pcIRM: Complex Ideal Ratio Masking for Speaker-Independent Monaural Source Separation with Utterance Permutation Invariant Training

Wen Zhang*†, Xiaoyong Li*¶, Aolong Zhou*†, Kaijun Ren*†, Junqiang Song*†
*College of Meteorology and Oceanology, National University of Defense Technology, China
†College of Computer Science and Technology, National University of Defense Technology, China
E-mail: {wenzhang109, sayingxmu, zhouaolong10, renkaijun, songjunqiang}@nudt.edu.cn

***Abstract*—Typical speech separation systems usually operate in the time-frequency (T-F) domain by enhancing the magnitude response and leaving the phase response unaltered. Recent studies, however, suggest that phase is important for perceptual quality, leading some researchers to consider magnitude and phase spectrum enhancements. The merging of the complex ideal ratio masking (cIRM) estimation and training with deep neural network (DNN) has been proved to be an effective way to improve speech separation. Furthermore, the label ambiguity (or permutation) problem has become a major barrier for speaker-independent multi-talker source separation, which prompts us to come up with new solutions. In this paper, to solve the problem of speaker-independent monaural source separation, we propose a novel method called *pcIRM*, which creatively achieves the cIRM estimation with the utterance-level permutation invariant training (uPIT). Specifically, pcIRM is implemented with the deep bidirectional LSTM (Bi-LSTM) RNN network, and evaluated with the WSJ0-2mix datasets. We report separation results for the proposed method and compare them to that of the existing state-of-the-art methods. Extensive experimental results demonstrate the advantages of our proposed pcIRM method in terms of the signal-to-distortion ratio (SDR) metric.**

***Index Terms*—speaker-independent, monaural source separation, cIRM, uPIT, Bi-LSTM, SDR.**

I. Introduction

Speech source separation is the task of extracting multiple speech signals, one for each speaker, from a mixture containing two or more voices [1], which is often referred to as the cocktail-party problem. Human has the remarkable ability to separate one sound source from others. In a cocktail party, it seems effortlessly for a person with normal hearing sense to separate the target speaker from other speakers and background interference, and easy to change to another target. However, the same tasks seem to be extremely difficult for automatic computing systems, especially when only a single microphone recording of the speech mixture is available.

The cocktail-party problem has raised great concern in recent years, due to its potential use in real-world applications such as robust automatic speech and speaker recognition, as well as hearing prosthesis and assisted living systems [2, 3], where speech overlapping is commonly observed.

Generally, we can categorize the source separation problems into monaural (i.e., single-channel) and array-based (i.e., multichannel) source separation problems, in terms of the number of microphones or channels. As for the former problem, researches extract the target speech or remove the interference signal from the mixed signal, mainly using the acoustic and statistical characteristics of the target speech and the interference signal. While in the latter problem, the spatial information of the signal is also available. The monaural speech separation problem still remains been very challenging, as only one speech recording is available and the spatial information that can be extracted is limited [4].

Many approaches have been developed to address the monaural source separation problem since the 1960s. Before the deep learning era, classic single-channel speech separation methods can be classified into three categories: the model-based method, the blind source separation (BSS) method [5], and the computational auditory scene analysis (CASA) [6] method. However, these methods have limited effectiveness when processing acoustic sources in multi-talker mixed speech captured in real environments, such as with unseen noises in a mixture, in low signal-to-noise ratio (SNR), and with limited computational resources. Hence, in real-environment scenarios, it is difficult to obtain the target speech signal with high quality consistently via the aforementioned methods.

Recently, a deep neural network (DNN) has been adopted as a regression model to solve the source separation problem, especially for the monaural case. According to the training objectives, DNN-based monaural source separation methods can be grouped into three categories, namely the masking-based method [1], the mapping-based method [1], and the signal approximation (SA) based method [7, 8]. The targets in the masking-based method describe the time-frequency (T-F) relationships of targets speech to interference, with a value ranged in $[0, 1]$, while the targets in the mapping-based method demonstrate the spectral representations of clean speech, in which the value range of the spectrum at each T-F point is large, i.e., $[0, +\infty]$. While the SA-based method is the combination of the masking-based method and mapping-based method, which is to train a ratio masking estimator that

¶Corresponding author

This work was supported by the National Key R&D Program of China (Grant No. 2018YFB0203801), the National Natural Science Foundation of China (Grant Nos. 61572510, 61702529, 61502511), and the China National Special Fund for Public Welfare (Grant No. GYHY201306003).

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

minimizes the difference between the spectral magnitude of target speech and that of estimated speech. In comparison, the masking-based method can lead to a more accurate neural network model than the mapping-based method [1].

The first mask-based training target applied in supervised source separation is the ideal binary mask (IBM), which is inspired by the auditory masking phenomenon and the exclusive allocation principle in auditory scene analysis [6]. Many researchers exploited IBM as a training target and obtained promising separation results [9]. Because of the inflexible decisions on each T-F unit of the IBM, the separated speech signal from the IBM-based methods is distorted. Naturally, the ideal ratio mask (IRM) is proposed to optimize the performance of the IBM, in which the T-F unit is assigned as the ratio of the target source energy to mixture energy [10]. The target speech signal separated by IRM-based methods often achieves better quality, compared with the IBM.

Although these DNN-based methods obtained state-of-the-art performance, the IBM and the IRM only use the magnitude information of the target signals when separating and synthesizing the clean speech signal, as the phase spectrum is considered unimportant for speech separation [11]. Nevertheless, Erdogan et al. have shown that the phase information is beneficial to predict an accurate mask and the estimated source [12], they develop the phase-sensitive masking (PSM) based method, which significantly outperforms the IBM and the IRM in terms of SDR. In addition, in [13], Williamson et al. employ both the magnitude and phase spectra to estimate the complex IRM (cIRM) by operating in the complex domain.

In source separation, if the target speakers are not allowed to change from training to testing, then it is called in the speaker-dependent situation. While if interfering speakers are allowed to change, but the target speaker is fixed, then it is called the target-dependent source separation. Similarly, the speaker-independent source separation is defined if none of the speakers are demanded to be the same between training and testing, which is the least constrained case. The label ambiguity (or permutation) problem [14] is the biggest obstacle for previous work perform poorly on speaker-independent multi-talker source separation. In speaker-independent situation, since the source separation model has multiple outputs, one for each mixing source, and they share the same input mixture, reference assigning can be tricky, especially when processing numerous utterances spoken by multiple speakers. In [15, 16], permutation invariant training (PIT) is proposed to solve this problem, and achieves great performance.

To address the aforementioned problems, we propose a novel method called *pcIRM*, which creatively achieves the cIRM estimation with the utterance-level permutation invariant training. Specifically, pcIRM adopts the Bi-LSTM RNN to estimate the cIRM, and further exploits the criterion of the utterance permutation invariant training (uPIT). The contributions of this paper are summarized as follows:

- We propose a Y-shaped Bi-LSTM RNN to predict the cIRM as the training target in our pcIRM model, making use of both the amplitude and the phase information of the clean speech signal.
- We exploit the utterance permutation invariant training to overcome the label ambiguity problem for speaker independent multi-talker source separation, which is the first job integrating the cIRM estimate and the utterance PIT as a whole model.
- We conduct extensive experiments on different training targets to validate the effectiveness and efficiency of our proposals.

The rest of the paper is organized as follows. In Section II, we describe the background knowledge related to the training targets in recent monaural source separation methods. Section III introduces the novel criterion of utterance permutation invariant training and the proposed pcIRM-based source separation method. Section IV presents the experimental settings and results. Finally, the conclusions and future work are given in Section V.

II. Masking-based Training Targets

As the IRM, PSM and cIRM are the targets often chosen in the existing state-of-the-art masking-based DNN methods, we briefly describe them in the next subsections, respectively.

A. Ideal Ratio Mask (IRM)

Let us denote the target speech signal, the interference, and the mixed source signal sequences as $s(m)$, $i(m)$, and $y(m) = s(m)+i(m)$ at discrete time m, respectively. The corresponding short-time Fourier transformation (STFT) of these signals are $S(t,f)$, $I(t,f)$, and $Y(t,f) = S(t,f) + I(t,f)$, respectively, where f is the index of the frequency bins and t is the index of the time frames. In addition, given $Y(t,f)$, the goal of monaural speech separation is to recover each target source $S(t,f)$. By adopting the ideal T-F mask $M(t,f)$, the spectrum of the target speech can be reconstructed as follows:

$$S(t,f) = Y(t,f) * M(t,f) \tag{1}$$

where '$*$' indicates complex multiplication. The $M(t,f)$ for time frame t and frequency f can be expressed as:

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

where β is a tunable parameter to scale the mask, while $|S(t,f)|$ and $|I(t,f)|$ denote the magnitude spectrum of the target speech signal and the magnitude spectrum of interference, respectively. In addition, $|S(t,f)|^2$ and $|I(t,f)|^2$ represent the target speech power spectrum and the interference speech power spectrum within a T-F unit, respectively. Generally, β is selected as 0.5.

Obviously, in IRM, only magnitude information is exploited, while the phase information of the target speech signal is not used in speech reconstruction. To overcome this drawback, PSM and cIRM are proposed.

B. Phase-Sensitive Mask (PSM)

In polar coordinates, the STFT of speech signal can be defined as Equation (3).

$$S(t,f) = |S(t,f)|e^{j\theta_{S(t,f)}} \tag{3}$$

where $|S(t,f)|$ denotes the magnitude response, and $\theta_{S(t,f)}$ denotes the phase response of the STFT speech signal at time t and frequency f, which is commonly used when enhancing or separating the STFT of noisy speech. In polar coordinates, it is easy to understand the PSM, which extends the IRM by incorporating a measure of phase:

$$PSM(t,f) = \frac{|S(t,f)|}{|Y(t,f)|} cos(\angle Y - \angle S) \tag{4}$$

where $\angle Y$ and $\angle S$ denote the mixture speech phase and the target speech phase within the T-F unit, respectively. The inclusion of the phase difference between the mixture speech and the target speech in PSM brings about a higher SNR, and tends to produce a better estimate of target speech than the IRM. Apparently, the values of $\frac{|S(t,f)|}{|Y(t,f)|}$ and $|cos(\angle Y - \angle S)|$ are bounded within the range of $(0,1)$, and the latter may take a negative value.

C. Complex Ideal Ratio Mask (cIRM)

The cIRM is a complex T-F mask, which is calculated by using the real and imaginary components of the STFTs of the target and mixture speech signals. The STFTs of the mixture, clean signal, and the cIRM can be defined as:

$$Y(t,f) = Y_r(t,f) + jY_c(t,f) \tag{5}$$

$$S(t,f) = S_r(t,f) + jS_c(t,f) \tag{6}$$

$$cIRM(t,f) = cIRM_r(t,f) + jcIRM_c(t,f) \tag{7}$$

where $j = \sqrt{-1}$, and the subscripts r and c indicate the real and imaginary components, respectively. The index of the frequency bins f and the index of the time frames t are omitted for convenience below, but Y, S, and $cIRM$ are defined for each T-F unit. Thus, in the complex domain, Equation (1) can be further rewritten as:

$$S_r + jS_c = (Y_r + jY_c) * (cIRM_r + jcIRM_c) \tag{8}$$

$$S_r = cIRM_r * Y_r - cIRM_c * Y_c \tag{9}$$

$$S_c = cIRM_r * Y_c + cIRM_c * Y_i \tag{10}$$

Using Equations (9) and (10), the real and imaginary components of cIRM can be derived as follows:

$$cIRM_r = \frac{Y_r S_r + Y_c S_c}{{Y_r}^2 + {Y_c}^2} \tag{11}$$

$$cIRM_c = \frac{Y_r S_c - Y_c S_r}{{Y_r}^2 + {Y_c}^2} \tag{12}$$

Therefore, we can obtain the definition for the cIRM as:

$$cIRM = \frac{Y_r S_r + Y_c S_c}{{Y_r}^2 + {Y_c}^2} + j\frac{Y_r S_c - Y_c S_r}{{Y_r}^2 + {Y_c}^2} \tag{13}$$

It is worth noting that the value range of Y_r, Y_c, S_r and S_r are $\mathbb{R}$, meaning that $cIRM_r \in \mathbb{R}$ and $cIRM_c \in \mathbb{R}$, whose values are unbounded. As aforementioned, IRM gets with a range in $[0,1]$, which is favorable for supervised learning with DNNs. Therefore, we compress the cIRM with the following hyperbolic tangent function:

$$\begin{aligned} {cIRM_x}' &= K \cdot tanh(C \cdot cIRM_x) \\ &= K\frac{1 - e^{-2C \cdot cIRM_x}}{1 + e^{-2C \cdot cIRM_x}} \end{aligned} \tag{14}$$

where x is r or c, meaning the real or imaginary components. This compression operation limits mask values within $[-K, K]$, and C controls its steepness. Several pairs of values for K and C are evaluated in this study, and we find that when $K = 10$ and $C = 0.05$, the DNN-based source separation model performs best empirically. During the testing stage, the DNN outputs are the estimations of the compressed masks instead of the original masks, we apply the following inverse function to recover the estimation of the uncompressed mask.

$$cIRM_x = \frac{1}{C} arctanh(\frac{O_x}{K}) = -\frac{1}{2C} log(\frac{K - O_x}{K + O_x}) \tag{15}$$

where $cIRM_x$ represents the estimation of the uncompressed mask, and O_x is the DNN output.

In [13], Williamson finds that structures exist in both real and imaginary components of the cIRM in Cartesian coordinates, while in polar coordinates, structures exist in the magnitude spectrum, but not in the phase spectrogram. Figure 1 shows the comparisons of these two circumstances. Conspicuous and similar structures can be observed in child diagrams (a), (c) and (d). According to the results of Lee [17], direct phase estimation is difficult without a clear structure. Moreover, an estimation of the cIRM provides both the amplitude and phase estimate. In theory, cIRM is superior to PSM for more accurate estimates of source speech.

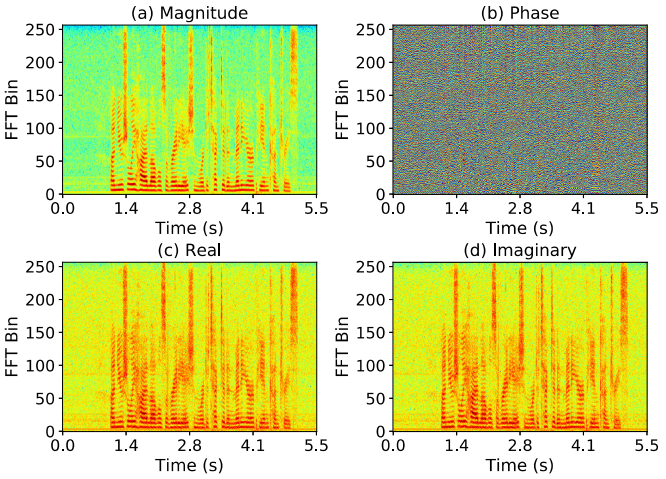

Fig. 1. Example magnitude (a) and phase (b) spectrograms, and real (c) and imaginary (d) spectrograms, for a clean speech signal. The real and imaginary spectrograms describe temporal and spectral structure and are close to the magnitude spectrogram. Little structure exists in the phase spectrogram.

III. Proposed Method

A. Network Architectures

In monaural source separation, most of the previous works are based on DNNs or RNNs, due to their flexibility and effectiveness. Moreover, the deep LSTM RNNs are capable of operating the utterance frame-by-frame over the whole past history information at each layer. Besides, the research in [18] has proved that the speaker generalization ability of the source separation method can be improved with the LSTM RNN. With deep bidirectional LSTM (called Bi-LSTM) [19], the information from the past and future (across the whole utterance) is stacked at each layer and fed into the next layer, which performs better than unidirectional LSTM when involving the processing of temporal sequences. Thus, the Bi-LSTM RNN is used as the framework of our proposed method.

Since the training target is the complex ideal ratio masking, the outputs of the Bi-LSTM RNNs are dual, one for real component and the other for the imaginary component of the prediction, which is a Y-shaped network structure. In contrast to this circumstance, the outputs of IRM-based and PSM-based LSTM RNN models are both single output. The architecture of the Y-shaped neural network is depicted in Figure 2, where the input features are the STFT spectrum of the mixture speech, and the outputs of real component and imaginary component are optimized individually. The specific settings of the network parameters are presented in section IV in detail.

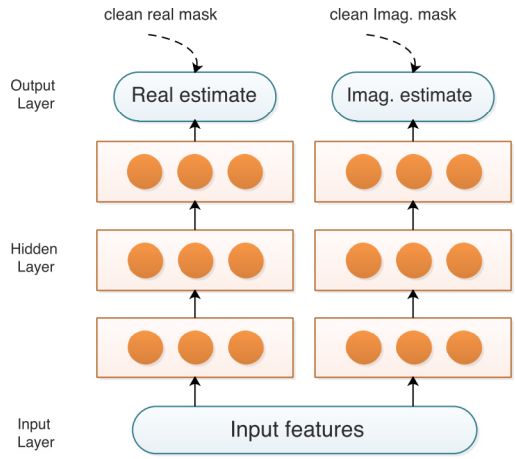

Fig. 2. The Y-shaped Bi-LSTM RNNs for cIRM estimation

B. Utterance Permutation Invariant Training

We take the two-speaker source separation with utterance permutation invariant training (uPIT) based cIRM model as an example, and the model is shown in Figure 3, which demonstrates the training process to predict the cIRM real component $\widehat{cIRM}_r$. The behavior of uPIT is marked with the dashed rectangle in Figure 3. Assume that the mixture speech Y is made up of source S_1 and source S_2, and S denotes the number of speakers. As for the targets input1 $S_1\_cIRM'_r$ and input2 $S_2\_cIRM'_r$, they are obtained by using Equations (13) and (14), regarding as the labels of this supervised learning model. Note that, $S_1\_\widehat{cIRM}_r$ and $S_2\_\widehat{cIRM}_r$ are the elements in the output1 of Bi-LSTM RNN network, referring to the real component predictions of S_1 and S_2, respectively. We compute the mean-square error (MSE) between the Bi-LSTM RNN outputs in uPIT module and the compressed targets masks of the clean speech signal as the cost function. Hence, the cost function for real components of the pcIRM-based method can be defined as:

$$J_r^{\phi^*} = \frac{1}{B}\sum_{i=1}^{S}\Big[\sum_t\sum_f S_i\_\widehat{cIRM}_r - S_{\phi^*(i)}\_cIRM'_r\Big]^2 \quad (16)$$

where $B = T \times N \times S$ is the total number of T-F units over all sources, T is the total number of frames over all source utterances, N is the window length (or frame length), and ϕ^* is the permutation that minimizes the utterance-level separation error, which can be defined as

$$\phi^* = \arg\min_{\phi\in\wp}\sum_{i=1}^{S}\Big[\sum_t\sum_f S_i\_\widehat{cIRM}_r - S_{\phi(i)}\_cIRM'_r\Big]^2 \quad (17)$$

Note that, $\wp$ in Equation (17) is the symmetric group of degree S, which is the set of all $S!$ permutations [15]. Similarly, the training process to predict the cIRM imaginary component $\widehat{cIRM}_c$ and the imaginary component cost function $J_c^{\phi^*}$ are both the same as the real component counterpart. Therefore, we can similarly define the cost functions of uPIT-based IRM model and the uPIT-based PSM model with Equations (18) and (19), respectively.

$$J_{IRM}^{\phi^*} = \frac{1}{B}\sum_{i=1}^{S}\Big[\sum_t\sum_f S_i\_\widehat{IRM} - S_{\phi^*(i)}\_IRM\Big]^2 \quad (18)$$

$$J_{PSM}^{\phi^*} = \frac{1}{B}\sum_{i=1}^{S}\Big[\sum_t\sum_f S_i\_\widehat{PSM} - S_{\phi^*(i)}\_PSM\Big]^2 \quad (19)$$

As for those masking-based methods without using uPIT, the order of targets source is fixed, and there is only one permutation for estimated speech and target speech pairs, whose cost functions have the same form with the uPIT ones.

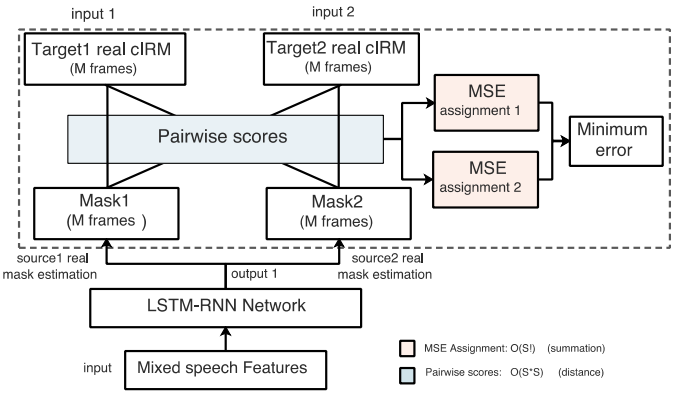

Fig. 3. uPIT-based cIRM estimate model

C. pcIRM Method

It is worthy noting that our proposed method pcIRM is inspired by the works [15, 20] and Bi-LSTM RNN, whose main idea is shown in Figure 4.

In the training stage, the STFTs of speech source and mixture are obtained in the feature extraction module. Then, the real and imaginary components of STFT of the speech source are used to calculate the compressed real mask $cIRM_r'$ and the compressed imaginary mask $cIRM_c'$ as the training targets for Bi-LSTM RNN1 and Bi-LSTM RNN2, respectively. During each iteration, the estimated T-F mask is optimized to minimize the MSE between the compressed targets masks of the clean speech signal and the Bi-LSTM RNN outputs in the uPIT module.

In the testing stage, the STFT of the mixed speech obtained in feature extraction is the input of the trained Bi-LSTM RNN1 and Bi-LSTM RNN2. In the compound module, we recover the output of these two networks by using Equation (15), which are the T-F real and imaginary masks of the estimated source speech, respectively. The real and imaginary components of the estimated signal are obtained by multiplying the estimated T-F real mask and the imaginary mask with the STFT of the mixture speech. Then, the separated speech signal is reconstructed in the reconstructed module.

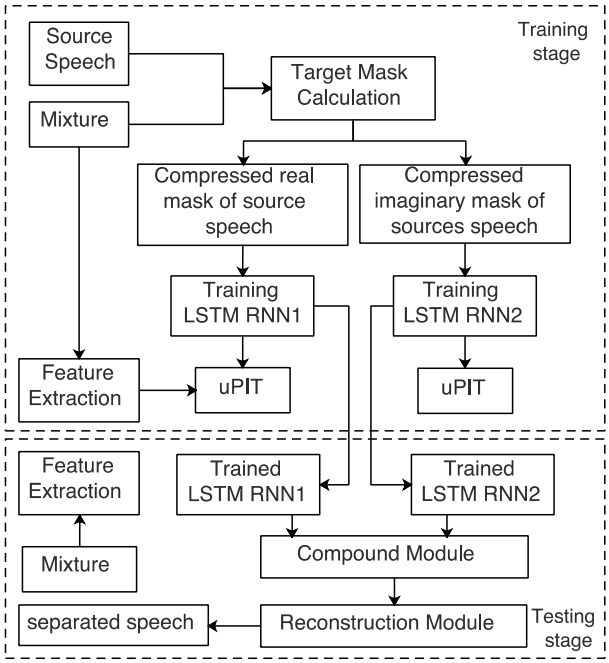

Fig. 4. The architecture diagram of the proposed pcIRM method with the example of two-talker source separation.

Compared with the IRM-based and PSM-based methods without using uPIT, the proposed pcIRM-based method has two following advantages:

1) A Y-shaped Bi-LSTM RNNs is exploited to predict the cIRM as the training target, the magnitude and phase information of the speech signal are be effectively utilized.
2) uPIT is integrated into the Bi-LSTM RNNs, elegantly solving the label permutation problem and speaker tracing problem in one shot.

IV. Experiments

We compare the pcIRM method with IRM-based method and PSM-based method using the uPIT or conventional training approach to show the advantage of the utterance permutation invariant training. Moreover, we also evaluate the cIRM-based uPIT speech separation model that is implemented with the vanilla DNN or the Bi-LSTM RNN to validate the superiority of our proposals.

A. Datasets

We evaluate the proposed monaural source separation model on the WSJ0-2mix datasets, using 129-dimensional STFT complex spectral computed with a sampling frequency of 16KHz. The WSJ0-2mix datasets are derived from WSJ0 corpus [21], which are introduced in [14]. WSJ0 corpus is made up of a training set namely set si_tr_s and two validation sets, namely set si_dt_05 and set si_et_05. Note that, the set si_tr_s consists of 101 speakers and each speaking contains about 140 or 90 utterances with a duration of approximately 5 seconds.

As for WSJ0-2mix datasets, the 30h training set and the 10h validation set are both constructed by randomly selecting two speakers and utterances from the WSJ0 training set si_tr_s, which includes 49 males and 51 females, and then mixing them at a various SNRs ranging from 0 dB to 5 dB. The 5h testing set is generated using utterances of 18 speakers, including 7 females and 11 males, from the WSJ0 validation set si_dt_05 and set si_et_05, with the same construction method as the 30h training set. As these 18 speakers in the testing set are not included in the training set, we conduct our experiments in the speaker-independent situation.

B. Network Architecture

All the vanilla DNN-based methods evaluated have 3 hidden layers with 1792 units each, and all the Bi-directional LSTM RNN based methods have 3 hidden layers, each of which has 896 units, so that all models have similar number of parameters. To avoid the overfitting, all models contain random dropouts when fed from a lower layer to a higher layer with a dropout rate of 0.5. There are $|S|$ output streams for $|S|$-speaker mixed speech, and in our study $|S|$ is set to 2, which is the same as in most of the existing studies. Then, data is poured into a $|S| \times 1792$-unit linear layer and a $|S| \times 1792$-unit rectified linear unit (ReLU) layer successively, aiming at avoiding the gradient vanishing problem.

The input to the network is the stack (over multiple frames) of the 129-dimensional STFT spectrum of the mixture speech, with a frame length of 16ms and an 8ms shift. The input data is a three-dimensional tensor shaped as $(D \times T \times 129)$, and each dimension is the size of batch (or the number of utterances in a batch), the maximum of frames in a batch, the number of frequency bins, respectively. The output consists of $|S|$ output

masks/streams, and each output mask vector has a dimension of $T \times 129$.

As for the training target of the cIRM, the corresponding neural networks outputs are the real components estimation and the imaginary components estimation of the predicted cIRM. Two Bi-LSTM RNNs are trained separately with the MSE cost functions $J_r^{\phi^*}$ and $J_c^{\phi^*}$, respectively. Adam optimization algorithm [22] is used both in the DNN and the Bi-LSTM RNN models with a weight decay of 10^{-5}, while the learning rate varies. The training process is terminated when the learning rate gets below 10^{-10}. In addition, the batch size is 8, meaning that each minibatch load 8 utterances randomly selected from datasets. The number of the epoch is set to 100. Note that, in our study the training data is used to train the model, and the validation set is only used to control the learning rate.

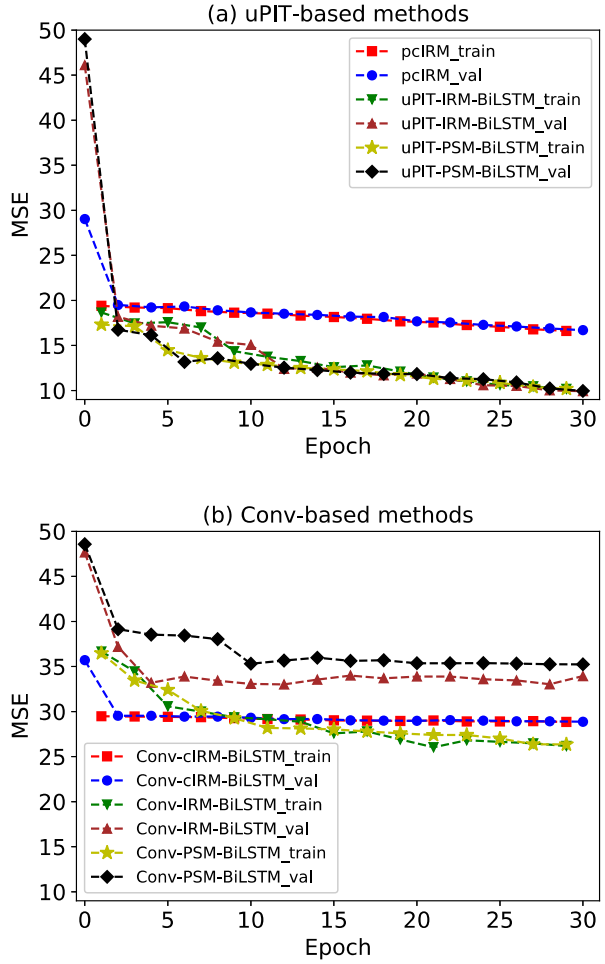

Fig. 5. The performances of the uPIT-based source separation methods (a) and the conventional training based source separation methods (b).

C. Training Behavior

In Figure 5, we compare the training progress of cIRM-based, IRM-based, PSM-based source separation methods with the uPIT approach and conventional training approach on the mixed speech datasets described in Section IV-A, measured by the MSE on the training and validation sets. Apparently, the MSE of the conventional training methods decreases slowly, whose remains almost unchanged since 10th epoch, showed in the subgraph (b), and that is probably a consequence of the permutation problem. In contrast, the MSE converges quickly when uPIT is used. The big gap between the MSE of the conventional training approach based cIRM and the MSE of pcIRM demonstrates the latter's effectiveness of the solving the label permutation problem discussed in [14].

In Figure 6, we display the training progress of vanilla DNN based and the pcIRM source separation methods with uPIT, on the same datasets as experiments in Figure 5. Note that, from Figure 6 we can see that the training MSE and validation MSE of these two methods decrease quickly and show almost the same trend, and the values of Bi-LSTM-based method are much smaller than the vanilla DNN ones, which indicates that pcIRM methods are more effective in processing the long-range context than other vanilla DNN-based methods.

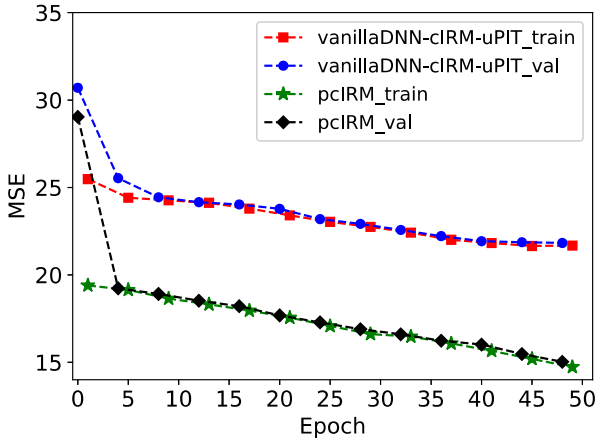

Fig. 6. Training progress of vanilla DNN based source separation method and the pcIRM-based source separation method.

D. Signal-to-Distortion Ratio Improvement

Generally, the separation performance is often evaluated with three measurements, including the short-time objective intelligibility (STOI), the perceptual evaluation of speech quality (PESQ), and the signal-to-distortion ratio (SDR) [23]. The STOI and the PESQ measure the intelligibility scores and human speech quality scores, respectively. Since the SDR is used to indicate the overall separation performance, we evaluate our proposed method on its potential to improve the SDR in this paper.

Table I summarizes the SDR improvement in dB from different separation methods for two-speaker mixed speech in the speech-independent situation, where bold indicates the best results. In our experiments, we reconstruct each frame by averaging over all output metaframes involved in the same frame. In the uPIT, it is assumed that the permutation of speakers keep constant across the utterance, which is not true in the real-world situation.

TABLE I
SDR IMPROVEMENTS(DB) FOR DIFFERENT SOURCE SEPARATION METHODS USING UPIT ON THE SPEECH-INDEPENDENT SITUATION

Index Model	*Opposite gender*	*Same gender*	*Average*
Conv-IRM-BiLSTM	6.51	5.73	6.13
Conv-PSM-BiLSTM	**7.19**	**6.30**	**6.75**
Conv-cIRM-BiLSTM	6.47	5.82	6.15
uPIT-IRM-BiLSTM	10.39	7.19	8.81
uPIT-PSM-BiLSTM	10.47	**7.69**	**9.10**
pcIRM	**10.48**	7.26	8.89
uPIT-cIRM-vanillaDNN	9.24	6.78	8.03
Oracle IRM	12.86	12.27	12.57
Oracle PSM	15.79	15.20	15.50
Oracle cIRM	**73.09**	**73.56**	**73.33**

From Table I, we can make several observations. Firstly, with the current experiment settings, the proposed pcIRM method achieves the best performance in the opposite-gender situation. At the same time, PSM-based Bi-LSTM RNN method almost obtains the maximum SDR improvement, with using conventional training methods or uPIT, which demonstrates the effectiveness of phase information in promoting the source separation results. Secondly, comparing with the SDR values of conventional training methods and uPIT-based methods, the latter show great advantages among different types of training targets. In addition, we can see that Bi-LSTM RNNs get better scores than vanilla DNNs in this task, which demonstrates great advantages of the Bi-LSTM RNN in capturing sequence information.

Moreover, Table I reports SDR (dB) improvements on test sets of WSJ0-2mix divided into opposite-gender and same-gender. From Table I, we can clearly see that our approach achieves much better SDR improvements on the opposite-gender mixed speech than the same-gender mixed speech, though the gender information is not explicitly used in our model and training procedure. With more training epochs, IBM-based and PSM-based methods would be more close to the oracle IRM and oracle PSM results for the opposite-gender condition. These results are consistent with breakdowns from other works [24] and generally indicate that same-gender mixed speech separation is a harder task.

Furthermore, it is worth discussing the problem that, the SDR improvements of oracle cIRM are over six times that of oracle IRM and are almost five times than of oracle PSM, whereas the performance of cIRM-based methods are inferior to PSM-based methods besides the instances of pcIRM's achievement in the opposite-gender situation. On one hand, as aforementioned, the values $cIRM_r \in \mathbb{R}$ and $cIRM_c \in \mathbb{R}$ are both unbounded, and we compress the cIRM with the hyperbolic tangent function as well as recover the estimation using Equation (15), which corrodes the accuracy of the pcIRM. On the other hand, depending on the difference of the real part and the imaginary part of the spectrum, we propose a Y-shaped Bi-LSTM RNNs, and the architecture of the Y-shaped neural network is depicted in Figure 2, where the output of the real component and the imaginary component are optimized individually. Figure 7 shows the total MSE loss, the real component MSE loss and the imaginary component MSE loss of pcIRM on the training and test stage in the first 30 epochs, respectively. The total MSE loss is the summation of the real component MSE loss and and the imaginary component MSE. Specifically, the values of loss for 3rd, 15th, and 29th epoch are marked in this line chart. The loss of the real part decreases continuously over the period, while the imaginary part is almost unchanged, whose reduction is tiny. In terms of these results, the Y-shaped neural network for the imaginary part makes a limited difference in optimizing the MSE loss. In theory, the real part can be understood as the projection of the spectrum, which is another expression of the PSM, but it is difficult to understand the imaginary part. Consequently, the model of the imaginary component is difficult to be well trained than the model of the real component. Compared with the results of cIRM-based methods in [13, 20], our proposed pcIRM method obtains slightly higher SDR improvement, displaying the effectiveness of the pcIRM method.

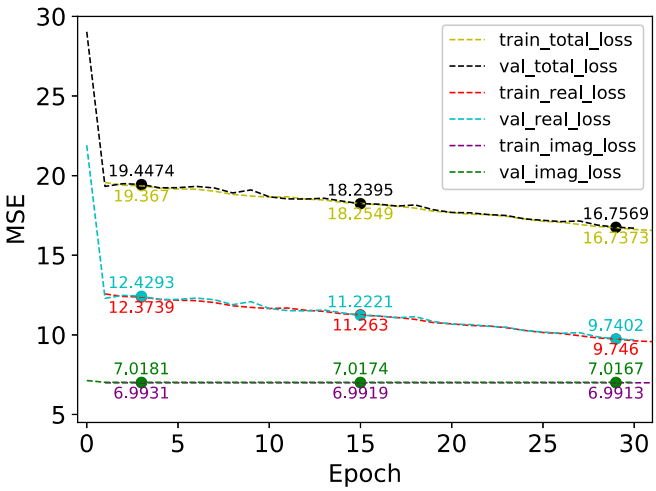

Fig. 7. The total MSE loss, the real component MSE loss and the imaginary component MSE loss of pcIRM method on the training and test stage in the first 30 epochs.

V. CONCLUSION AND FUTURE WORK

In this paper, the pcIRM method is proposed to address the speaker-independent monaural source separation problem. The proposed method achieves the cIRM estimation with the utterance-level permutation invariant training, and was implemented with a Y-shaped Bi-LSTM RNNs, where the output of real component and imaginary component are optimized individually. We report separation results for the proposed method and compare them to related systems with the WSJ0-2mix datasets. The experimental results show the importance of the phase information and the effectiveness of the uPIT method in the tasks of the source separation, in terms of the SDR metric.

To the best of our knowledge, this is the first study to integrate uPIT method and cIRM method as a whole model to address speech separation, there will likely be room for future

improvement. For example, effective features for such a task should be systematically examined and new features may need to be developed. Additionally, a more sophisticated network may need to be introduced for a more effective complex masking estimate.

REFERENCES

[1] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[2] A. Aroudi and S. Doclo, "Cognitive-driven binaural lcmv beamformer using eeg-based auditory attention decoding," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 406–410.

[3] F. Haider and S. Luz, "A system for real-time privacy preserving data collection for ambient assisted living," *Proc. Interspeech 2019*, pp. 2374–2375, 2019.

[4] Y. Sun, W. Rafique, J. A. Chambers, and S. M. Naqvi, "Underdetermined source separation using time-frequency masks and an adaptive combined gaussian-student's t probabilistic model," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4187–4191.

[5] S. Makino, T.-W. Lee, and H. Sawada, *Blind speech separation*. Springer, 2007, vol. 615.

[6] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*. Springer, 2005, pp. 181–197.

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

[8] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 1562–1566.

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

[10] Y. Wang, A. Narayanan, and D. 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.

[11] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 4, pp. 679–681, 1982.

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

[13] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, pp. 483–492, 2015.

[14] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 31–35.

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

[16] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[17] Y.-S. Lee, C.-Y. Wang, S.-F. Wang, J.-C. Wang, and C.-H. Wu, "Fully complex deep neural network for phase-incorporating monaural source separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 281–285.

[18] J. Chen and D. 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.

[19] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking," *arXiv preprint arXiv:1810.04826*, 2018.

[20] Y. Sun, Y. Xian, W. Wang, and S. M. Naqvi, "Monaural source separation in complex domain with long short-term memory neural network," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 359–369, 2019.

[21] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "Csr-i (wsj0) complete ldc93s6a," *Web Download. Philadelphia: Linguistic Data Consortium*, vol. 83, 1993.

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

[23] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[24] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," *arXiv preprint arXiv:1607.02173*, 2016.