Complex Neural Spatial Filter: Enhancing Multi-Channel Target Speech Separation in Complex Domain

Rongzhi Gu, Shi-Xiong Zhang, Yuexian Zou, Senior Member, IEEE, and Dong Yu

Abstract—To date, mainstream target speech separation (TSS) approaches are formulated to estimate the complex ratio mask (cRM) of target speech in time-frequency domain under supervised deep learning framework. However, the existing methods are designed in the way that the real and imaginary parts of the cRM are separately modeled using real-valued training data pairs. The research motivation of this study is to design a deep model that fully exploits the temporal-spectral-spatial information of multi-channel signals for estimating cRM directly and efficiently in complex domain. As a result, a novel TSS network is designed consisting of two modules, a complex neural spatial filter (cNSF) and an MVDR. Essentially, cNSF is a cRM estimation model and an MVDR module is cascaded to the cNSF module to reduce the nonlinear speech distortions introduced by neural network. Specifically, to fit the cRM target, all input features of cNSF are reformulated into complex-valued representations. Then, to achieve good hierarchical feature abstraction, a complex deep neural network (cDNN) is delicately designed with U-Net structure. Experiments conducted on simulated multi-channel speech data demonstrate the proposed cNSF outperforms the baseline NSF by 12.1% scale-invariant signal-to-distortion ratio and 33.1% word error rate.

Index Terms—Target speech separation, complex neural networks, complex ratio mask estimation, MVDR.

I. INTRODUCTION

TARGET speech separation (TSS) aims to recover the target speech that is corrupted by interfering speech, noise and reverberation. With the entry into deep learning era, significant improvements have been witnessed in TSS performance [4]–[7]. Based on the time-frequency (T-F) sparsity assumption of speech signals [8], mainstream deep learning based TSS methods formulate TSS task as a supervised learning problem in T-F domain [3]. In the early study, the learning objective was to estimate a magnitude mask of the target speech from the mixture magnitude, leading to ratio mask (RM) based TSS methods [9], [10]. Essentially, these RM based methods work in the magnitude domain and ignore the phase estimation. Later research showed that the reuse of the mixture phase to reconstruct the target speech becomes a limiting factor to the performance [11]. To tackle this problem, phase-sensitive mask (PSM) and complex ratio mask (cRM) were further proposed to consider the phase information of multi-channel mixture signals, respectively. cRM is defined as the division of the complex spectrogram of clean speech and its mixture [12]:

\[ M = \frac{S}{Y} = \frac{|S| + j\angle(S)}{|Y|} \]

where \( M \) is cRM, \( S \) and \( Y \) denote the complex spectrogram of clean speech and its mixture, respectively. cRM based TSS methods have become the mainstream, where the real and imaginary parts of cRM are separately estimated. TSS with multiple microphones has drawn great attention recently, owing to the spatial information that can be exploited to further enhance the TSS performance. Previous investigations on interaural phase difference (IPD) and its derived directional features showed that these features can provide discriminative cues for cRM estimation [1], [2], [16]–[19]. Our previous work, spatial neural filter (NSF) [20], leveraged the direction information of the target speech and formulated two directional features to indicate the dominance of acoustic components from the target direction, which guides the deep neural network (DNN) model to learn better masks. Very recently, to reduce the nonlinear speech distortions brought by DNN model, based on the classic minimum variance distortionless response (MVDR) [21] beamforming technique, Xu et al. proposed to jointly train the DNN and MVDR beamformer to improve the robustness of MVDR coefficients estimation. Experimental results demonstrated that both the ASR accuracy and perceptual speech quality are improved.

There is also some exploring work related to cRM estimation for monaural speech enhancement (SE) task. [14] investigated to estimate the cRM using two-stream DNNs that process magnitude and phase respectively. Since the distribution of phase does not have an explicit structure as magnitude, to promote the learning of phase, a communication module was designed to facilitate the information flow between the magnitude and phase stream.
As discussed above, existing cRM based methods model the magnitude and phase (or real and imaginary) of the cRM separately with two sets of real-valued training pairs, which is sub-optimal. Triggered by the complex deep neural network (cDNN) models introduced recently to support phase-aware processing for SE task [23]–[25], this work proposes to directly estimate the cRM using cDNN for multi-channel TSS task. Specifically, we design a single-stream framework, which consists of a cRM estimation network (named as complex neural spatial filter, cNSF) and an MVDR module. cNSF works in complex domain, which is delicately designed to fully exploit the temporal-spectral-spatial information. Also, to suppress the nonlinear speech distortion introduced by DNN, an MVDR module is introduced into the TSS framework and jointly trained with cNSF. To verify our proposed cNSF-MVDR model, extensive experiments have been conducted with simulated multi-channel speech data. Experimental results demonstrate the proposed cNSF-MVDR outperforms the baseline NSF by 12.1% scale-invariant signal-to-distortion ratio (SI-SDR) and 33.1% word error rate (WER).

II. PROPOSED TSS FRAMEWORK

In this work, TSS aims to separate the target speech \( s \) from the \( U \)-channel mixture \( y \) with given direction \( \theta \) of the target speaker. We assume that the oracle direction of the target speaker is known by the separation system, via the camera or speaker localization frontends. Fig. 1 illustrates our proposed TSS framework, consisting three modules: 1) feature extraction (Section II-A), which leverages the temporal-spectral-spatial information in the multi-channel mixture signal to extract effective cues for TSS; 2) the backbone network (Section II-B) designed in the form of U-Net, including a complex encoder, complex BLSTM layers and a complex decoder; 3) target speech reconstruction (Section II-C). This module has two optional paths, one is the cRM based speech reconstruction (cNSF) and another is the cRM based MVDR beamforming (cNSF-MVDR).

A. Training Pair Formulation

Following the supervised learning paradigm, the training target is to estimate the cRM of the target speech at the reference channel (the first channel in this work) as (1). To fit the cRM target defined in complex domain, we reformulate all the input features into complex-valued representations. In order to fully exploit the temporal-spectral-spatial information, spectral, spatial and directional features are combined as input features, following our previous work neural spatial filter (NSF) [20]. To be specific, firstly, the reference channel of mixture spectrogram \( Y \in \mathbb{C}^{T \times F} \) is taken as the spectral feature, where \( T \) and \( F \) are the total number of frames and frequency bands, respectively. Then, since we assume a far-field scenario, IPD is chosen as the spatial feature considering that spatial information contained in interaural level difference (ILD) is relatively weak and noisy. IPD is computed as the phase difference between two channels of the complex spectrogram:

\[
\text{IPD}^{(p)}_{t,f} = \angle Y_{t,f}^{p1} - \angle Y_{t,f}^{p2}
\]

where \( t \) is the frame index, \( f \) denotes the frequency, \( Y \) is the complex spectrogram of multi-channel mixture, \( p_1 \) and \( p_2 \) are two microphones of the \( p \)-th microphone pair, \( P \) is the number of selected microphone pairs. In the ideal anechoic environment, IPD can be approximated as a linear function of frequency with a certain time delay [8]. Since T-F bins that dominated by the same directional source experience the same time delay, IPDs of these bins will form a cluster within each frequency band. This property enables IPDs to indicate the T-F bin groups that dominated by the same source. However, as Wang et al. [19] pointed out, due to the phase wrapping issue, IPDs are not continuous across frequencies, which brings difficulty for the model to learn the correlation across frequency. Moreover, the linear functions that belong to different sources may cross because of spatial aliasing. To alleviate the discontinuity and crossing problem, we regard IPD as a complex-valued representation, of which the real and imaginary components are cosIPD and sinIPD, respectively. The cosIPD along with sinIPD formulates a 3 d helix-like function [19], therefore reducing the crossing in 3 d space while promoting continuity.

Finally, to leverage the direction information we design a directional feature (DF). The formulation of DF aims to indicate the target speech dominance against spatially distributed interferences at each T-F bin, which is a discriminative feature for TSS [1], [20]. Specifically, define target-dependent phase difference \( TPD_f^{(p)}(\theta) = 2\pi f TPD_f^{(p)}(\theta) \) [29] as the phase delay for the plane wave from \( \theta \), with frequency \( f \), to traverse between
the $p$-th pair of microphones, where $\tau^{(p)}(\theta)$ is the corresponding pure delay. The design principle of DF is that, if a T-F bin is dominated by the target speech (from $\theta$), the modulus of Hermitian inner product between IPD$^{(p)}$ and $\theta$ related TPD$^{(p)}$ of this T-F bin will be close to 1, otherwise it will be 0. Following this concept, DF is computed as follows:

$$d_{t,f}(\theta) = \sum_{p=1}^{P} \text{TPD}^{(p)}(\theta) \text{IPD}^{(p)}$$

Since both TPD and IPD are complex, their inner product, i.e., the computed DF, is also complex-valued. From the above, all input features can then be obtained by concatenating them along the feature axis, i.e., $[Y, \text{IPD}^{(1)}, \ldots, \text{IPD}^{(P)}, d(\theta)]^T$.

### B. Network Design

Borrowing design ideas from [30], to achieve hierarchical feature abstraction, our network adopts a U-Net structure [26], which is featured with an encoder-decoder structure and skip connections. The convolutional and deconvolutional layers (yellow and green blocks in Fig. 1) are used to capture and retrieve the local information, respectively. The BLSTM layers are used to enhance the temporal sequence information modeling. Also, the densely connected layers (blue blocks) support efficient multi-level feature re-use [27].

In Fig. 1, the complex encoder is composed of alternating complex 2D convolutional (conv2d) layers and complex dense blocks. The complex convolution is defined between the complex input $X = X_r + jX_i$ and the complex kernel $W = A + jB$. As illustrated in Fig. 2(b), the real and imaginary parts of convolution operation can be written in matrix notation:

$$
\begin{bmatrix}
(W \otimes X)_r \\
(W \otimes X)_i
\end{bmatrix} =
\begin{bmatrix}
A & -B \\
B & A
\end{bmatrix}
\otimes
\begin{bmatrix}
X_r \\
X_i
\end{bmatrix}
$$

where $\otimes$ denotes the real-valued convolution operation. Each complex conv2d layer is specified in the format: $\text{height(time)} \times \text{width(freq)} @ \text{featureMaps}$, and the stride and padding is set as (1, 2) and (1, 0) for all layers. Each complex conv2d layer is followed by complex ReLU activation function along with the complex batch normalization [22]. Complex ReLU applies ReLUs on the real and the imaginary parts of a neuron separately. Complex batch normalization can be viewed as whitening 2-dimensional vectors, which makes the whitened complex input subject to the standard complex distribution with zero mean and unit covariance [22]. Each dense block $\text{featureMaps}$ contains 5 complex conv2d layers that are densely connected, and the stride and padding are set as (1, 1) and (1, 1). The complex BLSTM layer is implemented according to [24]. The complex decoder is composed of alternating complex 2D deconvolutional layers and complex dense blocks. At the last deconvolutional layer, the number of output channel is set as 1 to predict the cRM of the target speech.

### C. Target Speech Reconstruction

To reconstruct the target speech spectrogram based on the predicted cRM, we explore two ways. The first one (dashed blue box in Fig. 1), named as cNSF, obtaining the estimated target speech spectrogram $\hat{S}$ by directly multiplying the estimated complex mask $\hat{M}$ to $Y$:

$$\hat{S} = \hat{M} \cdot Y = (Y_r \hat{M}_r - Y_i \hat{M}_i) + j(Y_r \hat{M}_i + Y_i \hat{M}_r)$$

The iSTFT operation implemented as a complex 1D deconvolution layer then converts the target complex spectrogram $\hat{S}$ to waveform $\hat{s}$. To reduce nonlinear speech distortions while further attenuating the interferences from the undesired direction, the cNSF-MVDR is developed (dashed red box in Fig. 1). MVDR is a classic and effective spatial beamformer proposed in [31], optimized with a constraint that minimizes the noise level without distorting the target speech. In our work, we follow the MVDR computation method in [32], and the coefficients $w \in \mathbb{C}^U \times F$ can be computed as follows:

$$w(f) = \frac{\Phi^{-1}_{\text{nn}}(f)\Phi_{ss}(f)}{\text{Tr}(\Phi^{-1}_{\text{nn}}(f)\Phi_{ss}(f))} u$$

where $u \in \mathbb{R}^U$ is a one-hot vector that marks the reference channel, Tr(·) solves the trace of a matrix. $\Phi_{ss}(f)$ and $\Phi_{nn}(f)$ are the spatial correlation matrices (SCM) of target speech and remaining noise, respectively. Motivated by the joint training of DNN and MVDR beamformer [21], an MVDR neural beamformer is designed with the estimated cRM. Given the $M$, the SCM of the target speech can be computed as:

$$\Phi_{ss}(f) = \frac{\sum_{t=0}^{T-1} (\hat{M}(t,f)Y(t,f))(\hat{M}(t,f)Y(t,f))^H}{\sum_{t=0}^{T-1} (\hat{M}H(t,f)\hat{M}(t,f))}$$

where $\hat{M}$ is shared across all the signal channels, $H$ denotes the conjugate transpose matrix. In the same way, to compute $\Phi_{nn}(f)$, the cRM for remaining noise needs to be estimated as $\hat{M}_n$. Finally, the beamformed signal can be obtained by:

$$\hat{s}(t) = w^H(t)Y(t,f)$$

### D. Loss Function

In order to optimize the whole network in the end-to-end manner, we adopt the speech separation metric SI-SDR [33]. It aims to reduce the logarithm estimation error between estimated target speech $\hat{s}$ and ground truth $s$ in time domain, which is defined as $\text{SI-SDR} := 10 \log_{10}(||s_{\text{target}}||^2/||e_{\text{noise}}||^2)$, where $s_{\text{target}} := ((\hat{s}, s)/|\hat{s}|^2, e_{\text{noise}} := \hat{s} - s_{\text{target}}$.

### III. EXPERIMENTS AND RESULTS

#### A. Dataset

To evaluate the proposed TSS framework under a challenging scenario, we simulate a multi-channel noisy reverberant dataset using the simulation pipelines described in [29]. The original speech data is collected from Youtube [34], in which Mandarin accounts for the majority. The sampling rate is 16 kHz. The
simulated dataset contains 192,000, 3000 and 5000 mixtures for training, validation and testing, respectively. The speakers appear in the test set has no overlap with the training set. The average duration of the testing utterance is about 4.1 seconds. We use a 15-element non-uniform linear array, with spacing 6-5-4-3-2-1-1-2-3-4-5-6 cm. The multi-channel speech signals are generated by convolving single-channel signals with room impulse responses (RIRs) simulated by image-source method [35]. The room size is ranging from 4m-4m-2.5 m to 10m-8m-6 m (length-width-height). The reverberation time T60 is sampled in a range of 0.05 s to 0.7 s. The signal-to-interference ratio is ranging from −6 to 6 dB. Also, different types of noise are added with 18–30 dB signal-to-noise ratio.

### B. Training and Evaluation Configuration

For STFT setting, we use 32 ms square-root Hann window with 16 ms hop size. IPDs and TPDs are extracted between 9 microphone pairs (1,15), (2,14), (3,13), (1,7), (12,4), (11,5), (12,8), (7,10), and (8,9) to sample different microphone spacings. Real-valued input features are the concatenation of $[Y]$, cos IPD and $|d(\theta)|$, while complex-valued input features are as described in Section II-A.

The detailed hyper-parameters for cNSF is illustrated in Fig. 1 with 17.5 M parameters. As for NSF, to keep the same parameters with cNSF for fair comparison, the numbers of all the convolution channels and BLSTM cells are doubled. To evaluate the effectiveness of the proposed network structure (densely connected U-Net, DUNet), we also train a BLSTM based network for comparison [20], which includes four 256-cell BLSTM layers followed by two feedforward layers. Both the DUNet and BLSTM based network are trained with 4-second mixture chunks, using Ranger optimizer [36] with early stopping. The learning rate is initialized as 1e-3 and will be decayed by 0.5 when the validation loss has no improvement for consecutive 3 epochs.

The performance is evaluated under different speaker mixing conditions: 1 speaker, 2 speakers and 3 speakers, respectively accounts for 5%, 45% and 50% in the test set. SI-SDR and WER are adopted as the evaluation metrics. The WER is measured using Tencent commercial Mandarin speech recognition API [37].

### C. Results Analysis

Table I reports the results of the multi-channel TSS systems. Firstly, it is obvious that DUNet based models exhibit consistent improvements over BLSTM based ones, owing to the effective multi-level feature abstraction ability of U-Net. Then, we examine the effects of training pair formulation. Beginning with the real-valued input features with RM target, DUNet based NSF obtains the SI-SDR of 10.6 dB and WER of 26.0%. Alternating the output target from RM to cRM, the SI-SDR has a gain of 0.1 dB and WER reduces 0.5%. Then, forming both input features and output target in complex domain using a real-valued DNN, the model further obtains a 0.4 dB SI-SDR improvement and 3.0% WER reduction. Finally, the proposed cNSF outperforms NSF by a large margin of 1.0 dB SI-SDR and 4.4% absolute WER.

When an MVDR module is used to reconstruct the target speech, performances of both NSF-MVDR and cNSF-MVDR improve over the NSF and cNSF, especially on the WER metric. This is because the target speech is comparatively preserved well, thanks to the distortionless constraint of MVDR formulation. At last, cNSF-MVDR exhibits the best performance with SI-SDR of 12.0 dB and WER of 17.0%.

To further examine the accuracy of the estimated cRM, Fig. 3 illustrates the distributions of cRMs on a two-speaker mixture. From the ground truth cRM plot (a), it is observed that most of the values are gathered around 0 and 1, while cRM estimated by NSF (b) misses the values around 0. Also, the red circle marks some complex values that form pairs of complex conjugates with corresponding ground truth cRM values, which leads to a right magnitude yet completely wrong phase. The cRM estimated by cNSF (c) shows a similar yet more compact distribution pattern with the ground truth, which confirms the accuracy of phase estimation and effectiveness of phase-aware processing with the proposed cDNN.

### IV. Conclusion

In this work, we propose a novel multi-channel TSS framework, which directly models the cRM estimation in complex domain, and elaborately designed to fully exploit the temporal-spectral-spatial information. Experiments show that the target speech estimated by the proposed framework achieves better quality and intelligibility over the baseline.
REFERENCES


[34] D. Yu and D. Wang, “Dcrr: A large-scale audio-visual corpus for multimodal speaker diarization, speech separation and recognition,” unpublished.

