PAPER

A Two-Stage Phase-Aware Approach for Monaural Multi-Talker Speech Separation

Lu YIN[†,††], Junfeng LI[†,††a)], Yonghong YAN[†,††,†††], *Nonmembers*, *and* Masato AKAGI[††††], *Member*

SUMMARY The simultaneous utterances impact the ability of both the hearing-impaired persons and automatic speech recognition systems. Recently, deep neural networks have dramatically improved the speech separation performance. However, most previous works only estimate the speech magnitude and use the mixture phase for speech reconstruction. The use of the mixture phase has become a critical limitation for separation performance. This study proposes a two-stage phase-aware approach for multi-talker speech separation, which integrally recovers the magnitude as well as the phase. For the phase recovery, Multiple Input Spectrogram Inversion (MISI) algorithm is utilized due to its effectiveness and simplicity. The study implements the MISI algorithm based on the mask and gives that the ideal amplitude mask (IAM) is the optimal mask for the mask-based MISI phase recovery, which brings less phase distortion. To compensate for the error of phase recovery and minimize the signal distortion, an advanced mask is proposed for the magnitude estimation. The IAM and the proposed mask are estimated at different stages to recover the phase and the magnitude, respectively. Two frameworks of neural network are evaluated for the magnitude estimation on the second stage, demonstrating the effectiveness and flexibility of the proposed approach. The experimental results demonstrate that the proposed approach significantly minimizes the distortions of the separated speech.

key words: *speech separation, phase recovery, amplitude estimation, deep learning, mask estimation*

1. Introduction

While humans can communicate and converse with others under complex acoustic environments, the noise, reverberation and concurrent speech impact the ability of hearing-impaired persons and automatic speech recognition systems dramatically. To solve this problem, researchers have dedicated to speech separation task for several decades. The multi-talker speech separation refers to extracting every source speech from a mixture utterance. Generally, to address the speech separation task, the time-domain mixture signal is transformed into temporal-frequency domain using short-time Fourier transform (STFT), and then the spectral magnitude of target speech is estimated and combined with the mixture phase to reconstruct the source signal.

Various techniques have been proposed to estimate the spectral magnitude, such as Non-negative Matrix Factorization (NMF) [1]–[4], Computational Auditory Scene Analysis (CASA) [5]–[7], and Hidden Markov Model (HMM) [8], [9]. In recent years, neural network-based speech separation has attracted increasing attention. In [10], [11], the authors proposed to train a deep neural network to estimate time-frequency masks, which were multiplied to the mixture spectrum to recover the target spectrum. However, only one target source is considered in these works. In [12]–[14], the neural networks are trained to separate two speakers with different gender and specific relative energy ratios. These constraints are helpful for the neural network to trace the target speakers, but the label permutation problem is not successfully solved. To deal with the label ambiguous problem, the deep clustering (DPCL) method [15]–[17] and the Permutation Invariant Training (PIT) method [18], [19] are proposed. The DPCL based approach trains a deep recurrent neural network to map the mixture into an embedding space, where k-means clustering is used to assign the time-frequency bins to different speakers. The PIT method computes two losses by exchanging the target labels and uses just the lower one in the back-propagation process. However, in these approaches, only the magnitude is estimated and the mixture phase is used for signal reconstruction.

The reasons of only estimating magnitude while keeping the mixture phase unchanged are as follows. On the one hand, the early studies considered that spectral phase is less important than spectral magnitude [20]. On the other hand, the minimum mean square error estimation of the speech phase equals to the mixture phase [21], with a uniform prior distribution assumption on the phase. Similarly, [22], [23] proposed that the mixture phase is the optimal maximum a posterior (MAP) estimator of speech phase. Another reason is that the spectral phase of speech is randomly distributed and unstructured, which makes it difficult to recover the phase.

With more studies on speech, researchers realized that phase also plays an important role in speech perception and speech signal processing [24]–[26]. In [24] the result showed that spectral phase provided important information of speech. In [26], the oracle and non-oracle scenarios were considered for objective and subjective speech quality experiments. Their results showed that speech quality was sig-

Manuscript received September 26, 2019.
Manuscript revised March 3, 2020.
Manuscript publicized April 20, 2020.

[†]The author is with the University of Chinese Academy of Sciences, Beijing, China.

[††]The author is with the Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, Beijing, China.

[†††]The author is with the Xinjiang Laboratory of Minority Speech and Language Information Processing, Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Xinjiang, China.

[††††]The author is with the Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, Nomi-shi, 923–1292 Japan.

a) E-mail: lijunfeng@hccl.ioa.ac.cn (Corresponding author)

DOI: 10.1587/transinf.2019EDP7259

nificantly improved with clean phase spectrum. This confirmed the importance of phase recovery for speech signal processing.

Several methods have been proposed for phase recovery. The first one is the consistency-based phase recovery [27], [28]. Griffin and Lim proposed an iterative approach for phase recovery, which can generate a consistent STFT spectrum, by updating the spectral phase and fixing the spectral magnitude during the iteratively iSTFT and STFT procedure [27]. The authors in [28] proposed the Multiple Input Spectrogram Inversion (MISI) algorithm, which improved the Griffin-Lim method for multiple signals separation. In [29], the authors used a differentiable consistency constraint layer within a DNN to enforce the consistency of the recovered speech. Another phase recovery method is based on the sinusoidal model which models speech as a weighted superposition of several sinusoidals [30]–[32]. This sinusoidal model approach can only recover the voiced sounds, it cannot provide valid phase estimation for the unvoiced sounds. In [33]–[35], the authors proposed a geometry-derived analytic solution for phase recovery. The group delay and instantaneous constraints were used to solve the phase ambiguity of two possible combinations. In [36], the authors proposed an auditory scene analysis-based method, which used the four heuristic regularities proposed by Bregman as constraints, to extract the instantaneous amplitude and phase of the desired signal. Recent works [37]–[39] attempted to use deep learning methods for phase recovery. In [37], the instantaneous frequency deviation (IFD) was used as the training target to recover the phase. At the testing stage, the phase was recovered from the estimated IFD with a post-processing procedure. In [38], the phase loss was combined with the group delay loss to train a von-Mises-distribution DNN for phase recovery. In [39], the authors proposed to treat phase estimation as a classification problem by discretizing and re-encoding the phase values. However, these deep-learning-based methods were proposed and evaluated for the target speaker extraction tasks such as speech enhancement [37], [39] and music source separation [39].

Considering the importance of the phase and the limitation of using the mixture phase in source separation, this study proposes to separate the mixed speech by both estimating the magnitude and recovering the phase. Our study makes following major contributions.

- First, a two-stage approach is proposed for phase-aware multi-talker speech separation, of which the first stage is used to recover the phase and the second stage is used to estimate the magnitude.
- Second, few previous works evaluated the performance of the mask for phase recovery. This study investigates the effectiveness of the mask used for phase recovery and proposes that the IAM is the optimal mask for the mask-based MISI algorithm, which brings less phase distortion.
- Third, this study proposes an advanced mask for the magnitude estimation with the recovered phase. The mask is used to estimate a magnitude that compensates for the error of phase recovery and brings less signal distortion.
- Last, two different frameworks for the proposed approach are evaluated in this work. Both the two frameworks yield significant separation performance and bring more flexibility for applications.

1.1 Related Work

Recent works [37], [39] attempted to train deep neural networks to recovery the phase for speaker extraction task, of which only one target speaker is mixed in a mixture. In contrast, the multi-talker separation task, which aims to reconstruct all utterances from an utterance containing multi-speaker, is more complicated. The interference of other speakers makes it more difficult to recover the phase. Therefore the method proposed in [37], [39] may not work well for multi-talker speech separation.

In this study, the phase is recovered via the MISI algorithm due to its effectiveness and simplicity. We propose to implement the MISI algorithm based on the Ideal amplitude mask (IAM) estimated by deep neural networks. Similar work to recover the phase can be found in [40]–[42]. The difference is that, in these works, the phase-sensitive mask (PSM) was used for the MISI algorithm according to the conclusion in [18], [43]. In [18] and [43], the authors concluded that the PSM outperforms other masks for speech separation. However, the conclusion is only valid for the magnitude estimation on the scenario with the mixture phase. For the phase recovery or the phase-recovered magnitude estimation, the conclusion is not valid. In contrast, this study investigates the masks listed in Table 1 for the mask based-MISI algorithm and demonstrates that the IAM brings less phase distortion than the PSM. Additionally, this study proposes an advanced mask for the magnitude estimation after phase recovery, which minimizes the distortion of the reconstructed signal.

Another two related works published just recently are worth mentioning. The first paper [44] proposes a discretization algorithm referred to as the Phasebook and Friends, which attempts to discretize the magnitude and the phase and estimate them by an end-to-end learning framework. The second paper [45] proposes an updated version of the TasNet algorithm, which separates the mixture on the time domain and achieves a performance surpassing the state-of-the-art works on the WSJ0-2mix dataset. Differ-

Table 1 Mask definition. $|Y|, |S|$ and $|X|$ denote the spectral magnitudes of the mixture, speech and noise, respectively.

Mask	Formula
Ideal Ratio Mask [46] :	$(\|S\|^2/(\|S\|^2+\|X\|^2))^{0.5}$
Ideal Amplitude Mask:	$\|S_j\|/\|Y\|$
Phase-sensitive Mask [43]:	$\|S_j\|/\|Y\| \cdot cos(\theta_{\rm Y}-\theta_j)$

ent to these works, this study recovers the phase with the mask-based MISI algorithm. While these methods are different to our approach, several techniques could be adopted in our framework as future work, such as multi-task training with deep-clustering head in the loss function and end-to-end training with unfolded MISI on waveform approximation [44], and convolution layers instead of bi-directional long short-term memory (BLSTM) [45].

2. Signal Model

The observed mixture can be modeled as a summation of each utterance

$$y(n) = \sum_{j=1}^{J} s_j(n) \tag{1}$$

where $y(n)$ and $s_j(n)$ denote the observed mixture and the individual utterance in the time domain, respectively. J is the number of speakers and j is the index of each speaker. The multi-talker speech separation task aims to extract every source speech $s_j(n)$ from the given mixture $y(n)$.

Generally, speech separation is processed in the frequency domain. Let $w(n)$ denote the analysis window used for STFT. Then, the analyzed signal $s(n)$ can be transformed into frequency domain as

$$S(k,l) = \sum_{n=0}^{R-1} s(n)w(n - lH)e^{-i2\pi kn/T} \tag{2}$$

where $S(k,l)$ is the frequency spectrum of $s(n)$. k and l denote the frequency index and the frame index, respectively. R is the length of $s(n)$. H denotes the shift size of analysis window. T indicates the length of the discrete Fourier transform (DFT).

With Eq. (2), Eq. (1) can be transformed to the frequency domain as

$$Y(k,l) = \sum_{j=1}^{J} S_j(k,l) \tag{3}$$

where $Y(k,l)$ and $S_j(k,l)$ denote the STFT spectra of $y(n)$ and $s_j(n)$, respectively. Both $Y(k,l)$ and $S_j(k,l)$ are complex numbers which can be represented in the polar coordinate system as the combination of the magnitude and the phase.

$$|Y(k,l)|e^{i\theta_Y(k,l)} = \sum_{j=1}^{J} |S_j(k,l)|e^{i\theta_j(k,l)} \tag{4}$$

where $|\cdot|$ is the magnitude function and θ_j denotes the phase of the jth source. θ_Y is the phase of the mixture. For clarity, the frequency index k and the time index l are omitted in the following sections.

3. Proposed Method

Most previous works only estimate the magnitude and use the mixture phase for speech separation. By this approach, the mask is first estimated and then multiplied with the mixture magnitude to recover the source $\hat{S}_j = \hat{\mathcal{M}}_\mathrm{j} \cdot |Y| \cdot e^{i\theta_Y}$. Here $\hat{\mathcal{M}}_\mathrm{j}$ denotes the estimated mask of the jth speaker. However, only separating the magnitude cannot exactly recover the source, and the mixture phase destroys the estimated magnitude due to the consistency constraint. By the approach proposed in this study, both the magnitude and the phase are estimated and the source is reconstructed as $\widetilde{S}_j = \widetilde{\mathcal{M}}_\mathrm{j} \cdot |Y| \cdot e^{i\hat{\theta}_j}$. Here $\hat{\theta}_j$ denotes the recovered phase. The phase is recovered with the IAM and the magnitude is estimated with the proposed mask, on the phase recovery stage and the magnitude estimation stage, respectively. The motivation of the two-stage approach is that the optimal mask used for phase recovery and the magnitude estimation are different, it is a reasonable way to estimate them respectively. Moreover, the source speech is not sufficiently separated by a single deep neural network and still contains some components of the interference speakers. A second stage deep neural network will further improve the separation performance [19]. Hence, the magnitude estimation stage is cascaded, rather than paralleled, with the phase recovery stage.

Following sections describe the motivation and the method of the phase recovery and the magnitude estimation, and then describe the approach of the mask estimation.

3.1 Phase Recovery Stage

As shown in Eq. (2), the analysis window is typically overlapped which introduces dependency between STFT frames. The length of analysis window introduces dependency between the STFT frequency channels. These dependencies between frequency bins impose certain constraint on the STFT spectrum. This means that the values of the time-frequency bins are not independently distributed but subject to a specific inner-relationship. By iteratively performing inverse STFT and STFT [28], the MISI algorithm utilizes the inner-relationship to recover the phase, which produces a consistent STFT spectrum. However, the MISI algorithm assumes that the ideal speech magnitude is given. While for the speech separation, the ideal magnitude of speech cannot be exactly known and few previous works evaluate the performance of the MISI algorithm with the estimated magnitude. Therefore, this study proposes to recover the phase based on the estimated mask. Moreover, this study investigates various masks for the mask-based MISI algorithm and demonstrates that the IAM is the optimal mask which brings less phase distortion than the IRM and PSM.

In this paper, the IAM is proposed to recover the phase. First, the IAM is estimated by a neural network. Then the IAM is used in the iSTFT and STFT iterative procedure. During the iteration, the phase is updated while the mask is fixed. After each iteration, the error between the summation of the reconstructed signals and the observed mixture is re-distributed to each source. Let δ denote the error of the summation of the reconstructed signals in the time domain,

$\delta = y - \sum_{j=1}^{J} \hat{s}_j^m$. Here y is the observed mixture and $\hat{s}_j^m$ is the reconstructed signal of the jth source. The mask-based MISI algorithm can be presented as the iterative procedure of Eq. (5) to Eq. (7):

$$\hat{s}_j^m = iSTFT(\hat{M}_j \cdot |Y| \cdot e^{i\hat{\theta}_j^m}) \tag{5}$$

$$\hat{S}_j^{m+1} = STFT(\hat{s}_j^m + \delta/J) \tag{6}$$

$$\hat{\theta}_j^{m+1} = \angle \hat{S}_j^{m+1} \tag{7}$$

where $\hat{M}_j$ denotes the estimated IAM of the jth source and $\hat{\theta}_j^{m+1}$ denotes the recovered phase after $(m+1)$ iterations. The mixture phase is used as the initial value of $\hat{\theta}_j^{(0)}$.

3.2 Magnitude Estimation Stage

For the phase-recovered speech separation, the target magnitude is obtained by multiplying the estimated mask to the mixture magnitude. An advanced mask is proposed in this section to estimate the magnitude, which compensates for the error of phase recovery and improves the separation performance.

Various masks have been used for magnitude estimation. In [43], the authors proposed to use the real part of the complex ideal ratio mask [47], [48] to estimate the magnitude. The study concluded that the PSM outperformed other masks for speech separation, which was widely used by later works. However, this conclusion was only valid for the magnitude estimation scenario without phase recovery. Moreover, The authors did not give the mathematical explanation that why the PSM outperformed other masks. In this paper, we propose an advanced mask, which takes the error of phase recovery into consideration by minimizing the mean square error (MSE) of the separated signal to the target speech. On the phase recovered separation scenario, the reconstructed signal is denoted as $\widetilde{S}_j = \widetilde{\mathcal{M}}_{\mathrm{j}} \cdot |Y| \cdot e^{i\hat{\theta}_j}$ in the complex domain. Hence, the complex distance between the reconstructed signal $\widetilde{S}_j$ and the target speech S_j is presented as:

$$\begin{aligned}\mathcal{D} = (\,\widetilde{\mathcal{M}}_{\mathrm{j}} \cdot |Y| cos\hat{\theta}_j - |S_j| cos\theta_j)^2 \\ +(\,\widetilde{\mathcal{M}}_{\mathrm{j}} \cdot |Y| sin\hat{\theta}_j - |S_j| sin\theta_j)^2\end{aligned} \tag{8}$$

With the product-to-sum formulas, Eq. (8) can be transformed as:

$$\begin{aligned}\mathcal{D} = (\,\widetilde{\mathcal{M}}_{\mathrm{j}} \cdot |Y| - |S_j| cos(\hat{\theta}_j - \theta_j))^2 \\ + |S_j|^2 (1 - cos^2(\hat{\theta}_j - \theta_j))\end{aligned} \tag{9}$$

After phase recovery, the value of the parameter $\hat{\theta}_j$ is determined. To minimize $\mathcal{D}$, the optimal solution is:

$$\widetilde{\mathcal{M}}_{\mathrm{j}} = \frac{|S_j|}{|Y|} cos(\hat{\theta}_j - \theta_j) \tag{10}$$

The optimal solution is denoted as phase-recovered mask (PRM). The PRM-based magnitude estimation compensates for the error of phase and optimally minimizes the error of the estimated speech to the target speech. It can be easily seen that, the PSM is a generalized mask which connects the conventional IAM and PSM. For $\hat{\theta}_j = \theta_j$, the PRM equals to the IAM. For $\hat{\theta}_j = \theta_Y$, the PRM equals to the PSM. This shows that for the separation with oracle phase, the IAM is the optimal mask to estimate the magnitude, while with the mixture phase, the PSM is the optimal mask to estimate the magnitude, which also explains why the PSM brings better separation results than IAM and IRM in [19], [43].

3.3 Mask Estimation

In this study, the masks used for phase recovery and magnitude estimation are estimated with deep neural networks. Figure 1 shows the diagram of the training procedure of

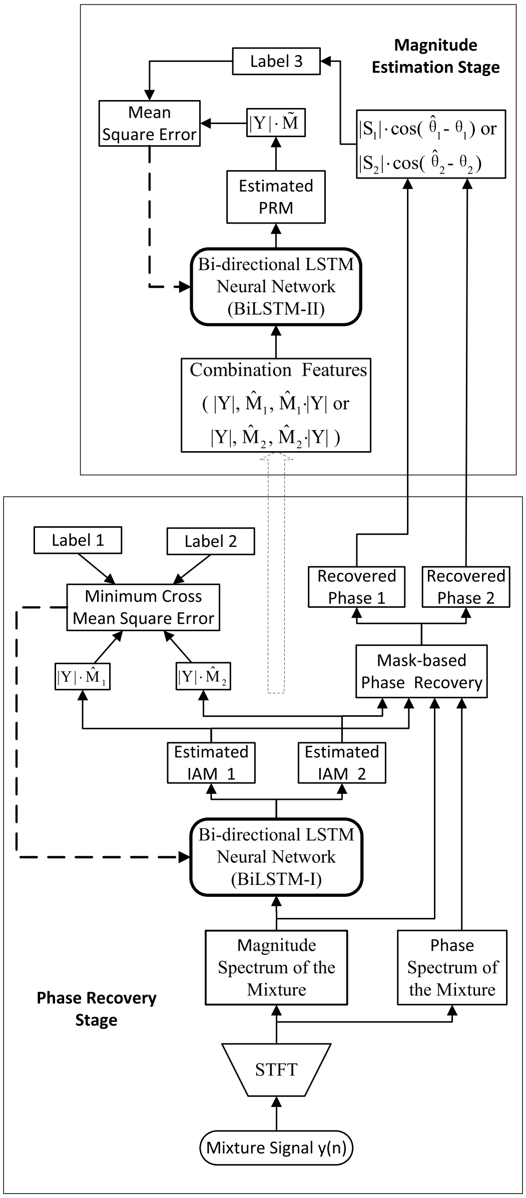

Magnitude Estimation Stage
Label 3
Mean Square Error
$|Y|\cdot\tilde{M}$
$|S_1|\cdot\cos(\hat{\theta}_1 - \theta_1)$ or
$|S_2|\cdot\cos(\hat{\theta}_2 - \theta_2)$
Estimated PRM
Bi-directional LSTM Neural Network (BiLSTM-II)
Combination Features
($|Y|, \hat{M}_1, \hat{M}_1\cdot|Y|$ or
$|Y|, \hat{M}_2, \hat{M}_2\cdot|Y|$)
Label 1
Label 2
Minimum Cross Mean Square Error
Recovered Phase 1
Recovered Phase 2
$|Y|\cdot\hat{M}_1$
$|Y|\cdot\hat{M}_2$
Mask-based Phase Recovery
Estimated IAM 1
Estimated IAM 2
Bi-directional LSTM Neural Network (BiLSTM-I)
Magnitude Spectrum of the Mixture
Phase Spectrum of the Mixture
Phase Recovery Stage
STFT
Mixture Signal y(n)

Fig. 1 The block diagram of the training procedure of the two-stage approach for phase-aware multi-talker speech separation.

the proposed phase-aware speech separation system, which mainly consists of a phase recovery module and a magnitude estimation module. The phase recovery module trains a neural network to estimate the IAM, and uses the IAM to recover the phase. The magnitude estimation module trains a neural network to estimate the PRM for magnitude estimation.

3.3.1 Estimating the IAM for Phase Recovery

A bi-directional LSTM neural network is used to estimate the IAM for the mask-based phase recovery algorithm. The neural network is denoted as BiLSTM-I in Fig. 1, of which the input feature is the magnitude spectrum of the observed mixture and the outputs are the estimated masks corresponding to every source. During the training procedure shown in Fig. 1, the minimum cross mean square loss (MCL) is used as the loss function for back propagation. The MCL is calculated as the minimum error from Normal Mean Square Error (NMSE) and Cross Mean Square Error (CMSE) of the estimated masks to the labels [18], [49]. Let $loss_{\text{NMSE}}$ and $loss_{\text{CMSE}}$ denote NMSE and CMSE, respectively. Hence,

$$loss_{\text{NMSE}} = \sum_{(k,l)} (\|\hat{\mathcal{M}}_1 \cdot |Y| - |\mathcal{S}|_1\|^2 + \|\hat{\mathcal{M}}_2 \cdot |Y| - |\mathcal{S}|_2\|^2) \tag{11}$$

$$loss_{\text{CMSE}} = \sum_{(k,l)} (\|\hat{\mathcal{M}}_1 \cdot |Y| - |\mathcal{S}|_2\|^2 + \|\hat{\mathcal{M}}_2 \cdot |Y| - |\mathcal{S}|_1\|^2) \tag{12}$$

where $\hat{\mathcal{M}}_1$ and $\hat{\mathcal{M}}_2$ denote the estimated ideal amplitude masks, $\mathcal{S}_1$ and $\mathcal{S}_2$ denote the magnitude of each target speaker. The minimum loss function is defined as:

$$loss_{\text{MCL}} = \lambda \cdot loss_{\text{NMSE}} + (1 - \lambda) \cdot loss_{\text{CMSE}} \tag{13}$$

where λ is a chosen determiner, if $loss_{\text{NMSE}} \leq loss_{\text{CMSE}}$, $\lambda = 1$, otherwise $\lambda = 0$. Here, λ ensures that the smaller one between the $loss_{\text{NMSE}}$ and $loss_{\text{CMSE}}$ is used as the final training loss function in the back propagation procedure. The signal approximation method [50] and the utterance-level training are used in Eq. (11) and Eq. (12) to improve the training performance. Note that, although the loss is calculated on the signal level, the outputs of the neural network are the expected masks.

3.3.2 Estimating the PRM for Magnitude Estimation

This study propose two frameworks for PRM estimation, which are denoted with BiLSTM-II and BiLSTM-III and shown in Fig. 2 and Fig. 3, respectively. The difference between the BiLSTM-II and BiLSTM-III is that for the BiLSTM-II, multiple neural networks shared parameters are parallel used corresponding to target speakers, while for the BiLSTM-III, a single neural network is used to enhance every target speaker.

(1) BiLSTM-II Neural Network

As shown in Fig. 2, multiple BiLSTM-II neural networks shared parameters are used to estimate the PRMs for magnitude estimation. Each of the BiLSTM-II works as an speech enhancement neural network extracting one of the target magnitudes in a mixture. The input of BiLSTM-II is a combination of the mixture magnitude and one of the masks from BiLSTM-I. For better generalization, during training, every utterance in a mixture is separately used to train a BiLSTM-II, which is shown in Fig. 1. The BiLSTM-

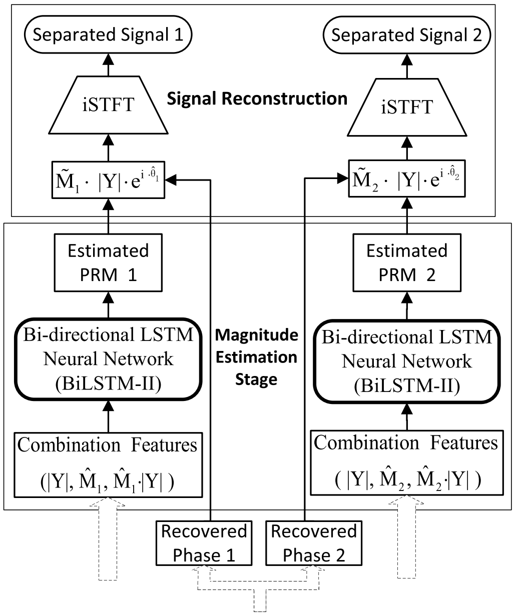

Separated Signal 1
Separated Signal 2
iSTFT
Signal Reconstruction
iSTFT
$\tilde{M}_1 \cdot |Y| \cdot e^{i\hat{\theta}_1}$
$\tilde{M}_2 \cdot |Y| \cdot e^{i\hat{\theta}_2}$
Estimated PRM 1
Estimated PRM 2
Bi-directional LSTM Neural Network (BiLSTM-II)
Magnitude Estimation Stage
Bi-directional LSTM Neural Network (BiLSTM-II)
Combination Features ($|Y|, \hat{M}_1, \hat{M}_1 \cdot |Y|$)
Combination Features ($|Y|, \hat{M}_2, \hat{M}_2 \cdot |Y|$)
Recovered Phase 1
Recovered Phase 2

Fig. 2 The block diagram of the testing procedure with the BiLSTM-II for magnitude estimation. The masks of every speaker are estimated by multiple BiLSTM-II neural networks.

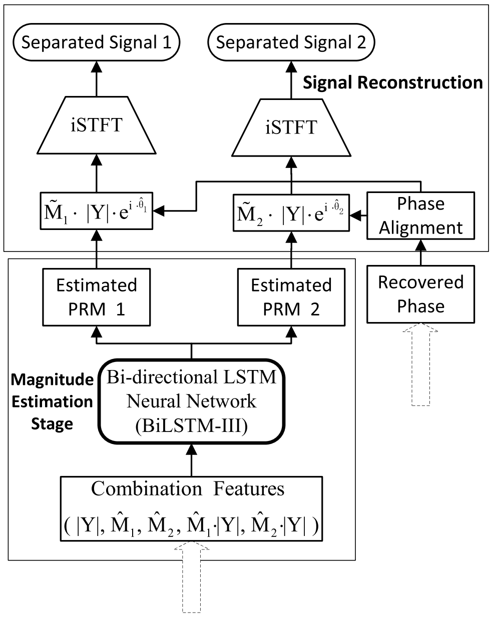

Separated Signal 1
Separated Signal 2
Signal Reconstruction
iSTFT
iSTFT
$\tilde{M}_1 \cdot |Y| \cdot e^{i\hat{\theta}_1}$
$\tilde{M}_2 \cdot |Y| \cdot e^{i\hat{\theta}_2}$
Phase Alignment
Estimated PRM 1
Estimated PRM 2
Recovered Phase
Magnitude Estimation Stage
Bi-directional LSTM Neural Network (BiLSTM-III)
Combination Features
($|Y|, \hat{M}_1, \hat{M}_2, \hat{M}_1 \cdot |Y|, \hat{M}_2 \cdot |Y|$)

Fig. 3 The block diagram of the testing procedure with the BiLSTM-III for magnitude estimation. The masks of every speaker are simultaneously estimated by a single BiLSTM-III.

II is trained with the signal approximation loss function:

$$loss = \sum_{(k,l)} (\|\widetilde{\mathcal{M}}_j \cdot |Y| - |\mathcal{S}|_j cos(\hat{\theta}_j - \theta_j)\|^2) \qquad (14)$$

During testing, the parameters of the trained BiLSTM-II are shared with other BiLSTM-II neural networks..

(2) BiLSTM-III Neural Network

The BiLSTM-III simultaneously estimates every PRM of target speaker, as shown in Fig. 3. The input of this neural network is the combination of the mixture magnitude and the masks of every speaker from BiLSTM-I. The purpose of the BiLSTM-III is to further improve the mask estimation performance by exploiting the underlying relationship of the signals. Therefore, the masks of every speaker are fed to the BiLSTM-III. Since the multiple outputs of BiLSTM-III bring the permutation problem, the minimum cross square loss is utilized during training procedure to solve the label permutation problem. The loss function ($loss_{\text{MCL-III}}$) for BiLSTM-III is the minimum value of $loss_{\text{NMSE-III}}$ and $loss_{\text{CMSE-III}}$, here $loss_{\text{NMSE-III}}$ and $loss_{\text{CMSE-III}}$ are respectively defined as:

$$\begin{aligned} loss_{\text{NMSE-III}} = & \sum_{(k,l)} (\|\widetilde{\mathcal{M}}_1 \cdot |Y| - |\mathcal{S}|_1 cos(\hat{\theta}_1 - \theta_1)\|^2 \\ & + \|\widetilde{\mathcal{M}}_2 \cdot |Y| - |\mathcal{S}|_2 cos(\hat{\theta}_2 - \theta_2)\|^2) \end{aligned} \qquad (15)$$

$$\begin{aligned} loss_{\text{CMSE-III}} = & \sum_{(k,l)} (\|\widetilde{\mathcal{M}}_1 \cdot |Y| - |\mathcal{S}|_2 cos(\hat{\theta}_2 - \theta_2)\|^2 \\ & + \|\widetilde{\mathcal{M}}_2 \cdot |Y| - |\mathcal{S}|_1 cos(\hat{\theta}_1 - \theta_1)\|^2) \end{aligned} \qquad (16)$$

During testing, the estimated magnitudes are not aligned with the recovered phases due to the permutation problem. Since the recovered phases are aligned with the outputs of BiLSTM-I, this issue is addressed in this study by aligning the outputs of BiLSTM-III with the outputs of BiLSTM-I, with the minimum mean square error between the IAMs and the PRMs. The alignment between the estimated PRMs and the phases is called phase alignment in Fig. 3.

4. Experimental Setup and Evaluation Result

4.1 Dataset

The proposed approach is first trained and evaluated on the Wall Street Journal (WSJ0) corpus [51]. The WSJ0 mixtures are generated with the WSJ0-2mix list [15], which is widely used for speech separation task. The training set, validation set and evaluation set consist of 20000 utterances, 5000 utterances and 3000 utterances, respectively. The speakers in the validation set are seen (closed-condition CC) in the training set, and are unseen (open-condition OC) in the evaluation set. All utterances are mixed at signal-to-signal ratios (SSRs) uniformly chosen between 0 dB and 5 dB. For further evaluation of generalization, the trained neural networks are evaluated on the TIMIT corpus [52]. The TIMIT mixtures are mixed in two sets with the SSRs of 0dB and 5dB, respectively. Each of the TIMIT evaluation set contains 200 utterances. All the utterances are re-sampled to 8000 Hz before mixing. The analysis window is hanning window with the length of 256 samples (32ms) and the overlap of 128 samples (16ms).

4.2 Neural Networks

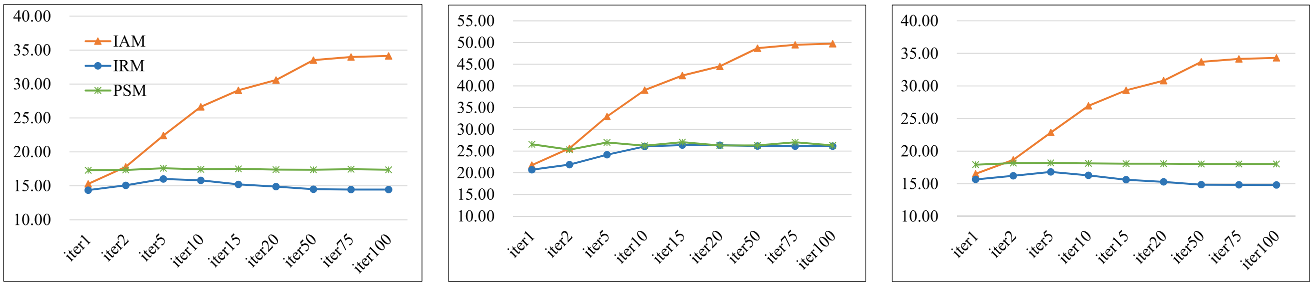
Each of the neural network BiLSTM-I, BiLSTM-II and BilSTM-III has three hide-layers. Each hide-layer has 896 bi-directional LSTM units with Tanh activation function. The output layers of each neural network are feed forward networks with Relu units. The output size of BiLSTM-I and BiLSTM-III is J*129, where J is the number of target speakers. Similarly, the output layer of BiLSTM-II has 129 units. The size of the input layers of BiLSTM-I, BiLSTM-II and BiLSTM-III are 129, 129*3, 129*(2*J+1) units, respectively. During training, the Adam learning algorithm [53] were used to train the neural networks with the initial learning rate of 0.0005. For each training epoch, if the loss of the validation set increases, the learning rate will be scaled by 0.7. The value of dropout [54] was set to 0.5.

4.3 Evaluation Metrics

To investigate the separation performance, the results are evaluated in three aspects: the overall distortion of the separated speech compared with the source speech, the remaining interference in the separated speech, and the distortion introduced by the separation procedure. To evaluate these aspects, the results are measured in the signal-to-distortion ratio (SDR), the signal-to-interferences ratio (SIR), and the signal-to-artifacts ratio (SAR), respectively, which are proposed for the performance measurement in blind source separation [55]. The values are united in dB. For the phase recovery and magnitude estimation, the approach with higher score brings less distortion and is deemed to be more effective.

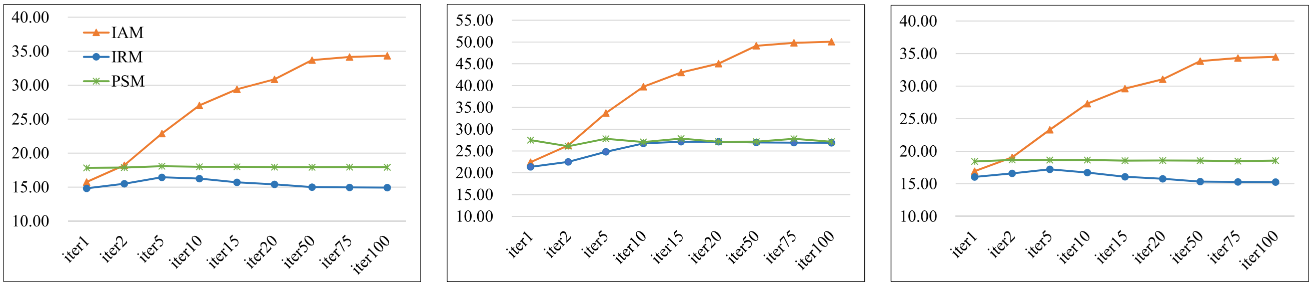
4.4 Evaluations on the Performance of the IAM for Phase Recovery

In this paper, the IAM is proposed to be the optimal mask for the phase recovery, compared with the IRM and the PSM. To evaluate the effectiveness of the IAM for the mask-based phase recovery algorithm, the performance of various oracle masks used for phase recovery is investigated first. The oracle mask shown in Table 1 is calculated and used in the iteration procedure (Eq. 5 to Eq. 7) to recover the phase. After phase recovery, to remove the effects of the magnitude on the results, the recovered phase is combined with the oracle magnitude to reconstruct the source signal. The result with higher score indicates less distortion of the reconstructed signal as well as less distortion of the recovered phase. The corresponding mask is therefore more appropriate for phase recovery. The results of the mask-based phase recovery on the WJS0 mixtures with the SSRs at 0 dB and 5



(a) SDR-0dB (b) SIR-0dB (c) SAR-0dB

Fig. 4 The performance of different oracle masks used for the mask-based phase recovery on 0 dB WSJ0 mixtures. The vertical axis denotes the score united in dB and the horizontal axis denotes the number of iterations.



(a) SDR-5dB (b) SIR-5dB (c) SAR-5dB

Fig. 5 The performance of different oracle masks used for the mask-based phase recovery on 5 dB WSJ0 mixtures. The vertical axis denotes the score united in dB and the horizontal axis denotes the number of iterations.

Table 2 The evaluations on the performance of the IAM for the mask-based phase recovery. To remove the effects of magnitude, the signals are reconstructed from the oracle magnitude and the phase recovered with different masks. The results are evaluated in terms of SDR, SIR and SAR in the units of *dB*. Boldface highlights the best result.

SSR	Iteration Index	SDR			SIR			SAR		
		IAM	IRM	PSM	IAM	IRM	PSM	IAM	IRM	PSM
0dB	iter0	12.81	12.81	12.81	18.48	18.48	18.48	14.37	14.37	14.37
	iter1	13.36	13.27	13.62	20.27	20.22	21.37	14.53	14.45	14.60
	iter3	14.07	13.62	13.86	22.14	21.41	22.17	**14.95**	14.57	14.71
	iter6	**14.19**	13.59	13.88	23.25	22.22	22.38	14.88	14.36	14.69
	iter10	14.00	13.19	13.85	23.77	22.74	22.45	14.58	13.81	14.64
	iter15	13.72	12.63	13.82	**24.01**	22.93	22.48	14.24	13.16	14.60
5dB	iter0	13.31	13.31	13.31	19.21	19.21	19.21	14.80	14.80	14.80
	iter1	14.10	13.91	14.58	21.16	20.98	22.39	15.23	15.05	15.53
	iter3	14.94	14.28	14.84	23.21	22.19	23.21	**15.79**	15.21	15.67
	iter6	**15.11**	14.25	14.86	24.44	23.05	23.42	15.75	15.00	15.65
	iter10	14.91	13.84	14.83	25.00	23.61	23.49	15.46	14.44	15.61
	iter15	14.65	13.29	14.80	**25.25**	23.85	23.52	15.13	13.78	15.57
Avg.	iter0	13.06	13.06	13.06	18.84	18.84	18.84	14.58	14.58	14.58
	iter1	13.73	13.59	14.10	20.72	20.60	21.88	14.88	14.75	15.06
	iter3	14.51	13.95	14.35	22.68	21.80	22.69	**15.37**	14.89	15.19
	iter6	**14.65**	13.92	14.37	23.85	22.63	22.90	15.32	14.68	15.17
	iter10	14.45	13.51	14.34	24.39	23.18	22.97	15.02	14.13	15.13
	iter15	14.18	12.96	14.31	**24.63**	23.39	23.00	14.69	13.47	15.08

dB are shown in Fig. 4 and Fig. 5, respectively. As is shown, with the iterations, the phase recovery using the IAM brings significant improvement in SDR, SIR and SAR. While the SIR of the phase recovery using the PSM and IRM is only slightly increased, the SDR and the SAR using the IRM begin to decline after 5 iterations, which means that using the IAM for the mask-based phase recovery brings less phase distortion.

To further evaluate the performance, the BiLSTM-I neural networks are trained to estimate the mask to be used

Table 3 The performance of the PRM for magnitude estimation. For comparison, the magnitudes are estimated with different oracle masks and the phase is recovered with the IAM.

	SSR	IAM	IRM	PSM	PRM (*proposed*)
SDR	0dB	14.19	13.31	14.38	**17.03**
	5dB	15.11	14.10	15.21	**17.80**
	Avg.	14.65	13.71	14.80	**17.42**
SIR	0dB	23.25	20.97	24.61	**28.03**
	5dB	24.44	22.01	25.50	**28.87**
	Avg.	23.85	21.49	25.06	**28.45**
SAR	0dB	14.88	14.29	14.91	**17.46**
	5dB	15.75	15.05	15.72	**18.22**
	Avg.	15.32	14.67	15.32	**17.84**

for phase recovery. As a comparison, the PSM and the IRM are also estimated with the BiLSTM-I neural networks and used to recover the phase. The results are listed in Table 2 with iterations of *iter*0 - *iter*15. With the improvement of the result, the signal distortion is reduced, due to the reduction of phase distortion. As it can be seen, for every evaluation item, the phase recovery using the IAM has the best results and brings the least phase distortion. For the phase recovery using the IAM, the SDR, SIR and SAR results are saturated after 6 iterations, 15 iterations and 3 iterations, respectively. Considering to minimize the overall distortion, the epoch for the phase recovery is set to 6 in the following section. Comparing the results of separation using the phase recovery (IAM-*iter*6) with the results using the mixture phase (*iter*0), the average (Avg.) SDR, SIR and SAR are improved by 1.59 dB, 5.01 dB and 0.74 dB, respectively. The results in Table 2 show that the IAM is more effective for the phase recovery than the PSM and IRM.

4.5 Evaluations on the Effectiveness of the PRM for Magnitude Estimation

This paper proposes the PRM to estimate the magnitude for the phase-recovered speech separation. To study the performance of the PRM, the magnitude estimated with the PRM is combined with the recovered phase to reconstruct the source signal. The phase is recovered using the IAM as evaluated in Table 2. For comparison, the PSM, IRM and IAM are also used to estimate the magnitude. For the different separations of a specified mixture, only the magnitudes are different while the phases are the same. Therefore, the differences of the results are caused by the differences of the magnitude and the higher score indicates less magnitude distortion. The evaluation results are shown in Table 3. As shown, with the recovered phase, the magnitude estimated with the PRM brings the best results in terms of SDR, SIR and SAR. The results demonstrate that, with the recovered phase, the proposed PRM achieves better performance for magnitude estimation than conventional masks. This is because the optimization of the PRM for magnitude estimation compensates for the error of phase recovery.

Table 4 The SDR, SIR and SAR improvements of the proposed two-stage phase-aware speech separation. The results titled IAM_I and PSM_II are listed as comparisons. Boldface highlights the best result.

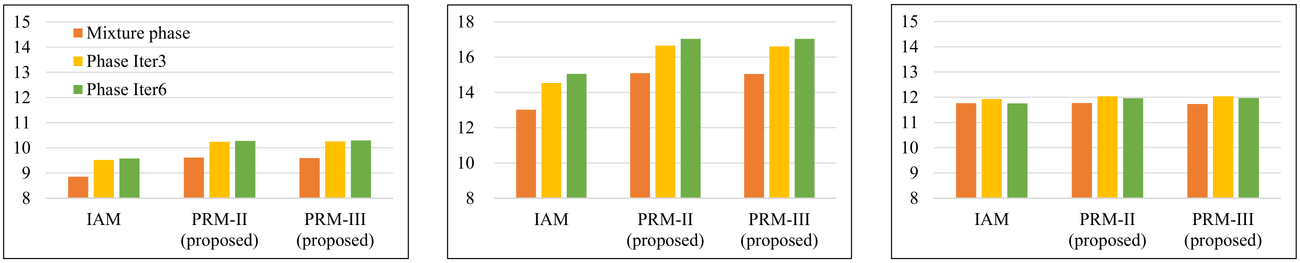
	Phase	IAM_I	PSM_II	PRM_II	PRM_III
SDR	Mixture	9.82	11.20	10.55	10.54
	Recovered	10.96	11.55	**11.81** (*proposed*)	11.78
SIR	Mixture	15.34	19.04	17.50	17.51
	Recovered	18.26	20.35	**20.67** (*proposed*)	20.64
SAR	Mixture	11.78	12.41	11.96	11.96
	Recovered	12.30	12.56	**12.82** (*proposed*)	12.79

4.6 Results of the Proposed Two-Stage Phase-Aware Speech Separation

Table 4 shows the results of the proposed two-stage speech separation. The target speech is reconstructed from the phase recovered with IAM and the magnitude estimated with PRM. The columns titled PRM_II and PRM_III are the results corresponding to the magnitude estimation with different neural networks of BiLSTM-II and BiLSTM-III, respectively. The best results of one-stage approach titled IAM_I are listed in Table 4 as a comparison. The BiLSTM-II neural network is also used to estimate the PSM for magnitude estimation and the results are titled with PSM_II. As can be seen, the performance of the proposed approach is significantly improved, compared to the one-stage approach (IAM_I). With the recovered phase, the SDR, SIR and SAR of the proposed approach are improved by 0.85 dB, 2.41 dB and 0.52 dB respectively, compared with IAM_I. The SDR of the conventional stacking model using mixture phase is 10.0 dB in [19], as a comparison, the neural network proposed in this study (PSM_II with mixture phase) brings 1.2 dB improvement, and with the recovered phase the SDR is further improved by 0.35 dB. Comparing the proposed mask PRM with PSM, the SDR, SIR and SAR are improved by 0.26 dB, 0.32 dB and 0.26 dB respectively for the scenario with phase recovery. The higher scores demonstrate the higher efficiency of the PRM over the PSM. For the results of the BiLSTM-II and BiLSTM-III, only slightly differences are observed. The SDR, SIR and SAR of the separation using recovered phase are improved by 1.26 dB, 3.13 dB, and 0.84 dB respectively, compared with the results using mixture phase, which clearly demonstrates the effectiveness of the mask-based phase recovery.

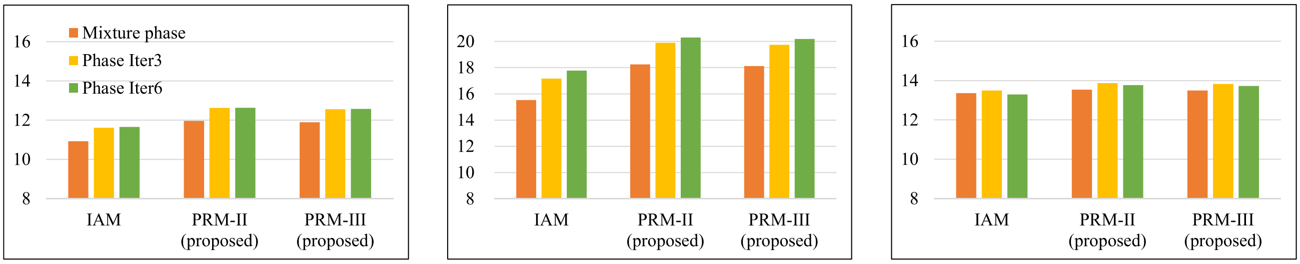
4.7 Evaluations of Generalization on Untrained Datasets

To further investigate the generalization abilities, the neural networks trained with the WSJ0 corpus are evaluated on the TIMIT dataset. The results in Fig. 6 and Fig. 7 show that the performance of the proposed approach (PRM_II and PRM_III) are better than the one stage approach (IAM) on the untrained TIMIT mixtures. For the proposed approach, the SDR and SIR of the results using phase recovery are signif-



(a) SDR improvements on 0 dB mixtures (b) SIR improvements on 0 dB mixtures (c) SAR improvements on 0 dB mixtures

Fig. 6 The results of the two-stage phase-aware speech separation on 0 dB TIMIT mixtures.



(a) SDR improvements on 5 dB TIMIT mixtures (b) SIR improvements on 5 dB TIMIT mixtures (c) SAR improvements on 5 dB TIMIT mixtures

Fig. 7 The results of the two-stage phase-aware speech separation on 5 dB TIMIT mixtures.

icantly improved, compared with the results using mixture phase. While the SAR of the separation using the phase recovered after 6 iterations is slightly declined, compared to the result using the phase recovered after 3 iterations, it is greatly improved than the result using the mixture phase, which is consistent to the previous result in Sect. 4.4. The results in Fig. 6 and Fig. 7 show the effectiveness and good generalization of the proposed approach on the untrained dataset.

4.8 Comparisons with Other Speech Separation Methods on WSJ0-2mix

Table 5 Comparisons with other methods on WSJ0-2mix.

Approaches	SDR	
	CC	OC
uPIT-ST [19]	10.0	10.0
Chimera++ [40] + MISI [40]	11.1 11.4	11.2 11.5
cIRM	-	8.6
TasNet [56]	-	11.1
Proposed(Phase-IAM+Magnitue-PSM) Proposed+(Phase-IAM+Magnitue-PRM)	11.52 11.76	11.55 11.81

The results of different methods evaluated on WSJ0-2mix utterances are given in Table 5. The two-stage phase-aware approach was compared with uPIT stacking (uPIT-ST) model [19], complex ideal ratio mask (cIRM) model [48], chimera++ model [40] and time-domain audio separation network (TasNet) [56]. Comparison against uPIT-ST helps to evaluate the improvements of the proposed phase-recovered separation over the conventional speech separation. The uPIT-ST is an approach that stacks two BLSTM models, of which the training target is the PSM and the final mask is computed as the average mask from the two models. By uPIT-ST model, the target signal is constructed with the mixture phase. As shown in Table 5, the SDR of the proposed approach improves by 1.81 dB, comparing with that of uPIT-ST, which shows a significant advantage of the proposed approach over the conventional speech separation. In [48], the cIRM is used for the noisy speech enhancement. As a comparison, a neural network was trained in this study to learn the cIRM for multi-talker speech separation, on the WSJ0-2mix utterances. The SDR of the separation using cIRM is only 8.6 dB, which is lower than the proposed approach. This is mainly due to the ambiguity of the real part and the imaginary part of different speakers. Even though the label ambiguity problem is solved with PIT method [18] during the training procedure, the real part and imaginary part is still unaligned during the testing procedure. The minimum square error metric is used to solve the problem, but it does not work well. Another reason that decreases the performance of the cIRM is the unclear structure on the imaginary spectrum, which makes the imaginary part of the cIRM difficult to be estimated. The chimera++ network combines the deep clustering with mask-inference in a multi-task training approach, which leverages the performance of mask inference. In [40], the chimera++ model is used to infer the PSM and gains 11.2 dB SDR with the mixture phase and 11.5 dB SDR with the recovered phase, respectively. However, using the PSM for phase recovery is not an optimal way. Different from the frequency domain speech separation, the TasNet directly operates on the sound waveforms. The proposed two-stage approach achieves 0.31 dB improvement over the chimer++ model and 0.71 dB improvement over the TasNet. The comparisons show that the proposed speech separation brings better performance compared to the current state-of-the-art works.

5. Concluding Remarks

Generally, mixed utterances are separated in the frequency domain and only the spectral magnitude is separated, while the mixture phase remains unchanged. Given that the magnitude and phase are two important parts of speech signal, this paper proposes a two-stage phase-aware approach for multi-talker speech separation, which optimally estimates the magnitude and recovers the phase. Also, this study proposes that the optimal mask for the mask-based phase recovery algorithm is the IAM, rather than the IRM and the PSM. Furthermore, to compensate for the error of the phase recovery, the PRM is proposed for magnitude estimation, which minimizes the complex distance between the separated speech and the source speech. The first stage is used to infer the IAM for phase recovery and the second stage is used to infer the PRM for magnitude estimation. Two different structures for the second stage are given in this study, with which similar separation performance are observed. The results demonstrate that the proposed approach brings better separation performance, compared with the state-of-the-art works. Future works include combining phase information into the input of neural networks for mask estimation, jointly training the two-stage neural networks for better performance and directly training neural networks to recover the phase for multi-talker separation.

References

[1] D.D. Lee and H.S. Seung, “Learning the parts of objects by nonnegative matrix factorization,” Nature, vol.401, no.6755, pp.788–790, 1999.

[2] M.N. Schmidt and R.K. Olsson, “Single-channel speech separation using sparse non-negative matrix factorization,” The International Conference on Spoken Language Processing, pp.2614–2617, 2006.

[3] F. Weninger, J.L. Roux, J.R. Hershey, and S. Watanabe, “Discriminative NMF and its application to single-channel source separation,” INTERSPEECH, pp.865–869, 2014.

[4] J.L. Roux, J.R. Hershey, and F. Weninger, “Deep NMF for speech separation,” 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, pp.66–70, 2015.

[5] G.J. Brown and D.L. Wang, Separation of Speech by Computational Auditory Scene Analysis, Springer Berlin Heidelberg, 2005.

[6] P. Li, Y. Guan, B. Xu, and W. Liu, “Monaural speech separation based on computational auditory scene analysis and objective quality assessment of speech,” First International Conference on Innovative Computing, Information and Control - Volume I, pp.742–745, 2006.

[7] Y. Shao, S. Srinivasan, Z. Jin, and D.L. Wang, “A computational auditory scene analysis system for speech segregation and robust speech recognition,” Computer Speech and Language, vol.24, no.1, pp.77–93, 2010.

[8] A.P. Varga and R.K. Moore, “Hidden Markov model decomposition of speech and noise,” International Conference on Acoustics, Speech, and Signal Processing, vol.2, pp.845–848, 1990.

[9] A. Ozerov, C. Févotte, and M. Charbit, “Factorial scaled hidden Markov model for polyphonic audio representation and source separation,” IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp.121–124, 2009.

[10] Y. Wang, A. Narayanan, and D.L. Wang, “On training targets for supervised speech separation,” IEEE/ACM Trans. Audio Speech Lang. Process., vol.22, no.12, pp.1849–1858, 2014.

[11] X. Zhang and D. Wang, “Deep learning based binaural speech separation in reverberant environments,” IEEE/ACM Trans. Audio Speech Lang. Process., vol.25, no.5, pp.1075–1084, 2017.

[12] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, “Deep learning for monaural speech separation,” 2014 IEEE International Conference on Acoustics, Speech and Signal Processing, pp.1562–1566, 2014.

[13] C. Weng, D. Yu, M.L. Seltzer, and J. Droppo, “Deep neural networks for single-channel multi-talker speech recognition,” IEEE/ACM Trans. Audio Speech Lang. Process., vol.23, no.10, pp.1670–1679, 2015.

[14] Y. Wang, J. Du, L.-R. Dai, and C.-H. Lee, “Unsupervised single-channel speech separation via deep neural network for different gender mixtures,” Signal and Information Processing Association Summit and Conference, pp.1–4, 2016.

[15] J.R. Hershey, Z. Chen, J.L. Roux, and S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation,” 2016 IEEE International Conference on Acoustics, Speech and Signal Processing, pp.31–35, 2016.

[16] Y. Isik, J.L. Roux, Z. Chen, S. Watanabe, and J.R. Hershey, “Single-channel multi-speaker separation using deep clustering,” Interspeech 2016, pp.545–549, 2016.

[17] Z. Chen, Y. Luo, and N. Mesgarani, “Deep attractor network for single-microphone speaker separation,” 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, pp.246–250, 2017.

[18] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, “Permutation invariant training of deep models for speaker-independent multi-talker speech separation,” 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, pp.241–245, 2017.

[19] M. Kolbaek, D. Yu, Z.-H. Tan, and J. Jensen, “Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks,” IEEE/ACM Trans. Audio Speech Lang. Process., vol.25, no.10, pp.1901–1913, 2017.

[20] D.L. Wang and J. Lim, “The unimportance of phase in speech enhancement,” IEEE Trans. Acoust., Speech, Signal Process., vol.30, no.4, pp.679–681, Aug. 1982.

[21] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” IEEE Trans. Acoust., Speech, Signal Process., vol.32, no.6, pp.1109–1121, 1984.

[22] P.J. Wolfe and S.J. Godsill, “Simple alternatives to the ephraim and malah suppression rule for speech enhancement,” Proc. 11th IEEE Signal Processing Workshop on Statistical Signal Processing, pp.496–499, 2001.

[23] T. Lotter and P. Vary, “Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model,” EURASIP J. Adv. Signal Process., vol.2005, no.7, pp.1–17, 2005.

[24] K.K. Paliwal and L.D. Alsteris, “On the usefulness of STFT phase spectrum in human listening tests,” Speech Commun., vol.45, no.2, pp.153–170, 2005.

[25] M. Kazama, S. Gotoh, M. Tohyama, and T. Houtgast, “On the significance of phase in the short term fourier spectrum for speech intelligibility,” The Journal of the Acoustical Society of America, vol.127, no.3, pp.1432–1439, 2010.

[26] K. Paliwal, K. Wójcicki, and B. Shannon, “The importance of phase in speech enhancement,” Speech Commun., vol.53, no.4, pp.465–494, 2011.

[27] D. Griffin and J.S. Lim, “Signal estimation from modified short-time fourier transform,” IEEE Trans. Acoust., Speech, Signal Process.,

vol.32, no.2, pp.236–243, 1984.

[28] D. Gunawan and D. Sen, "Iterative phase estimation for the synthesis of separated sources from single-channel mixtures," IEEE Signal Process. Lett., vol.17, no.5, pp.421–424, 2010.

[29] S. Wisdom, J.R. Hershey, K. Wilson, J. Thorpe, M. Chinen, B. Patton, and R.A. Saurous, "Differentiable consistency constraints for improved deep speech enhancement," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.900–904, 2019.

[30] J. Jensen and J.H.L. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," IEEE Trans. Speech Audio Process., vol.9, no.7, pp.731–740, 2001.

[31] M. Krawczyk and T. Gerkmann, "STFT phase improvement for single channel speech enhancement," International Workshop on Acoustic Signal Enhancement, pp.1–4, 2012.

[32] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," IEEE/ACM Trans. Audio Speech Lang. Process., vol.22, no.12, pp.1931–1940, 2014.

[33] P. Mowlaee, R. Saiedi, and R. Martin, "Phase estimation for signal reconstruction in single-channel speech separation," INTERSPEECH, pp.1548–1551, 2012.

[34] P. Mowlaee and R. Saeidi, "Time-frequency constraints for phase estimation in single-channel speech enhancement," 2014 14th International Workshop on Acoustic Signal Enhancement, pp.337–341, 2014.

[35] Z.-Q. Wang, K. Tan, and D. Wang, "Deep learning based phase reconstruction for speaker separation: A trigonometric perspective," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.71–75, 2019.

[36] M. Unoki and M. Akagi, "A method of signal extraction from noisy signal based on auditory scene analysis," Speech Commun., vol.27, no.3-4, pp.261–279, 1999.

[37] N. Zheng and X.-L. Zhang, "Phase-aware speech enhancement based on deep neural networks," IEEE/ACM Trans. Audio Speech Lang. Process., vol.27, no.1, pp.63–76, 2019.

[38] S. Takamichi, Y. Saito, N. Takamune, D. Kitamura, and H. Saruwatari, "Phase reconstruction from amplitude spectrograms based on von-mises-distribution deep neural network," 2018 16th International Workshop on Acoustic Signal Enhancement, pp.286–290, 2018.

[39] N. Takahashi, P. Agrawal, N. Goswami, and Y. Mitsufuji, "Phasenet: Discretized phase modeling with deep neural networks for audio source separation.," Interspeech 2018, pp.2713–2717, 2018.

[40] Z.-Q. Wang, J.L. Roux, and J.R. Hershey, "Alternative objective functions for deep clustering," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, pp.686–690, 2018.

[41] Z.-Q. Wang, J.L. Roux, D. Wang, and J. Hershey, "End-to-end speech separation with unfolded iterative phase reconstruction," Interspeech 2018, pp.2708–2712, 2018.

[42] G. Wichern and J.L. Roux, "Phase reconstruction with learned time-frequency representations for single-channel speech separation," 2018 16th International Workshop on Acoustic Signal Enhancement, pp.396–400, 2018.

[43] H. Erdogan, J.R. Hershey, S. Watanabe, and J.L. Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, pp.708–712, 2015.

[44] J.L. Roux, G. Wichern, S. Watanabe, A. Sarroff, and J.R. Hershey, "Phasebook and friends: Leveraging discrete representations for source separation," IEEE J. Sel. Top. Signal Process. vol.13, no.2, pp.370–382, 2019.

[45] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," IEEE/ACM Trans. Audio Speech Lang. Process., vol.27, no.8, pp.1256–1266, 2019.

[46] A. Narayanan and D.L. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp.7092–7096, 2013.

[47] D.S. Williamson, Y. Wang, and D.L. Wang, "Complex ratio masking for joint enhancement of magnitude and phase," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing, pp.5220–5224, 2016.

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

[49] L. Yin, Z.T. Wang, R.S. Xia, J.F. Li, and Y.H. Yan, "Multi-talker speech separation based on permutation invariant training and beamforming," Interspeech 2018, pp.851–855, 2018.

[50] F. Weninger, J.R. Hershey, J.L. Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," 2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp.577–581, 2014.

[51] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete LDC93S6A," Philadelphia: Linguistic Data Consortium, 1993.

[52] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, and D.S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," NASA STI/Recon Technical Report N, vol.93, Feb. 1993.

[53] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," International Conference on Learning Representations, 2015.

[54] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," Journal of Machine Learning Research, vol.15, no.1, pp.1929–1958, 2014.

[55] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," IEEE Trans. Audio Speech Lang. Process., vol.14, no.4, pp.1462–1469, 2006.

[56] Y. Luo and N. Mesgarani, "TaSNet: time-domain audio separation network for real-time, single-channel speech separation," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, pp.696–700, 2018.

Lu Yin received the M.S. degree in Electronics Science and Technology from Beijing Institute of Technology, China, in 2017. He is currently pursuing the Ph.D. degree in Information and Signal Processing at the Institute of Acoustics, Chinese Academy of Sciences and the University of Chinese Academy of Sciences. His currently research interests include speech signal processing, deep learning, speech enhancement, noise reduction and speech separation.

Junfeng Li received the Ph.D. degree in Information Science from Japan Advanced Institute of Science and Technology (JAIST) in March 2006. From April 2006, he was a post-doctoral research fellow at Research Institute of Electrical Communication (RIEC), Tohoku University. From April 2007 to July 2010, he was an Assistant Professor in School of Information Science, JAIST. Since August 2010, he has been a Professor in Institute of Acoustics, Chinese Academy of Sciences. His research interests include speech signal processing and 3D audio technology. Dr. Li received the Best Student Award from the Acoustical Society of America in 2006, and the Best Paper Award from JCA2007 in 2007, and the Itakura Award from the Acoustical Society of Japan in 2012. Dr. Li is now serving as Subject Editor for Speech Communication and Editor for IEICE Trans. on Fundamentals of Electronics, Communication and Computer Sciences.

Yonghong Yan received the B.E. degree in Electronic Engineering from Tsinghua University, China, in 1990, and the Ph.D degree in Computer Science and Engineering from Oregon Graduate Institute of Science and Technology, USA, in 1995. Currently he is a professor at the Speech Acoustics and Content Understanding Laboratory, Chinese Academy of Sciences. His research interests include speech processing and recognition, language/speaker recognition and human computer interface.

Masato Akagi received the B.E. from Nagoya Institute of Technology in 1979, and the M.E. and Ph.D. Eng. from the Tokyo Institute of Technology in 1981 and 1984. He joined the Electrical Communication Laboratories of Nippon Telegraph and Telephone Corporation (NTT) in 1984. From 1986 to 1990, he worked at the ATR Auditory and Visual Perception Research Laboratories. Since 1992 he has been on the faculty of the School of Information Science of JAIST and is now a full professor. His research interests include speech perception, modeling of speech perception mechanisms in human beings, and the signal processing of speech.