Monaural Source Separation in Complex Domain with Long Short-Term Memory Neural Network

Yang Sun, *Student Member, IEEE,* Yang Xian, *Student Member, IEEE,*
Wenwu Wang, *Senior Member, IEEE,* and Syed Mohsen Naqvi, *Senior Member, IEEE*

***Abstract*—In recent research, deep neural network (DNN) has been used to solve the monaural source separation problem. According to the training objectives, DNN-based monaural speech separation is categorized into three aspects, namely masking, mapping and signal approximation (SA) based techniques. However, the performance of the traditional methods is not robust due to variations in real-world environments. Besides, in the vanilla DNN-based methods, the temporal information cannot be fully utilized. Therefore, in this paper, the long short-term memory (LSTM) neural network is applied to exploit the long-term speech contexts. Then, we propose the complex signal approximation (cSA) which is operated in the complex domain to utilize the phase information of the desired speech signal to improve the separation performance. The IEEE and the TIMIT corpora are used to generate mixtures with noise and speech interferences to evaluate the efficacy of the proposed method. The experimental results demonstrate the advantages of the proposed cSA-based LSTM RNN method in terms of different objective performance measures.**

***Index Terms*—Deep neural networks, Monaural speech separation, Long short-term memory, Complex signal approximation**

I. INTRODUCTION

Source separation has attracted a remarkable amount of attention due to its potential use in several real-world applications such as automatic speech recognition (ASR), assisted living systems and hearing aids [1]–[6]. In these applications, well separated signals are required for the system to work properly. According to the number of channels, the source separation problem is classified into multichannel, binaural-channel and single-channel (monaural) categories. The monaural source separation problem still remains an important research challenge, because only one recording is available and the spatial information that can be extracted is limited [7].

Many approaches have been developed to address the monaural source separation problem. For example, in signal processing-based methods, Loizou estimated the ideal Wiener filter and reconstructed the target signal in the minimum mean squared error (MMSE) sense [8]. While in model-based methods, the non-negative matrix factorization (NMF) [9] is exploited to separate signals from a single channel mixture [10]. Grais and Erdogan modelled the noisy observations based on weighted sums of non-negative sources [11]. However, these methods are limited when dealing with acoustic mixtures captured in real environments, for instance, in low signal-to-noise ratio (SNR) conditions, with unseen noises in the mixture and limited computational resources. Therefore, in real-environment scenarios, it is difficult to obtain the target speech signal with high quality consistently by using the above mentioned methods [12].

Recently, the DNN-based techniques have been introduced, where the trained neural network model is used to reconstruct the desired speech signals. According to the training objectives, DNN-based monaural speech separation is categorized into three aspects, namely masking, mapping and signal approximation (SA) based techniques.

In masking-based DNN approach, the ideal time-frequency (T-F) mask is applied as the training target of the neural network models. The T-F mask predicted by the trained model is applied to the mixture to reconstruct the desired speech signal. The predicted T-F mask can be categorized as a binary or soft mask. In the binary mask, each T-F unit of the mask was assigned as 1 or 0 according to the criterion for the active source [13], [14]. For example, Jin and Wang exploited an ideal binary mask (IBM) as training target, and obtained promising separation results [15]. However, due to the hard decisions from the IBM, the separated speech signal of the IBM-based method is distorted. In the soft mask, also known as ideal ratio mask (IRM), the T-F unit was assigned as the ratio of target source energy to mixture energy [12]. By using the IRM, Zhang and Wang proposed a deep ensemble method to further improve the performance of the IRM [12]. Compared with the IBM, the desired speech signal separated by IRM often has better quality, e.g. with less musical noise artefacts. Although these DNN-based techniques offer state-of-the-art performance, the masks including the IBM and the IRM do not utilize the phase information of the target signal when synthesizing the clean speech signal. Wang and Lim considered phase information to be unimportant in speech enhancement [16], but Erdogan et al. have shown that the phase information is beneficial to predict an accurate mask and the estimated source [17]. Consequently, in [18], Williamson et al. employed both the magnitude and phase spectra to estimate the complex IRM (cIRM) by operating in the complex domain.

In mapping-based DNN approach, the training target is the spectrum of the clean speech signal and the neural network model is trained to estimate the clean spectrum of desired speech signal. In [19], the DNN model was trained to learn the

Y. Sun, Y. Xian, 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; Y.Xian2@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-mail: W.Wang@surrey.ac.uk)

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

relationship between the spectrum of the mixture and the clean spectrum of the target signal to address the dereverberation and denoising problems. However, compared with the masking- and SA-based approaches, it is more difficult to obtain a well trained neural network model, due to the large value ranges in the spectrum of the clean speech signal at each T-F point.

In the SA-based DNN approach, the training target is the spectrum of the clean speech signal, which is indirectly obtained by the T-F mask estimated by a trained model which minimizes the discrepancy between the estimated spectrum and the spectrum of the clean speech signal. The original SA-based (oSA) DNN method does not utilize the phase information to reconstruct the target signal [20]. Moreover, with different environmental noises, the separation performance of the DNN-based methods, which are trained with either an IRM or clean signal spectra, are not robust.

In this architecture, the temporal information cannot be fully used, hence, the recurrent neural network (RNN) is introduced as the framework of the monaural source separation. Huang et al. have shown that the recurrent unit is beneficial to predict an accurate mask and improve separation performance [1]. By using the LSTM block instead of the regular network units, Chen and Wang utilized the LSTM neural network in the monaural source separation and the evaluations confirmed the improvement of the separation performance [21]. Sun et al. compared the mapping- and masking-based LSTM RNN methods in speech enhancement with different SNR levels and background noise [22]. However, these LSTM-based methods are applied with SA or IRM, where the clean phase information was not used.

To address the above mentioned issues, we propose an improved method where the LSTM neural network is used to estimate the cIRM, and then a cSA-based LSTM RNN method is presented to recover the desired speech signal from the cIRM.

In summary, the contributions of this paper are:

(1) A Y-shaped LSTM RNN is exploited to predict the cIRM as the training target, in order to utilize the phase information of the clean speech signal.

(2) The cSA-based LSTM RNN method is proposed, where both real and imaginary components of the spectrum are used as the training targets.

(3) Several complex domain separation methods with different neural network architectures are compared.

The rest of the paper is organized as follows. In Section II, the background knowledge related to the training targets in recent monaural source separation methods is described. Section III introduces the LSTM-based method and the proposed cSA-based source separation method. Section IV presents the experimental settings and results with the IEEE and the TIMIT corpora [23], [24]. The conclusions and future work are given in Section V.

II. Monaural source separation with Neural Networks

Recently, neural networks have been adopted as a regression model to solve the source separation problem, especially, in the monaural case. In this section, some background of the network architectures and training targets will be described.

A. Network Architectures

Generally, there are three fundamental and commonly used neural network architectures: DNN, RNN and convolutional neural network (CNN) [25]. All the above mentioned methods are based on the vanilla DNN, which is a feed-forward neural network model, and in this paper, all the DNNs are referred to the vanilla DNN. In monaural source separation, most of the approaches are based on DNNs or RNNs due to their relatively low complexities and effectiveness in solving the source separation problem. In addition, some advanced architectures have also been investigated, such as deep recurrent neural network (DRNN) [1], [26] and LSTM RNN [21], [22], [27]. Huang et al. applied the DRNN as neural network model to solve the monaural source separation problem where only specified hidden layers have connected units [1]. Compared with the DNN and RNN, the DRNN has a better trade-off between computational cost, storage space and the ability to employ temporal information. The LSTM RNN is able to store information in memory cells over a long period and the temporal information can be utilized more efficiently than the vanilla RNN [28]. By using the LSTM RNN, the speaker generalization ability of the source separation method can be improved, which is confirmed in [21]. Hence, the LSTM RNN is used as the framework of the proposed method.

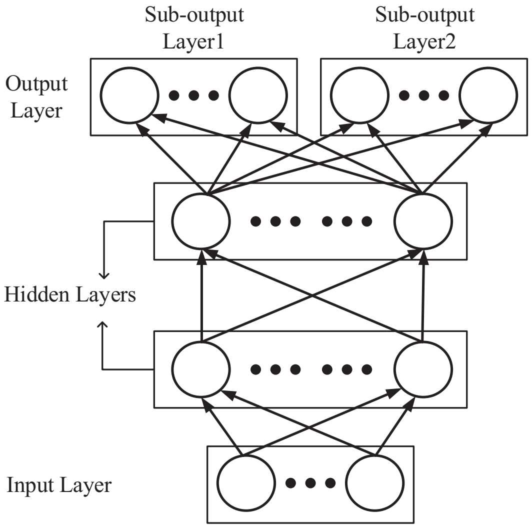

Fig. 1. The Y-shaped neural network architecture, which has two sub-output layers. The sub-output layer 1 and the sub-output layer 2 yield the real and imaginary components of the estimation, respectively.

If the training targets are given in the complex domain i.e. cIRM, the outputs of the DNN or the LSTM RNN are dual, with two sub-output layers, one for the real component and the other for the imaginary component of the estimation. Therefore, the shapes of DNN and the LSTM RNN will be changed with the types of training objectives. The architecture of the Y-shaped neural network is depicted in Figure 1, where the output predictions are jointly optimized [29].

B. Training Targets

Based on the training targets, the monaural source separation technique can be divided into three categories: masking-, mapping- and signal approximation (SA)-based, respectively. Both mapping- and SA-based approaches use the spectrum of the clean speech signal as a training target. In mapping-based approach, the value range of the spectrum at each T-F point is large, i.e. [0, +∞). In SA-based approach, however, the spectrum of clean speech signal is obtained by the predicted T-F mask, with a value range in [0, 1]. In comparison, the SA-based approach can lead to more accurate neural network model than the mapping-based method.

There are two differences between SA- and masking-based approaches. First, the training target of the masking-based approach is an ideal T-F mask, which is calculated by using the target signal and the speech mixture, while in the SA-based approach, the training target is the spectrum of the clean speech signal. Second, although the T-F mask is estimated in both SA- and masking-based approaches, in SA-based approach, the estimated T-F mask is exploited to minimize the discrepancy between estimated spectrum and the spectrum of the clean speech signal. The T-F mask is not directly used as the training target which is the main difference between the SA- and masking-based approaches. In this subsection, two state-of-the-art T-F masks are described. The IRM and the cIRM are the two training targets often chosen in masking-based approach.

1) Ideal Ratio Mask: Assume at discrete time m, the clean speech signal is $s(m)$, the interference is $i(m)$, and the mixture is $y(m) = s(m) + i(m)$. After applying the short time Fourier transform (STFT), the mixture is expressed as:

$$Y(t, f) = S(t, f) + I(t, f) \tag{1}$$

where f is the index of the frequency bins and t is the index of the time frames; $Y(t, f)$, $S(t, f)$ and $I(t, f)$ are the Fourier transforms of the mixture, clean signal and interference, respectively. Besides, employing the ideal T-F mask $M(t, f)$, the spectrum of the clean speech can be reconstructed as:

$$S(t, f) = Y(t, f)M(t, f) \tag{2}$$

The $M(t, f)$ as an IRM can be defined as:

$$M_{IRM}(t, f) = \left(\frac{|S(t, f)|^2}{|S(t, f)|^2 + |I(t, f)|^2}\right)^{\beta} \tag{3}$$

where β is a tunable parameter to scale the mask, $|S(t, f)|$ denotes the magnitude spectrum of the clean speech signal and $|I(t, f)|$ denotes the magnitude spectrum of the interference signal, respectively.

In IRM, only magnitude information is exploited, however, phase information is also important [30].

2) Complex Ideal Ratio Mask: Since the phase information of the spectrum is important, the cIRM was proposed [18], [31]. To calculate the cIRM, the STFTs of the mixture, clean signal and the cIRM are written as:

$$Y(t, f) = Y_r(t, f) + jY_c(t, f) \tag{4}$$

$$S(t, f) = S_r(t, f) + jS_c(t, f) \tag{5}$$

$$M_{cIRM}(t, f) = M_{cIRM_r}(t, f) + j{\cdot}M_{cIRM_c}(t, f) \tag{6}$$

where $j \triangleq \sqrt{-1}$ and the subscripts 'r' and 'c' indicate the real and the imaginary components in the STFTs, respectively. The $M_{cIRM}(t, f)$ is the T-F unit of the cIRM, which is defined as:

$$\begin{aligned} M_{cIRM}(t, f) = & \frac{Y_r(t, f)S_r(t, f) + Y_c(t, f)S_c(t, f)}{Y_r^2(t, f) + Y_c^2(t, f)} \\ & + j\frac{Y_r(t, f)S_c(t, f) - Y_c(t, f)S_r(t, f)}{Y_r^2(t, f) + Y_c^2(t, f)} \end{aligned} \tag{7}$$

The cost function of the cIRM-based DNN is expressed as:

$$\begin{aligned} J_{cIRM} = & \sum_t \sum_f \Big[\Big(\hat{M}_{cIRM_r}(t, f) - M_{cIRM_r}(t, f) \Big)^2 \\ & + \Big(\hat{M}_{cIRM_c}(t, f) - M_{cIRM_c}(t, f) \Big)^2 \Big] \end{aligned} \tag{8}$$

where the $\hat{M}_{cIRM}(t, f)$ is the T-F unit of the estimated cIRM.

In the cIRM-based approach, both the magnitude and phase responses are obtained to recover the target signal [18].

III. Algorithm Description

A. Complex Signal Approximation

In the mapping-based approach, the training target is the spectrum of the clean speech signal. The cost function of the mapping-based approach is written as:

$$J_{mapping} = \sum_t \sum_f (|\hat{S}(t, f)| - |S(t, f)|)^2 \tag{9}$$

where $\hat{S}(t, f)$ is the STFT of the estimated source. Hence, the clean spectrum of the target signal can be estimated by minimizing the error between the estimated spectrum and the spectrum of clean speech signal. While, due to the large value range of the T-F points in the spectrum, the network model is difficult to train [18].

The SA-based approach combines the mapping- and masking-based approaches. The training target in the oSA-based method is the spectral magnitude of clean speech, which is equivalent to the mapping-based approach. The cost function in the oSA-based method can be written as:

$$J_{oSA} = \sum_t \sum_f (|Y(t, f)\hat{M}_{oSA}(t, f)| - |S(t, f)|)^2 \tag{10}$$

where the predicted T-F mask in the oSA-based method is $\hat{M}_{oSA}(t, f)$, which is used to obtain the estimated spectrum $\hat{S}(t, f)$. The T-F mask is predicted in the oSA-based neural network to minimize the discrepancy between the magnitude spectrum of mixture and that of the clean speech signal, which is similar to masking-based approaches. Hence, using the magnitude spectrum of the clean signal as the training target can increase the accuracy of the estimated T-F mask and improve separation performance.

However, the oSA-based method has the same problem as the IRM-based method where the phase information of the target signal is not used when reconstructing the desired signal. Therefore, inspired by the cIRM, the cSA-based method is proposed, which replaces the IRM by cIRM in the training process to estimate both real and imaginary components of the clean speech signal. One could use the magnitude and phase information, instead of the real and imaginary components, as training targets, are exploited. However, our empirical tests show that using the real and imaginary components as trainning targets offers better separation performance. Hence, in the cSA-based method, the real and imaginary components of the desired clean speech signal are used as training targets.

In the cSA-based method, the estimated spectrum of the clean signal is obtained by applying the predicted complex T-F mask, defined as $\hat{M}_{cSA}$.

Similarly, the real component of the estimated clean spectrum in the cSA is expressed as:

$$\hat{S}_r(t,f) = \hat{M}_{cSA_r}(t,f)Y_r(t,f) - \hat{M}_{cSA_c}(t,f)Y_c(t,f) \tag{11}$$

The imaginary component of the estimated clean spectrum is calculated as:

$$\hat{S}_c(t,f) = \hat{M}_{cSA_r}(t,f)Y_c(t,f) + \hat{M}_{cSA_c}(t,f)Y_r(t,f) \tag{12}$$

In the proposed cSA-based LSTM RNN method, when the Y-shaped neural network model is used, the shared weights in the hidden layers cannot be fully used for both components, and this may have negative impacts on the estimations, and thus the separation performance. Our empirical tests show that using two networks performs better than stacking the two components in one network. In the cSA-based method, the real and imaginary components are estimated separately and two neural network models are trained with real and imaginary components of the cIRM. The cost functions can be expressed in the complex domain with the real and imaginary components. According to (11) and (12), the expanded cost functions of the cSA-based method are:

$$\begin{aligned} J_{1} = \sum_t \sum_f \Big[\Big(& \hat{M}_{cSA_r}(t,f)Y_r(t,f) \\ & - \hat{M}_{cSA_c}(t,f)Y_c(t,f) \Big) - S_r(t,f) \Big]^2 \end{aligned} \tag{13}$$

$$\begin{aligned} J_{2} = \sum_t \sum_f \Big[\Big(& \hat{M}_{cSA_r}(t,f)Y_c(t,f) + \\ & \hat{M}_{cSA_c}(t,f)Y_r(t,f) \Big) - S_c(t,f) \Big]^2 \end{aligned} \tag{14}$$

The architecture of the neural network model for the cSA-based method is shown in Figure 2, it has two output layers, the T-F mask is obtained in the additional output layer and the estimated component of the clean spectrum is obtained with the final output layer. If the training target is the imaginary component, the T-F mask is employed to estimate the imaginary component.

However, in the vanilla DNN, the temporal information cannot be fully used, which impacts on the separation performance. To address this limitation, the vanilla RNN and its improved version e.g. the LSTM RNN, which uses the LSTM block in the vanilla RNN, has been used for the challenging monaural source septation problem [27], [28]. In the cIRM- and the proposed cSA-based methods, the LSTM RNN is applied in this work for monaural source separation. The frameworks of the cIRM- and the cSA-based LSTM RNN methods are discussed in the following subsection.

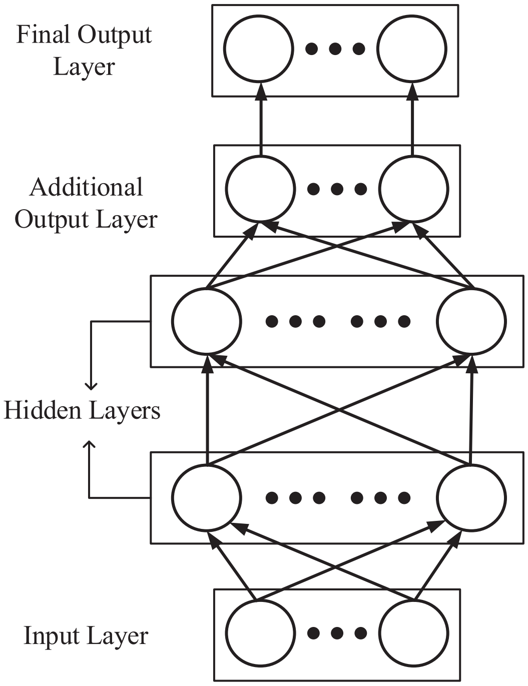

Fig. 2. Proposed neural network architecture, where a linear output layer is added before the final output layer to obtain the estimated speech signal. The output of the neural network model is related to the training target.

B. *LSTM RNN-based Methods in the Complex Domain*

Different from the vanilla DNN, which can only use context window to capture temporal dependencies, the LSTM RNN stores the temporal information in the cell, therefore, the long temporal dependencies can be utilized. In the DNN-based method, the neural network model is trained with backward propagation algorithm [18] but in the LSTM RNN-based method, the backward propagation through time algorithm is exploited [28]. The LSTM block in the proposed method is composed of a cell, an input gate, an output gate and a forget gate, similar to the structure in [21].

After the hidden states are obtained from the LSTM blocks, the output layer is added to generate the output of the LSTM RNN. The activation function of the output layer is selected as a linear function. For complex domain monaural source separation, the estimated phase information of clean speech signal is used to recover the desired speech signal. Then, by introducing the LSTM RNN, the temporal information is utilized. Besides, if the training target of the LSTM RNN is the cIRM, the neural network is Y-shape and two sub-output layers are added as shown in Figure 1. In the cSA-based LSTM RNN method, two LSTM RNNs are exploited to predict the real and imaginary components in parallel and both LSTM RNNs have the same configuration.

In the proposed cSA-based LSTM RNN method, inspired by [18], [25] and vanilla DNN methods, the feature combination is given to the input layer to increase the efficiency of the networks and system. The amplitude modulation spectrogram

(AMS) [33], relative spectral transform and perceptual linear prediction (RASTA-PLP) [34], mel-frequency cepstral coefficients (MFCC), cochleagram response and their deltas are extracted by a 64-channel gammatone filterbank to obtain the compound feature [35]. Furthermore, in the oSA- and the cSA-based methods, the spectra of the mixture and the clean signal are given to calculate the spectrograms of the predicted clean signal and the training objective, respectively.

The flow diagram of the proposed cSA-based LSTM RNN method is shown in Figure 3.

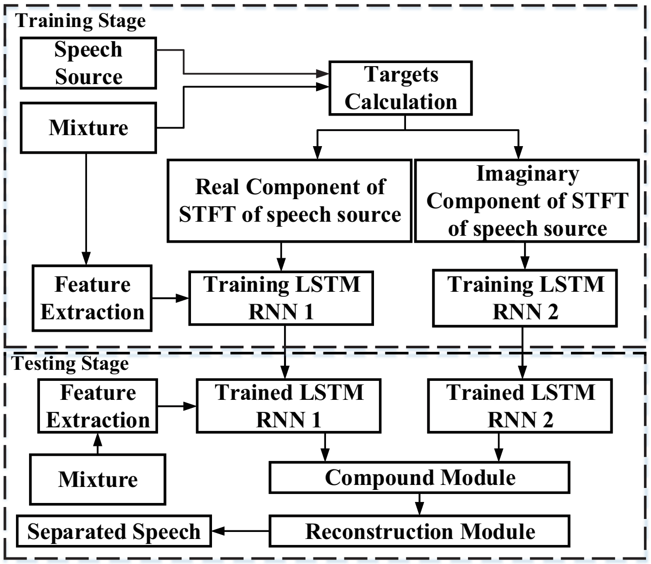

Fig. 3. The block diagram of the proposed complex signal approximation (cSA)-based LSTM RNN method. Two LSTM RNNs are trained with the separate training targets, e.g. the real and the imaginary components of the STFT of clean speech signal.

In the training stage, by using the targets calculation module, the STFTs of speech source and mixture are obtained. Then, the real and imaginary components of STFT of the speech source are used as the training targets for LSTM RNN 1 and LSTM RNN 2, respectively. The outputs of the LSTM RNN models are obtained by multiplying the estimated T-F mask with the STFT of the mixture. After each iteration, the estimated T-F mask is trained to minimize the discrepancy between the spectrum of the clean speech signal and that of the estimated source signal.

In the testing stage, the trained LSTM RNNs can output the real and imaginary components of the estimated speech signal when the feature combination of the mixture is used as input. Then, the STFT of the separated speech is obtained in the compound module and the separated speech signal is reconstructed in the reconstruction module.

Compared with the oSA-based DNN method, the proposed cSA-based LSTM RNN method has two advantages:

(1) In traditional oSA-based DNN method, the noisy phase information is used to synthesise the desired speech signal. However, in the proposed cSA-based LSTM RNN method, both clean magnitude and phase information are estimated.

(2) The LSTM blocks are introduced with the RNN, the temporal information can be better utilized and the trained LSTM RNN models have better generalization ability.

IV. Evaluations and results

In this section, we evaluate the cIRM- and oSA-based method with the vanilla DNN and the LSTM RNN to show the advantage of LSTM RNN over the vanilla DNN. Then, we show the results of the proposed cSA-based LSTM RNN method. Firstly, the interference is selected as the noise, in both seen and unseen scenarios. Then, the interference is chosen as the undesired speech signal which is unseen in the training stage. Therefore, the generalization ability of these methods can be evaluated.

A. Experimental Settings

1) Datasets: The speech sources are selected randomly from the IEEE and the TIMIT corpora [23], [24]. 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 speaker-independent. We randomly select 1000, 100 and 200 clean utterances from the IEEE and the TIMIT corpora to generate the training, development and testing datasets.

The interferences are categorized into two aspects, the noise interference and the undesired speech interference. In the seen noise interference cases, these clean speech utterances are mixed with five different noise types at three different SNR levels (-3 dB, 0 dB and 3 dB). These five noise scenes are named as *factory*, *babble*, *cafe*, *f16* and *tank*. The names of these noise signals indicate their recording situations. The above mentioned noise signals are selected from the NOISEX database [36]. Each noise sequence is four minutes long, which is truncated randomly from the first two minutes to match the lengths of the speech signals to generate the training mixtures. The last two minutes are used to generate the development and testing mixtures. In this case, although the noise interference in the testing dataset is unseen, the noise type is known.

In the unseen noise interference cases, 50 different noise signals are used to generate the training, development and testing datasets and 50 noise signals are only used to generate the testing data. These non speech sounds contain many different types of noise, e.g. animal sounds, tooth brushing sounds and machine noise [37]. Finally, the number of mixtures in training, development and testing data is 12,000, 1200 and 2400, respectively. The training speech duration is around 10 hours and 100 types of different noise signals are used in the unseen cases.

In our evaluation studies where the interference is undesired speech signal, 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

TABLE I

Separation performance comparison in terms of STOI with different training targets, noises and neural network architectures, the SNR of these mixtures is -3 dB. Each result is the average value of 200 experiments. *Italic* shows the proposed methods. **BOLD** indicates the best result.

STOI	Unprocessed	cIRM-DNN [18]	*cIRM-LSTM*	oSA-DNN [1]	*oSA-LSTM*	*cSA-LSTM*
Factory	60.35%	68.31%	70.59%	70.21%	72.42%	**73.57**%
Babble	57.04%	69.22%	70.00%	68.33%	74.12%	**76.70**%
Cafe	58.07%	65.45%	68.62%	66.11%	69.03%	**75.44**%
F16	62.54%	71.11%	72.58%	72.02%	74.17%	**75.20**%
Tank	70.93%	75.48%	79.04%	76.11%	85.35%	**86.77**%
Averaged	61.79%	69.91%	72.17%	70.56%	75.01%	**77.54**%

TABLE II

Separation performance comparison in terms of STOI with different training targets, noises and neural network architectures, the SNR of these mixtures is 0 dB. Each result is the average value of 200 experiments. *Italic* shows the proposed methods. **BOLD** indicates the best result.

STOI	Unprocessed	cIRM-DNN [18]	*cIRM-LSTM*	oSA-DNN [1]	*oSA-LSTM*	*cSA-LSTM*
Factory	67.42%	74.20%	77.92%	76.33%	78.92%	**79.59**%
Babble	64.22%	73.87%	76.81%	72.91%	78.99%	**79.47**%
Cafe	63.21%	70.36%	75.38%	71.38%	75.44%	**77.61**%
F16	65.31%	74.20%	77.26%	74.87%	79.77%	**80.13**%
Tank	75.34%	80.92%	83.75%	81.25%	87.51%	**88.03**%
Averaged	67.10%	74.74%	78.22%	75.35%	80.12%	**80.96**%

TABLE III

Separation performance comparison in terms of STOI with different training targets, noises and neural network architectures, the SNR of these mixtures is 3 dB. Each result is the average value of 200 experiments. *Italic* shows the proposed methods. **BOLD** indicates the best result.

STOI	Unprocessed	cIRM-DNN [18]	*cIRM-LSTM*	oSA-DNN [1]	*oSA-LSTM*	*cSA-LSTM*
Factory	70.36%	81.39%	83.94%	81.95%	84.89%	**85.99**%
Babble	71.22%	80.01%	82.99%	80.03%	85.28%	**86.03**%
Cafe	70.47%	79.20%	81.14%	79.30%	80.97%	**82.06**%
F16	72.45%	81.34%	82.62%	81.66%	84.02%	**84.71**%
Tank	79.66%	84.37%	87.77%	84.66%	89.20%	**89.26**%
Averaged	72.83%	81.26%	83.69%	81.52%	84.87%	**85.61**%

network is able to differentiate the target and undesirable speech signals. Similarly, the SNR levels are -3 dB, 0 dB and 3 dB and the number of mixtures in training, development and testing data is 12,000, 1200 and 2400, respectively.

2) Network Architecture: Both the DNNs of the comparison group and the LSTM RNN have three hidden layers and each hidden layer has 512 units. The dimension for the input layer is 1722 ($246\times(3\times2+1)$). In terms of the DNN, according to [18], 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 [31]. In the LSTM RNN, the activation function for each hidden unit is selected as the sigmoid and the output layer has linear units. When the training target is the cIRM, the corresponding neural network outputs the estimates of real and imaginary components of the predicted cIRM. When the training target is the clean spectrum of the desired speech signal, two LSTM RNNs are trained separately. The DNN and the LSTM RNN are trained by using the RMSprop algorithm [38] with a learning rate of 0.001. The number of epochs is 100 and the batch size is 1024. Auto-regressive moving average (ARMA) filtering is applied to reduce the interference from the background noise, as in [39].

3) Comparisons and Performance Measures: In the experiments, the proposed cIRM- and cSA-based LSTM RNN methods are compared with DNN-based approaches: the cIRM [18] and the oSA estimation [20]. In the oSA-based method, the T-F mask is an IRM, which is estimated by minimizing the discrepancy between the estimated spectrum and the spectrum of the target speech signal. In oSA-based DNN and LSTM RNN methods, the target signal is reconstructed without using the phase information of the clean speech signal, meanwhile, the cIRM- and the cSA-based methods utilize both the amplitude and phase information from the clean signal. The proposed methods are shown in *italics*. The separation performance is evaluated with three measurements. The short-time objective intelligibility (STOI) [40], the perceptual evaluation of speech quality (PESQ) [41] and the SDR [42]. The values of the STOI are in the range of [0, 1] and the PESQ are in the range of [-0.5, 4.5]. The STOI and the PESQ indicate the intelligibility scores and human speech quality scores, respectively. The SDR is exploited to evaluate the overall separation performance. In this paper, we use SDR value of the separated speech signal and the SDR value of the unprocessed speech mixture to calculate the improvement of the SDR.

TABLE IV
SEPARATION PERFORMANCE COMPARISON IN TERMS OF PESQ WITH DIFFERENT TRAINING TARGETS, NOISES AND NEURAL NETWORK ARCHITECTURES, THE SNR OF THESE MIXTURES IS -3 dB. EACH RESULT IS THE AVERAGE VALUE OF 200 EXPERIMENTS. *Italic* SHOWS THE PROPOSED METHODS. **BOLD** INDICATES THE BEST RESULT.

PESQ	Unprocessed	cIRM-DNN [18]	*cIRM-LSTM*	oSA-DNN [1]	*oSA-LSTM*	*cSA-LSTM*
Factory	1.63	2.07	2.33	2.11	2.30	**2.41**
Babble	1.76	2.05	2.12	2.03	2.22	**2.28**
Cafe	1.75	2.03	2.16	2.10	2.14	**2.38**
F16	1.64	2.13	2.25	2.10	2.27	**2.38**
Tank	1.92	2.29	2.49	2.33	2.72	**2.74**
Averaged	1.74	2.11	2.27	2.13	2.33	**2.44**

TABLE V
SEPARATION PERFORMANCE COMPARISON IN TERMS OF STOI WITH DIFFERENT TRAINING TARGETS, NOISES AND NEURAL NETWORK ARCHITECTURES, THE SNR OF THESE MIXTURES IS 0 dB. EACH RESULT IS THE AVERAGE VALUE OF 200 EXPERIMENTS. *Italic* SHOWS THE PROPOSED METHODS. **BOLD** INDICATES THE BEST RESULT.

PESQ	Unprocessed	cIRM-DNN [18]	*cIRM-LSTM*	oSA-DNN [1]	*oSA-LSTM*	*cSA-LSTM*
Factory	1.80	2.34	2.54	2.41	2.50	**2.59**
Babble	1.89	2.19	2.37	2.14	2.49	**2.51**
Cafe	1.95	2.27	2.38	2.29	2.32	**2.49**
F16	1.79	2.30	2.47	2.25	2.49	**2.61**
Tank	2.01	2.58	2.67	2.59	2.88	**2.91**
Averaged	1.88	2.34	2.49	2.37	2.54	**2.62**

TABLE VI
SEPARATION PERFORMANCE COMPARISON IN TERMS OF STOI WITH DIFFERENT TRAINING TARGETS, NOISES AND NEURAL NETWORK ARCHITECTURES, THE SNR OF THESE MIXTURES IS 3 dB. EACH RESULT IS THE AVERAGE VALUE OF 200 EXPERIMENTS. *Italic* SHOWS THE PROPOSED METHODS. **BOLD** INDICATES THE BEST RESULT.

PESQ	Unprocessed	cIRM-DNN [18]	*cIRM-LSTM*	oSA-DNN [1]	*oSA-LSTM*	*cSA-LSTM*
Factory	1.98	2.61	2.73	2.63	2.71	**2.81**
Babble	1.96	2.40	2.56	2.29	2.69	**2.76**
Cafe	2.01	2.46	2.58	2.48	2.55	**2.62**
F16	1.97	2.42	2.64	2.37	2.67	**2.77**
Tank	2.19	2.69	2.88	2.70	3.12	**3.17**
Averaged	2.02	2.51	2.67	2.49	2.75	**2.82**

B. Experimental Results and Analysis

1) Experimental Results with Seen Noise Interference in terms of the STOI and PESQ: The separation results based on the STOI are shown in Tables I, II and III. The results based on PESQ are shown in Tables IV, V and VI. Each experimental result in Tables I - VI is the average value over 200 testing mixtures. In total, 43,200 tests are performed. The baseline is calculated by using the unprocessed mixture and the clean speech signal.

It can be observed in Tables I - VI that the performance of LSTM RNN-based methods is better than the DNN-based methods. This is because the memory component in the LSTM RNN can better exploit the temporal information. In addition, the phase information is also beneficial and cSA-based LSTM RNN method outperforms all other methods. Besides, both values of the STOI and PESQ are increased when the SNR level changes from -3 dB to 3 dB.

2) Experimental Results with Noise Interference in terms of the SDR: These experiments aim to evaluate how the variations of the training targets, types of neural network models and SNR levels affect the SDR. The experimental settings are consistent with Section IV-A. The SDR values with different training targets and SNR levels are shown in Figure 4. It is shown in Figure 4 that the proposed cSA-based LSTM RNN method achieves the largest SDR improvement in all scenarios. When the vanilla DNN is trained, the cIRM- and oSA-based methods offer almost the same SDR improvement. While comparing the cIRM- and oSA-based methods with DNN and LSTM RNN, the performance of the LSTM RNN is again better than the DNN. By using the proposed LSTM RNN, the oSA-based method can gain 3.08, 3.11 and 2.58 dB more SDR improvements at -3, 0, and 3 dB SNR levels,

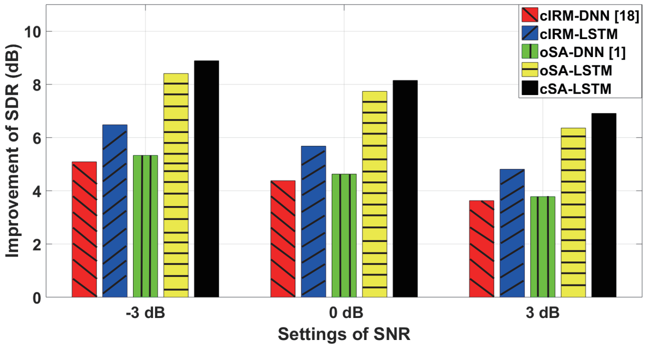

Fig. 4. Average SDR improvement (dB) for different training targets and neural network models with five types of seen noise. Each result is the average value of 200 experiments.

TABLE VII
SEPARATION PERFORMANCE COMPARISON IN TERMS OF STOI AND PESQ WITH DIFFERENT METHODS AND THE UNSEEN NOISES, THE SNR LEVELS OF THESE MIXTURES ARE -3, 0, AND 3 dB. EACH RESULT IS THE AVERAGE VALUE OF 200 EXPERIMENTS. *Italic* SHOWS THE PROPOSED METHODS. **BOLD** INDICATES THE BEST RESULT.

SNR level	STOI			PESQ		
	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB
Unprocessed	59.50%	66.16%	73.00%	1.61	1.80	2.01
cIRM-DNN [18]	64.33%	70.68%	76.92%	2.07	2.22	2.37
cIRM-LSTM	65.56%	72.78%	79.43%	2.17	2.34	2.53
oSA-DNN [1]	63.17%	69.06%	75.81%	2.09	2.25	2.36
oSA-LSTM	66.30%	75.99%	81.02%	2.24	2.35	2.47
cSA-LSTM	**75.14%**	**78.87%**	**83.52%**	**2.29**	**2.47**	**2.60**

respectively. In addition, the phase information of clean speech signal in complex domain provides further SDR improvement, e.g. by comparing with the oSA- and the cSA-based LSTM RNN methods.

3) Experimental Results with Unseen Noise Interference in terms of the STOI and PESQ: In the real-world environments where the situations varies, it is important to provide the generalization ability of the proposed methods. Therefore, the evaluation results based on the STOI and PESQ are shown in Table VII for unseen noise cases.

It can be known from Table VII that when the noise interference is unseen, the separation performance is decreased, compared with the seen noise interference case. It is difficult to obtain the accurate estimate in the testing stage with unseen noise interference. For example, when the noise interference is seen, in 0 dB SNR level, the cIRM-based DNN method can gain 7.64% improvement in terms of the STOI. However, if the noise interference is unseen, the improvement decreases to 4.83%.

Besides, in the unseen noise interference case, when the SNR level is increased, the separation performance is improved and the best separation performance is given by the proposed cSA-based LSTM RNN method. For instance, in -3 dB SNR level case, the cSA-based LSTM RNN method achieves 75.14% and 2.29 in STOI and PESQ, respectively. While the oSA-based DNN method only achieves 63.17% and 2.09, respectively.

Hence, if LSTM RNN is selected as the neural network model, the generalization of the related methods is enhanced, which has been confirmed by our experimental results similar to [21].

4) Experimental Results with Unseen Noise Interference in terms of the SDR: These experiments aim to evaluate how the variations of the SNR levels affect the SDR performance in terms of the proposed methods with unseen noise interference. Besides, the generalization ability is further evaluated. Figure 5 gives the SDR improvement with different training targets and neural network models.

It can be seen from Figure 5 that in the unseen noise case, compared with the cIRM-based DNN method, the cIRM-based LSTM RNN method gives more SDR improvement from -3 dB to 3 dB SNR levels. Similarly, the oSA-based LSTM RNN method achieves a higher SDR improvement than the oSA-based method by using the vanilla DNN. It is clear to observe that when the SA approach is operated in the complex domain and the LSTM RNNs are trained to predict the corresponding training targets, the separation performance outperforms others. For example, in the scenario, when the SNR level is -3 dB, the separation performance of oSA-based DNN method is 6.68 dB and the cSA-based LSTM RNN method gives 7.77 dB SDR improvement.

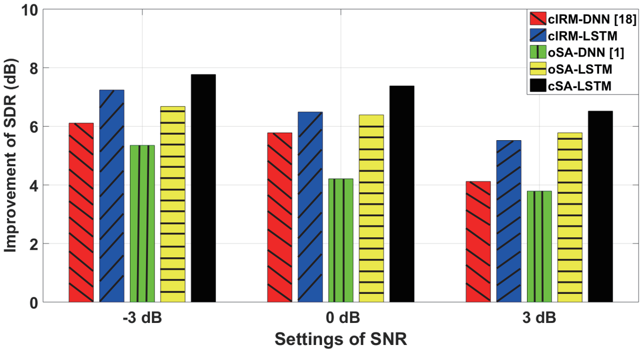

Fig. 5. Average SDR improvement (dB) for different training targets and neural network models with 100 types of unseen noise. Each result is the average value of 200 experiments.

From Tables I - VII and Figures 4 & 5, the best separation performance in noise interference case is given by the proposed cSA-based LSTM RNN method. There are two main reasons: (1) The phase information of clean speech signal is used to recover the desired speech signal; (2) the LSTM RNN exploits the temporal information and the generalization ability is enhanced. Besides, it can be seen from Table VII that by using the proposed cSA-based LSTM method, the best performance in terms of the STOI and PESQ is obtained in all SNR levels, although there are some discrepancies in the level of improvements across these performance metrics. One possible reason is that when the SNR level is low, by using the proposed cSA-based LSTM method, the intelligibility of the separated speech, as assessed by the STOI, is better improved, due to the time-frequency weighting of the speech spectrum. In a high SNR level, less processing is enforced on the separated speech signal. As a result, the level of artefacts introduced by the proposed cSA-based LSTM method is lower, as shown by the PESQ measure.

In summary, in the seen noise interference case, the separation performance is better than the unseen case. When the SNR level is changed from -3 dB to 3 dB, all of the methods achieve better separation performance. Moreover, compared

TABLE VIII
SEPARATION PERFORMANCE COMPARISON IN TERMS OF STOI AND PESQ WITH DIFFERENT METHODS AND THE SPEECH INTERFERENCE, THE SNR LEVELS OF THESE MIXTURES ARE -3, 0, AND 3 dB. EACH RESULT IS THE AVERAGE VALUE OF 200 EXPERIMENTS. *Italic* SHOWS THE PROPOSED METHODS. **BOLD** INDICATES THE BEST RESULT.

	STOI			PESQ		
SNR level	-3 dB	0 dB	3 dB	-3 dB	0 dB	3 dB
Unprocessed	64.84%	69.03%	76.62%	1.63	1.92	2.01
cIRM-DNN [18]	69.27%	73.82%	80.16%	2.02	2.23	2.37
cIRM-LSTM	69.13%	73.11%	80.33%	2.05	2.19	2.39
oSA-DNN [1]	70.84%	74.37%	81.79%	2.02	2.30	2.38
oSA-LSTM	72.84%	76.54%	82.25%	2.14	2.36	2.48
cSA-LSTM	**75.80%**	**79.26%**	**82.59%**	**2.32**	**2.54**	**2.57**

with the vanilla DNN, using the LSTM RNN as the neural network model, the proposed method provides improvement in all performance measures.

5) Experimental Results with Speech Interference in terms of the STOI and PESQ: When the interference is the undesired speech signal, the task is more difficult to address because the speech signals are highly non-stationary. In this subsection, the evaluations with undesired speech interferences are shown in Table VIII and Figure 6.

From Table VIII, it can be observed that when the interference is the undesired speech signal, compared with the noise interference cases, the separation performance decreases in all cases. The proposed cSA-based LSTM RNN method provides the highest values of both STOI and PESQ. Compared with the noise interference, when the interference is speech signal, because the indeterminacy of the speech interference, the related neural network model is more difficult to train, which effects on the overall separation performance.

After introducing the LSTM RNN, the separation performance is improved. For example, when the speech interference is used, in 0 dB SNR level, the oSA-based DNN method can gain 5.34% improvement in terms of the STOI, the oSA-based LSTM RNN method gives 7.51% improvement. In general, the phase information is beneficial and it can be observed that in -3 dB SNR level, the PSEQ value of oSA-based LSTM RNN method is 2.14 and cSA-based LSTM RNN method achieves 2.32.

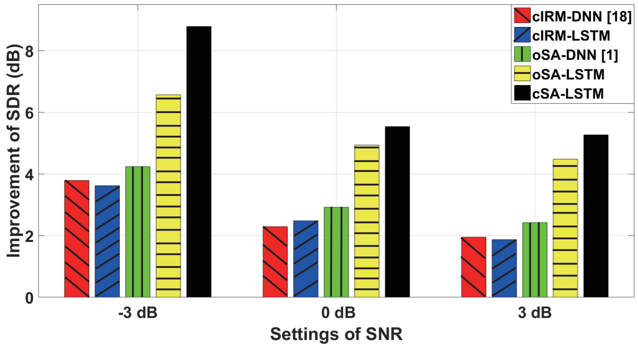

Fig. 6. Average SDR improvement (dB) for different training targets and neural network models with speech interferences. Each result is the average value of 200 experiments.

6) Experimental Results with Speech Interference in terms of the SDR: The variations of the SNR levels affect the SDR performance in terms of the proposed methods with speech interference is shown in Figure 6. It can be seen from Figure 6 that in the speech interference case, the cSA-based LSTM RNN method gives the largest SDR improvement over the other methods and SNR levels. It is shown that because the strong ability of using temporal information, the SDR improvement of the LSTM RNN-based method is always larger than the DNN-based methods. For instance, when the SNR level is -3 dB, the SDR improvement of the oSA-based DNN method is 4.11 dB and the improvement of the oSA-based LSTM RNN method is 6.24 dB.

However, in cIRM-based methods, due to the indeterminacy of the undesired speech signal, and the corresponding neural network is Y-shape, the T-F mask in the complex domain cannot be accurately estimated sometimes. For example, in Figure 6, when the SNR level is -3 dB, the cIRM-based DNN achieves higher SDR improvement than the cIRM-based LSTM RNN method. To address this issue, in the proposed cSA-based LSTM RNN method, two individual LSTM RNNs are used to estimate the real and imaginary components separately. It can be observed from Figure 6, when the SNR level is -3 dB, the performance of the proposed cSA-based LSTM RNN method is 8.91 dB, which confirms the efficacy of the proposed method.

In summary, in the speech interference case, the separation performance is less than the noise interference case. When the SNR level varies from -3 dB to 3 dB, all of these methods achieve better separation performance in both noise interference and speech interference cases. From Tables I to VIII and Figures 4 to 6, it is confirmed that the LSTM RNN is a better neural network model to utilize the long-term temporal information, which helps the trained model to obtain better separation performance.

It should be noted that although the phase information is helpful to improve the separation performance, which can be observed by comparing the results of the oSA-based method with those of the cSA-based method, the major improvement actually comes from the use of the SA-base method, which can be observed by comparing the performance of the oSA-based method with that of the cIRM-based method. The proposed cIRM-based LSTM RNN method not only has the benefits from the SA formulation but also the clean phase information.

V. Conclusions

In this paper, the cSA-based method with LSTM RNN was proposed to address the monaural source separation problem. By introducing cIRM, both real and imaginary components can be calculated and estimated in the cSA-based LSTM RNN method. Compared with oSA-based method, if the complex domain training targets were exploited, the phase information can be used in the SA-based approach. Hence, in the cSA-based method, both clean magnitude and phase information were utilized and the separation performance was further improved. The proposed method was evaluated using STOI, PESQ and SDR with two interfering cases. The unseen noise interference and undesired speech signal interference cases were evaluated to show the generalization ability of the proposed cSA-based LSTM RNN method. All the experimental results confirmed that the proposed method outperformed the oSA- and the cIRM-based approaches in all tested scenarios.

Acknowledgement

The authors would like to thank the Associate Editor and the anonymous reviewers for their valuable input to improving the 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] B. Wu, K. Li, M. Yang, and C.-H. Lee, “A reverberation time aware approach to speech dereverberation based on deep neural networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 102–111, 2017.

[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] Y. Sun, W. Rafique, J. A. Chambers, and S. M. Naqvi, “Underdetermined source separation using time-frequency masks and an adaptive combined Gaussian-Students t probabilistic model,” in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[8] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2007.

[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] J. Traa, P. Smaragdis, N. D. Stein, and D. Wingate, “Directional nmf for joint source localization and separation,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2015.

[11] 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.

[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] D. L. Wang, “On ideal binary mask as the computational goal of auditory scene analysis,” *Speech Sep. Humans Mach*, vol. 60, pp. 63–64, 2005.

[14] Y. Wang and D. L. Wang, “Towards scaling up classification-based speech separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

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

[22] L. Sun, J. Du, L. R. Dai, and C.-H. Lee, “Multiple-target deep learning for lstm-rnn based speech enhancement,” in *Hands-free Speech Communications and Microphone Arrays (HSCMA)*, 2017.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] F. Weninger, J.-L. Durrieu, F. Eyben, G. Richard, and B. Schuller, “Combining monaural source separation with long short-term memory for increased robustness in vocalist gender recognition,” in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

[28] E. Ceolini and S.-C. Liu, “Impact of low-precision deep regression networks on single-channel source separation,” in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[29] R. Caruana, “Multitask learning,” *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997.

[30] 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.

[31] 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.

[32] F. A. Gers, J. Schmidhuber, and F. Cummins, “Learning to forget: Continual prediction with lstm,” *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] G. Hu, "100 nonspeech sounds," http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html, 2014, online: accessed June 2018.

[38] S. Ruder, "An overview of gradient descent optimization algorithms," *in preprint arXiv: 1609.04747*, pp. 1–14.

[39] 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.

[40] 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.

[41] 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.

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

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.

Yang Xian (S'18) 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 2016. 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 enhancement and audio-video source separation.

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., in 2007, 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), artificial intelligence, and statistical anomaly detection. He has (co)-authored over 250 publications in these areas. He has been a Senior Area Editor (2019-) and an Associate Editor (2014-2018) for IEEE Transactions on Signal Processing. He is a Publication Co-Chair of ICASSP 2019, in Brighton, UK.

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 2002 to 2005.

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