the proposed two-stage approach, the experimental results demonstrate that using two DNNs in the proposed two-stage approach can further improve the separation performance.

b) Evaluations with the Unseen RIRs: In these experiments, the proposed two-stage approach is evaluated with unseen RIRs. The SNR_{fw} and the SDR performance of the proposed methods and the compared methods are given in Figures 7 & 8, respectively. The STOI performance of different methods using the IEEE and the TIMIT corpora with different noise and the unseen RIRs are shown in Table VI. In the experiments with the unseen RIRs, the RIRs used in the testing stage are different from those in the training stage.

Figure 7 shows the SNR_{fw} performance in terms of different methods with the unseen RIRs. It can be observed that compared with the IRM and the cIRM, the proposed methods, both single DNN and two DNNs, yield better performance. When the value of SNR level is increased, the performance of SNR_{fw} is refined. Besides, it is observed from the figure that when two DNNs are trained, the values of the SNR_{fw} become higher. For example, according to Figure 7, when the noise type is SSN and the SNR level is 3 dB, the SNR_{fw} value of the IRM-based method is 2.99 dB and the cIRM-based method is 3.32 dB, but the proposed approach with single DNN and

two DNNs achieve 3.66 dB and 4.78 dB, respectively.

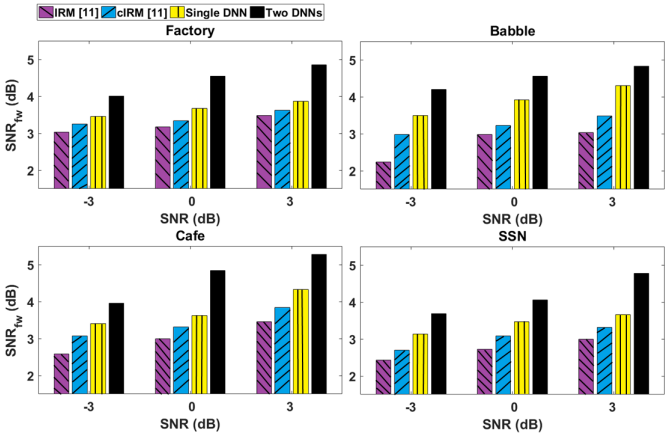


Fig. 7. The SNR$_{fw}$ (dB) in terms of different methods with the unseen RIRs. The X-axis is the SNR level, the Y-axis is the SNR$_{fw}$ (dB), each result is the average value of 120 experiments. The experimental results with four different types of noise are shown.

Figure 8 shows the SDR improvements over all types of noise with the unseen RIRs. It is observed that the proposed two-stage approach further refines the SDR performance (ΔSDR) when compared with the current state-of-the-art methods. In the situation where the RIRs are unseen, with increasing the SNR level, the improvement of the SDR becomes larger and the proposed two-stage approach provides the best performance. It is clear that by training two DNNs in the proposed two-stage approach, the value of the SDR improvement is increased significantly.

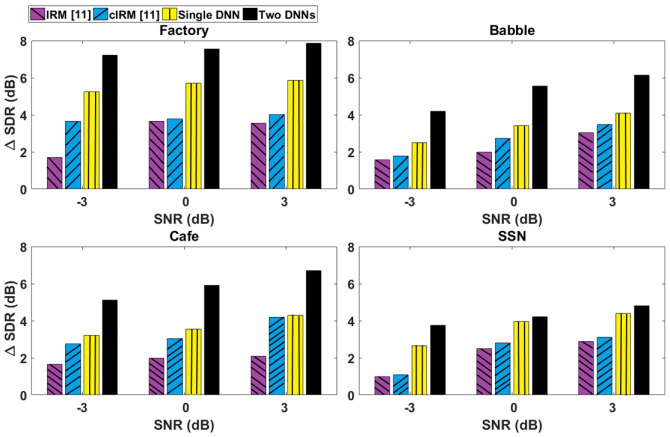


Fig. 8. The SDR improvement (dB) in terms of different methods with the unseen RIRs. The X-axis is the SNR level, the Y-axis is the SDR improvement (dB), each result is the average value of 120 experiments. The experimental results with four different types of noise are shown.

The experimental results in terms of the STOI are shown in three different SNR levels in Table VI. As the value of SNR level is increased, the performance of the STOI is improved. From Table VI, it is clear that with the same amount of training data and DNN configurations, when the RIRs are unseen, in terms of the STOI, the separation performance of the current state-of-the-art is not consistent and robust when the SNR levels and noise types are varied. *For all types of the noise, the value of the t-test in the STOI results with the unseen RIRs between the cIRM-based method and the proposed method with single DNN and two DNNs is 0.02 and 0.0004, respectively.* It confirms that the proposed two-stage approach outperforms the current state-of-the-art methods in terms of the STOI.

From Figures 7 & 8 and Table VI, it can be observed that the proposed two-stage approach can yield effective performance and using two DNNs in the proposed two-stage approach provides the best separation results. Using the noise and unseen RIRs, the proposed methods show better generalization ability. In the testing stage, since the RIR is unseen, compared with the seen RIRs case, the values of the corresponding SNR$_{fw}$, SDR and STOI are smaller.

2) Experimental Results with Speech Interference: After the evaluations of the proposed two-stage approach with noise interference, the undesired speech signal is exploited as the interference to generate the convolutive mixture.

a) Evaluations with the Seen RIRs: The interfering speech signal is chosen from the above mentioned corpora and both male and female speakers are used. The SNR$_{fw}$ and the SDR performance of the proposed methods and the comparison groups are given in Figures 9 & 10, respectively. The STOI performance of different methods are shown in Table VII.

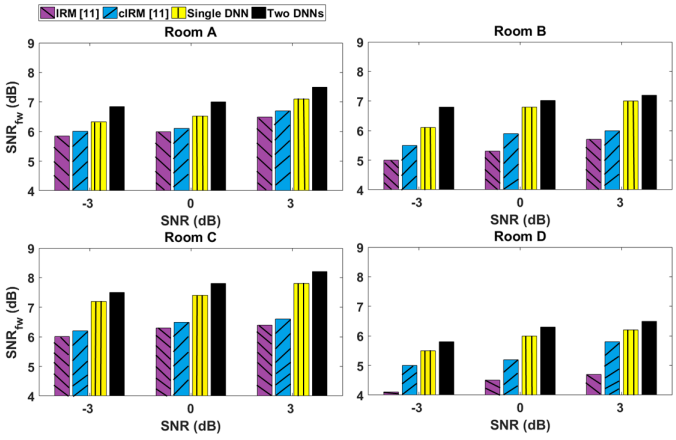


Fig. 9. The SNR$_{fw}$ (dB) in terms of different methods with various rooms i.e. different RT60s. The X-axis is the SNR level, the Y-axis is the SNR$_{fw}$ (dB), each result is the average value of 120 experiments. The interference is the undesired speech signal, respectively.

For the SNR$_{fw}$, shown in Figure 9, the proposed two DNN-based method further improves the performance relative to the separated desired speech signal. The largest SNR$_{fw}$ gains in all room environments are achieved by the proposed two DNN-based method. For example, at 3 dB SNR level, from Rooms A to D, the proposed method with two DNNs gives 16.1%, 21.8%, 22.3% and 13.7% more gain, respectively.

Besides, according to Figure 9, it confirms that the higher SNR level helps the two-stage approach to better separate the desired speech signal from the mixture with speech interference. Compared the performance with different SNR levels in terms of the SNR$_{fw}$, when the SNR levels increases (from -3 dB to 3 dB), the separation performance is improved, which is the same as the situations with noise interferences. For different RT60s, when the RT60 increases, e.g. Room A and Room D, the value of the SNR$_{fw}$ is decreased.

TABLE VI
SEPARATION PERFORMANCE COMPARISON IN TERMS OF STOI WITH THE UNSEEN RIRS. DIFFERENT TRAINING TARGETS, SNR LEVELS AND RT60S WITH ALL TYPES OF NOISE ARE EVALUATED. EACH RESULT IS THE AVERAGE VALUE OF 120 EXPERIMENTS. **BOLD** INDICATES THE BEST RESULT.

Noise Type	Factory			Babble			Cafe			SSN		
SNR Levels	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB
Mixture	0.46	0.48	0.50	0.47	0.49	0.52	0.49	0.51	0.54	0.50	0.53	0.55
IRM [11]	0.52	0.55	0.56	0.52	0.54	0.55	0.51	0.53	0.57	0.51	0.55	0.59
cIRM [11]	0.57	0.59	0.63	0.54	0.57	0.58	0.52	0.55	0.59	0.53	0.57	0.63
Single DNN	0.62	0.64	0.65	0.58	0.61	0.64	0.57	0.61	0.64	0.57	0.61	0.67
Two DNNs	**0.68**	**0.71**	**0.74**	**0.64**	**0.69**	**0.73**	**0.64**	**0.70**	**0.75**	**0.64**	**0.67**	**0.72**

TABLE VII
SEPARATION PERFORMANCE COMPARISON IN TERMS OF STOI WITH DIFFERENT TRAINING TARGETS, SNR LEVELS AND RT60S. THE INTERFERENCE IN THE EXPERIMENTS IS ***the undesired speech signal***. EACH RESULT IS THE AVERAGE VALUE OF 120 EXPERIMENTS. **BOLD** INDICATES THE BEST RESULT.

Speech Interference	Room A (0.32 s)			Room B (0.47 s)			Room C (0.68 s)			Room D (0.89 s)		
	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB
Mixture	0.58	0.63	0.67	0.54	0.59	0.63	0.58	0.64	0.66	0.48	0.50	0.51
IRM [11]	0.76	0.78	0.79	0.72	0.73	0.75	0.78	0.79	0.81	0.60	0.61	0.62
cIRM [11]	0.77	0.78	0.80	0.74	0.75	0.76	0.79	0.80	0.81	0.63	0.64	0.64
Single DNN	0.78	0.80	0.82	0.76	0.80	0.81	0.79	0.81	0.83	0.71	0.73	0.75
Two DNNs	**0.80**	**0.82**	**0.84**	**0.79**	**0.81**	**0.82**	**0.81**	**0.82**	**0.84**	**0.74**	**0.75**	**0.78**

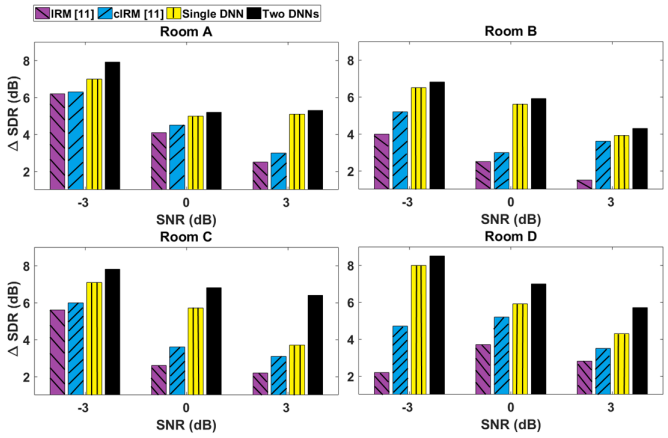


Fig. 10. The SDR improvement (dB) in terms of different methods with various rooms i.e. different RT60s. The X-axis is the SNR level, the Y-axis is the ΔSDR (dB), the improvements of the SDR. Each result is the average value of 120 experiments. The interference is the undesired speech signal, respectively.

Figure 10 displays the SDR improvements over all room environments. It is observed that the proposed two-stage approach significantly improves the SDR performance (ΔSDR), especially in the highly reverberant room environments such as Room C and Room D. With increasing the SNR level, the improvement of the SDR becomes smaller, but the proposed two DNN-based method still provides better results. In Room C, with 0.68 s RT60, compared with the cIRM, the proposed method with single DNN has 1.01 dB, 1.71 dB and 0.49 dB more improvements and the proposed method with two DNNs has 1.81 dB, 3.27 dB and 3.67 dB from -3 dB to 3 dB SNR levels, respectively.

From Table VII, it is clear that the two DNN-based method always gives the best performance in the case where the interference is a speech signal. For example, in Room D, the proposed method with two DNNs achieves 13.1%, 8.7% and 12.5% STOI improvements over the proposed method with single DNN (integrated training objective) at -3, 0 and 3 dB SNR levels, respectively. The two DNN-based method provides around 13.9% more STOI improvement in all scenarios. *When the undesired speech signal is the interference, the value of the t-test in the STOI results with the seen RIRs between the cIRM-based method and the proposed method with two DNNs is 0.008.* It proves that the proposed method with two DNNs yields better separation performance in terms of the STOI than the current state-of-the-art methods, e.g. cIRM-based method.

b) Evaluations with the Unseen RIRs: The interfering speech signal is chosen from the IEEE and the TIMIT corpora and both male and female speakers are used. The SNR_{fw} and the SDR performance of the proposed methods and the comparison groups are given in Figures 11 & 12, respectively. The STOI performance of different methods using the above mentioned corpora with different undesired speech signal and the unseen RIRs are shown in Table VIII.

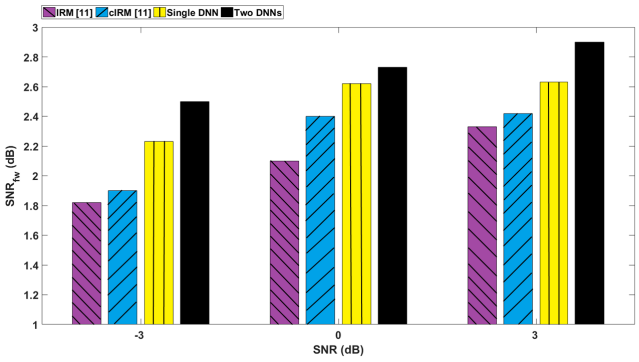


Fig. 11. The SNR_{fw} (dB) in terms of different methods with the unseen RIRs. The X-axis is the SNR level, the Y-axis is the SNR_{fw} (dB), each result is the average value of 120 experiments. The interference is the undesired speech signal, respectively.

For the SNR_{fw}, shown in Figure 11, the proposed two-stage

approach provides the largest performance improvements with the unseen RIRs scenarios. The largest SNR_{fw} gains in all SNR levels are achieved by the proposed two-stage approach with separate training targets. According to Figure 11, the proposed two-stage approach with integrate training target can achieve higher value of the SNR_{fw} and by training two DNNs in the proposed method, the separation performance is further improved.

Figure 12 shows the SDR improvements (ΔSDR) over all SNR levels with the unseen RIRs. It is observed that the proposed two-stage approach significantly improves the SDR performance, especially with higher SNR levels. With increasing the SNR level, the improvement of the SDR becomes larger and the proposed two DNN-based method achieves better separation results. For instance, when the SNR level is 3 dB, the value of ΔSDR of the proposed method with separate training objectives is 5.05 dB, while the value of the cIRM-based and the IRM-based method is 3.06 dB and 2.41 dB, respectively. It is clear that by training two DNNs in the proposed two-stage approach, the separation performance is increased significantly. In contrast to the evaluations with the seen RIRs, when the RIRs are unseen and the RT60 increases, the value of the SDR improvement increases, which are the same as the situations with noise interferences.

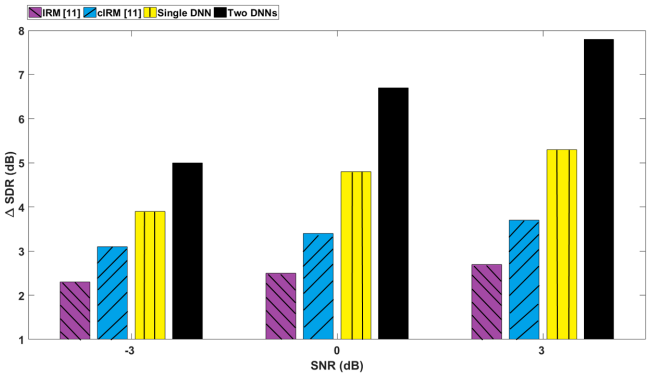


Fig. 12. The SDR improvement (dB) in terms of different methods with the unseen RIRs. The X-axis is the SNR level, the Y-axis is the ΔSDR (dB), the improvements of the SDR. Each result is the average value of 120 experiments. The interference is the undesired speech signal, respectively.

TABLE VIII
SEPARATION PERFORMANCE COMPARISON IN TERMS OF STOI WITH DIFFERENT TRAINING TARGETS, SNR LEVELS AND THE UNSEEN RIRS. THE INTERFERENCE IN THE EXPERIMENTS IS ***the undesired speech signal***. EACH RESULT IS THE AVERAGE VALUE OF 120 EXPERIMENTS. **BOLD** INDICATES THE BEST RESULT.

Speech Interference	STOI		
	-3 dB	0 dB	3 dB
Mixture	0.52	0.57	0.59
IRM [11]	0.56	0.59	0.64
cIRM [11]	0.59	0.61	0.66
Single DNN	0.65	0.69	0.73
Two DNNs	**0.70**	**0.72**	**0.76**

When the interference is the undesired speech signal, Table VIII, it is clear to observe that in terms of the STOI, the proposed two-stage approach outperforms current state-of-the-art. For example, compared with the cIRM, the proposed method with single DNN has 0.06, 0.08 and 0.07 improvements and the proposed method with two DNNs has 0.11, 0.11 and 0.1 improvements from -3 dB to 3 dB SNR levels, respectively. *When the undesired speech signal is the interference, the value of the t-test in the STOI results between the cIRM-based method and the proposed method with two DNNs is 0.01.* Hence, by using two DNNs in the proposed method, the value of STOI is the highest over all of the SNR levels.

3) Processing Time: Since two system structures of the proposed two-stage approach are exploited in this work, their processing time is different. In Section IV-A, the experimental settings in the proposed methods are the same, in order to evaluate their processing time, all of the DNN-based methods are executed ten times and their processing time is averaged. The evaluation results are shown in Table IX.

TABLE IX
AVERAGED PROCESSING TIME OF THE DNN-BASED METHODS WITH DIFFERENT TRAINING TARGETS. THE TIME OF TRAINING STAGE AND TESTING STAGE ARE SHOWN IN SECONDS.

Training Target in DNN-based Method	Processing Time (s)	
	Training Stage	Testing Stage
IRM [11]	8,398.8	37.4
cIRM [11]	8,655.4	43.1
IEM	8,443.4	39.8
DM & IRM	16,651.9	48.5

The codes of the IRM, cIRM and the proposed methods were written in MATLAB (R2015a version) without any optimization. The experiments were implemented on a desktop with an Intel i5 CPU with 3.5 GHz and 16 GB of memory without parallel processing. In the training and testing stages, no GPU was used.

It is observed from Table IX that in the training stage, the processing time of the proposed method with single training target (integrated objective) is half of the one with two training targets (separate objectives). Because in the second method, two DNNs are trained and these DNNs have the same architectures as the DNN in the first proposed method. While compared with the training stage, in the testing stage, the difference of the processing time with these methods can be ignored. The IRM-based method and the proposed IEM almost have the same processing time. Moreover, because the Y-shaped DNN was used in the cIRM-based method, its processing time is slightly higher than the IRM- and the IEM-based approaches. In the testing stage, all of these methods have a relative lower processing time.

Hence, the proposed two DNN-based method needs longer processing time and the computational cost is almost double than the single training target based method.

In summary, according to Figures 5 - 12 and Tables II - IX, the proposed two-stage approach outperforms state-of-the-art IRM- and the cIRM-based methods, particularly in reverberant room environments. When the RIRs are seen, the noise and undesired speech signal are used as the interferences in the mixture, all the experimental results further confirm that our proposed two-stage approach is effective in separating mixtures at various SNR levels and with different room environments. When the RIRs are unseen, the generalization

ability of the proposed method is evaluated, the results shown in Figures 7, 8, 11 & 12 and Tables VI & VIII confirm that the proposed method can better separate the desired speech signal from mixture than the IRM- and cIRM-based methods. There are two possible reasons that the proposed method has better generalization ability: (1) The compression and recovery modules are conducive for training the DNNs and thus leading to better prediction of the DM from the mixtures. (2) The use of DM can mitigate the adverse effect of acoustic reflections on the estimation of the IRM_{rev} and $cIRM$ for separating target speech from the mixture. As a result, the proposed method has better ability in adapting to unseen RIRs and leading to improved performance in such scenarios.

In addition, using the proposed two DNN-based method, the mixture can be better separated than just utilizing the IEM as integrated training target in the single DNN. From the results, it can be seen that the cIRM had worse performance than IRM in some cases. For example, in Table III, when the noise type is babble and the SNR level is -3 dB in Room B, the STOI performance of the cIRM is 0.67, while the IRM produces 0.68 STOI. It is our belief, this might be caused by the DNN architecture and how it is trained. To estimate the real and imaginary part of the cIRM jointly, the Y-shaped DNN was used. In this architecture, the weights of the hidden layers are shared by the real and imaginary parts of the cIRM and only two sub-output layers are used to distinguish the estimations of real and imaginary components of the cIRM. Hence, compared with the IRM, the cIRM-based DNN is more difficult to train, in order to provide balance for both the real and imaginary part. This can lead to degradation in separation performance.

It is worth noting that although the RT60 of Room C (RT60 = 680 ms) is higher than Room B (RT60 = 470 ms), the separation performance for Room C is better than that for Room B. This is mainly due to the difference in the Direct to Reverberant Ratio (DRR) where the DRR from Room C is higher than that for Room B.

From Table IX, in the proposed method with different training targets, when the DM and the IRM are trained individually, the computational cost is increased almost two times. Therefore, there is a trade-off between the computational cost and the separation performance. If two-DNNs are trained in the proposed two-stage approach, the separation performance is further refined, but more computational cost and storage space are required.

V. Conclusions and Future Work

In this paper, the two-stage approach with different training targets (integrated and separate) were proposed to address the monaural source separation problem. In the reverberant room environments, the separation performance was refined by adding the dereverberation stage before separating the desired speech signal from the mixture. The proposed methods were evaluated using the SNR_{fw}, SDR and STOI, for speech signals selected from the IEEE and the TIMIT databases with different interferences (the undesired speech signal, the stationary and the non-stationary noise). Besides, the RIRs are categorized into the seen and the unseen to evaluate the generalization ability of the proposed two-stage approach. Results showed that the proposed two-stage approach outperformed the IRM- and the cIRM-based approaches in all of the tested scenarios and the generalization ability of the proposed method was robust. Because the dereverberation stage was used to eliminate the reflections in the mixture, when the reverberant room environments had a higher RT60, the performance improvement of the proposed methods were more significant. In comparing the proposed methods with different training targets, the method with two DNNs gave further improvements, but the computational cost was almost doubled. Therefore, there is a trade-off between the computational requirement and the separation performance.

To further improve the performance, one direction is to explore the use of the advanced architecture neural networks such as the recurrent neural network (RNN), long-short term memory (LSTM) RNN and the DRNN to train the DM and the IEM, which exploits more temporal information in the models. Another direction is to apply the proposed DM in the complex domain and use the cIRM to separate the mixture.

Acknowledgement

The authors would like to thank the Associate Editor and the anonymous reviewers for their valuable input to improve this paper.

References

[1] 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.

[2] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.

[3] 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.

[4] B. Rivet, W. Wang, S. M. Naqvi, and J. A. Chambers, "Audiovisual speech source separation: An overview of key methodologies," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 125–134, 2014.

[5] M. S. Salman, S. M. Naqvi, A. Rehman, W. Wang, and J. A. Chambers, "Video-aided model-based source separation in real reverberant rooms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1900–1912, 2013.

[6] 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.

[7] Z. Y. Zohny, S. M. Naqvi, and J. A. Chambers, "Variational EM for clustering interaural phase cues in messl for blind source separation of speech," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[8] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, T. Hori, T. Nakatani, and A. Nakamura, "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the REVERB challenge," in *Proc. of REVERB Challenge*, 2014.

[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] E. M. Grais and H. Erdogan, "Single channel speech music separation using nonnegative matrix factorization and spectral masks," in *Proc. of IEEE International Conference on Digital Signal Processing (DSP)*, 2011.

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

[12] X. L. Zhang and D. L. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 967–977, 2016.

[13] 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.

[14] Y. Sun, L. Zhu, J. A. Chambers, and S. M. Naqvi, "Monaural source separation based on adaptive discriminative criterion in neural networks," in *Proc. of IEEE International Conference on Digital Signal Processing (DSP)*, 2017.

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

[16] Z. Jin and D. L. Wang, "A supervised learning approach to monaural segregation of reverberation speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 625–638, 2009.

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

[18] D. L. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, no. 4, pp. 679–681, 1982.

[19] H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[20] D. S. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.

[21] IEEE Audio and Electroacoustics Group, "IEEE recommended practice for speech quality measurements," *IEEE Transactions on Audio Electroacoust*, vol. 17, no. 3, pp. 225–246, 1969.

[22] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, *DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM*, 1993.

[23] 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," *Proc. of the 26th European Signal Processing Conference (EUSIPCO)*, 2018.

[24] Y. Wang, K. Han, and D. L. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 270–279, 2013.

[25] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listener," *Journal of the Acoustical Society of America*, vol. 126, pp. 1486–1494, 2009.

[26] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 2, no. 4, pp. 149–155, 1990.

[27] A. Varga and H. Steeneken, "Assessment for automatic speech recognition NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, pp. 247–251, 1993.

[28] S.-H. Jin and C. Liu, "English sentence recognition in speech-shaped noise and multi-talker babble for English-, Chinese-, and Korean-native listeners," *Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. 391–397, 2012.

[29] C. Hummersone, *Binaural Room Impulse Response Measurements*, Surrey University , United Kingdom , 2011.

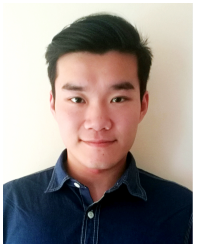
[30] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, pp. 2121–2159, 2011.

[31] C. Chen and J. A. Blimes, "MVA processing of speech features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 257–270, 2007.

[32] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[33] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transanctions on Audio Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[34] 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*, vol. 19, no. 7, pp. 2125–2136, 2011.



Yang Sun (S'17) received the B.Sc. degree in communication engineering from the Zhengzhou University, Zhengzhou, China, in 2014. The M.Sc. degree in communications and signal processing from Newcastle University, Newcastle Upon Tyne, U.K., in 2015. He is currently pursuing the Ph.D. degree within Intelligent Sensing and Communications (ISC) Research Group, School of Engineering, Newcastle University, U.K. His research areas of interest include audio signal processing, speech source separation based on deep learning.

Wenwu Wang (M'02-SM'11) was born in Anhui, China. He received the B.Sc. degree in 1997, the M.E. degree in 2000, and the Ph.D. degree in 2002, all from Harbin Engineering University, China. He then worked in Kings College London, Cardiff University, Tao Group Ltd. (now Antix Labs Ltd.), Creative Technology Ltd., before joining University of Surrey, Guildford, U.K., where he is currently a Reader in Signal Processing, and a Co-Director of the Machine Audition Laboratory, in the Centre for Vision Speech and Signal Processing.

His current research interests include blind signal processing, sparse signal processing, audio-visual signal processing, machine learning and perception, machine audition (listening), and statistical anomaly detection. He has (co)-authored over 200 publications in these areas.

Jonathon Chambers (S'83-M'90-SM'98-F'11) received the Ph.D. and D.Sc. degrees in signal processing from the Imperial College of Science, Technology and Medicine (Imperial College London), London, U.K., in 1990 and 2014, respectively. On 1st Dec 2017 he became the Head of the Engineering Department at the University of Leicester. He is also an International Honorary Dean and Guest Professor within the Department of Automation at Harbin Engineering University, China. His research interests include adaptive signal processing and machine learning and their application in communications, defence and navigation systems.

Dr. Chambers is a Fellow of the Royal Academy of Engineering, U.K., the Institution of Engineering and Technology, and the Institute of Mathematics and its Applications. He has served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING for three terms over the periods 1997-1999, 2004-2007, and as a Senior Area Editor 2011-2015.

Syed Mohsen Naqvi (S'07-M'09-SM'14) received the Ph.D. degree in Signal Processing from Loughborough University, Loughborough, U.K., in 2009 and his Ph.D. thesis was on the EPSRC U.K. funded project. He was a Postdoctoral Research Associate on the EPSRC U.K.-funded projects and REF Lecturer from 2009 to 2015. Prior to his postgraduate studies in Cardiff and Loughborough Universities U.K., he served the National Engineering and Scientific Commission (NESCOM) of Pakistan from Jan 2002 to Sep 2005.

Dr Naqvi is a Lecturer in Signal and Information Processing at the School of Engineering, Newcastle University, Newcastle, U.K. He has 100+ publications with the main focus of his research being on Multimodal (audio-video) Signal and Information Processing. He is Fellow of the Higher Education Academy (FHEA). His research interests include multimodal processing for human behaviour analysis, multi-target tracking, and source separation all; for machine learning. He organized special sessions on multi-target tracking in FUSION 2013&2014, delivered seminars and was a speaker at UDRC Summer School 2015-2017.