2023 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (RISP NCSP 2023), Feb. 28 - Mar. 3, 2023

Deep Complex-Valued Neural Network-Based Triple-Path Mask and Steering Vector Estimation for Multi-channel Target Speech Separation

Mohan Qin[1], Li Li[2], and Shoji Makino[1]

[1]Waseda University
2-7 Hibikino, Wakamatsu-ku, Kitakyushu, Fukuoka, Japan
E-mail: {q.mohan@asagi., s.makino@}waseda.jp

[2]NTT Communication Science Laboratories
Nippon Telegraph and Telephone Corporation, Japan
E-mail: lili-0805@ieee.org

Abstract

This paper proposes a deep complex-valued neural network-based beamforming framework for multi-channel target speech separation. The deep complex-valued neural network predicts steering vectors and complex ratio masks for speaker signals. The masked signals are then used to calculate the spatial covariance matrices that are needed for conducting Minimum Variance Distortionless Response (MVDR) beamforming filter. We propose a Triple-path modeling for mask estimation, which takes both intra-channel and inter-channel features into consideration. Our experimental results revealed that the proposed framework achieved better target speech separation performance than the baseline methods.

1. Introduction

In practical applications, the accuracy of the acoustic signal recognition (ASR) system is often affected by interference and reverberation. Therefore, target speech separation and speech enhancement algorithms are proposed to extract the desired target signal from the noisy mixture signal.

Recently, deep neural network (DNN)-based algorithms [1, 2] have achieved remarkable performance on target speech separation task. However, purely deep learning algorithms often import distortion when recovering the target signal due to the non-linear processing such as using the non-linear activation function. This distortion lowers the accuracy of the back-end ASR system. On the other hand, the beamforming algorithms [3] almost do not import distortion by using a linear filter to recover the target speech. However, beamforming algorithms often need prior knowledge of the target source. In this case, whether it is possible to combine the advantages of the two algorithms and benefit the ASR systems draws a lot of research interest.

DNN-based beamforming algorithms [4, 5, 6] have been proposed recently and proven to reconstruct the high-quality target speech from the mixture effectively. In these algorithms, a DNN is used to predict masks for the target and the interference speech, which are used to estimate the speech components for calculating the beamforming filter. However, these networks take the magnitude spectra or just simply concatenate the real the imaginary parts of the complex-valued spectra as the input since the real-valued DNN cannot handle the complex-valued input directly. That means the phase information, which is also essential for improving speech quality, is not fully utilized.

Phase estimation for acoustic signal reconstruction has been extensively studied due to its importance. One mainstream method [7] is to model magnitude and phase by defining two sub-networks and exchanging information between the two sub-networks during the inference. The result in [7] has demonstrated that the information exchange between magnitude and phase effective in improving the mask prediction accuracy. As another mainstream, the deep complex-valued neural network-based speech enhancement algorithms [8, 9] have been proposed and shown to achieve outstanding results on speech enhancement tasks. The complex-valued neural network extract features from both the real and imaginary parts. Therefore, the magnitude and phase of the signal can be reasonably estimated. These researches motivate us to use complex-valued neural networks to benefit the DNN-based beamforming and improve the speech enhancement performance.

For multi-channel speech tasks, spatial information should be utilized as the key for processing since it contains the inter-channel features. However, [4] processes each channel separately, which means the information of other channels is not utilized when estimating the masks for one specific channel.

To solve the above problems, we propose to use a complex-valued Triple-path mask estimation model which fully utilizes the phase and the spatial information. We evaluate our proposed method on the target speech separation task. The results show that proposed method effectively reduces interference noise and improves the speech evaluation metrics.

2. Problem formulation

In this research, we assume that the target source $s(t, f)$ is captured by M microphones. The observed signal $\boldsymbol{y}(t, f) = [y_1(t, f), ..., y_M(t, f)]^T \in \mathbb{C}^M$ at time frame t and frequency bin f can be modeled in the short-time Fourier transformation (STFT) domain as:

$$\boldsymbol{y}(t, f) = \boldsymbol{a}(f)s(t, f) + \boldsymbol{n}(t, f), \quad (1)$$

where $\boldsymbol{a}(f)$ denotes the transfer function from the target

source to microphones, which is also known as the steering vector. $\boldsymbol{n}(t,f) = [n_1(t,f), ..., n_M(t,f)]^T \in \mathbb{C}^M$ denotes the interference noise. The objective of the target speech separation is to find a filter $\boldsymbol{w}(f)$, which can separate the target speech $\hat{s}(t,f)$ from the observed mixture speech by

$$\hat{s}(t,f) = \boldsymbol{w}^H(f)\boldsymbol{y}(t,f), \tag{2}$$

where ^{H} denotes the Hermitian transpose.

3. Baseline method: DNN Beamformer

This paper takes the DNN-based MVDR Beamforming introduced in [4] as our baseline method. Hereafter, we refer to this method as "DNN Beamformer" and briefly introduce it in this section.

MVDR beamforming algorithm is widely used for multi-channel speech enhancement tasks. It extracts the target speech from the mixed signal by minimizing the power of the noise under the constraint of keeping the target speech distortionless, which can be formulated as:

$$\textbf{min } \boldsymbol{w}^H(f)\boldsymbol{\phi}_{nn}(f)\boldsymbol{w}(f) \qquad \textbf{s.t. } \boldsymbol{w}(f)^H\boldsymbol{a}(f) = 1, \tag{3}$$

where $\boldsymbol{\phi}_{nn}(f)$ denotes the spatial covariance matrix (SCM) of the noise. An optimal solution can be obtained by the Lagrange multiplier method, which is expressed as:

$$\boldsymbol{w}(f) = \frac{\boldsymbol{\phi}_{nn}^{-1}(f)\boldsymbol{a}(f)}{\boldsymbol{a}(f)^H\boldsymbol{\phi}_{nn}^{-1}(f)\boldsymbol{a}(f)}. \tag{4}$$

However, the $\boldsymbol{a}(f)$ and $\boldsymbol{\phi}_{nn}(f)$ are difficult to estimate without any prior knowledge. Therefore, the DNN Beamformer was proposed to obtain this knowledge with the ability of the DNN. The processing flow of the DNN Beamformer is shown in Fig. 1. A DNN is used to estimate masks for the target speech $\boldsymbol{m}^s(t,f) = [m_1^s(t,f), ..., m_M^s(t,f)]^T$ and the noise $\boldsymbol{m}^n(t,f) = [m_1^n(t,f), ..., m_M^n(t,f)]^T$. Then the masked signal is used to calculate the target SCM $\boldsymbol{\phi}_{ss}(f)$ by

$$\begin{aligned} \hat{\boldsymbol{s}}(t,f) &= \boldsymbol{m}^s(t,f) \cdot \boldsymbol{y}(t,f), \\ \boldsymbol{\phi}_{ss}(f) &= \sum_{t=1}^{T} \hat{\boldsymbol{s}}(t,f)\hat{\boldsymbol{s}}(t,f)^H, \end{aligned} \tag{5}$$

where $\cdot$ denotes element-wise multiplication. The steering vector is then obtained as the eigenvector corresponding to the maximum eigenvalue of the $\boldsymbol{\phi}_{ss}(f)$, which can be expressed as

$$\boldsymbol{a}(f) = \mathrm{EVD}(\boldsymbol{\phi}_{ss}(f)). \tag{6}$$

Here, EVD denotes the eigenvalue decomposition. The estimated steering vector is used to calculate the filter $\boldsymbol{w}(f)$, which recovers the target speech from the mixture.

One issue of the DNN Beamformer is that the spatial information is not utilized when mask estimation. It only concatenates the spectrum of all channels in the time dimension and processes it as a single-channel mask estimation problem.

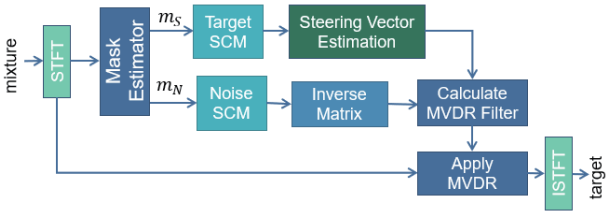

Figure 1: Processing flow of the DNN Beamformer

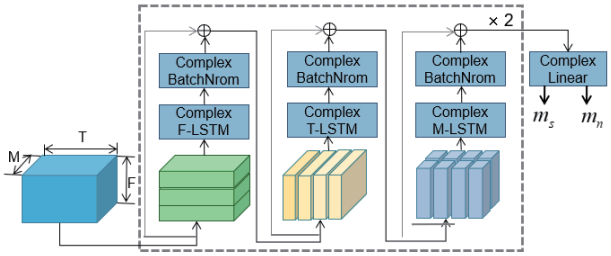

Figure 2: Triple-path mask estimation

Another limitation is the usage of real-valued network, which can only handle the real-valued input. Specifically, the bidirectional long short-term memory (BLSTM) network used in DNN Beamformer only takes the magnitude spectra of one channel as the input in each step.

4. Proposed method

4.1 Triple-path mask estimation

To deal with the drawback of the DNN Beamformer that do not utilize the spatial information when mask estimation, inspired by Dual-path Recurrent Neural Netwrok (RNN) [10], we propose a Triple-path modeling, whose composition is in Fig. 2. Each Triple-path block consists of a F-LSTM, T-LSTM and M-LSTM. The F-LSTM and T-LSTM are used to extract the inter-channel features. Denote the multi-channel mixture after STFT by $\boldsymbol{y} \in \mathbb{C}^{M\times T\times F}$, where M, T, and F are the number of channels, time frames, and frequency bins, respectively. The F-LSTM takes the $z_1 \in \mathbb{C}^{M\times T}$ as the input for each step, which is the information of all channels at frequency bin f. Then the F-LSTM uses such spatial information to model the sequential relationships between frequency bins. The T-LSTM takes the $z_2 \in \mathbb{C}^{M\times F}$ as the input for each step, which is the information of all channels at time frame t. Then T-LSTM model the sequential relationships between time frames. M-LSTM is used to model the intra-channel features. It takes the complex-valued spectra of one channel $z_3 \in \mathbb{C}^F$ as the input in each step, which is similar to the baseline method. By doing so, both intra-channel and inter-channel correlations are reasonably modeled.

Each BLSTM network in the Triple-path block is complex-valued, which consists of two real-valued networks including $\mathrm{BLSTM}_{\mathrm{real}}$ and $\mathrm{BLSTM}_{\mathrm{imag}}$. For a complex-valued input

$X = X_{real} + jX_{imag}$, where j denotes the imaginary unit, the output of the complex-valued BLSTM network F_{out} can be obtained by

$$\begin{aligned} F_{out} = &(\text{BLSTM}_{\text{real}}(X_{real}) - \text{BLSTM}_{\text{imag}}(X_{imag})) \\ &+j(\text{BLSTM}_{\text{real}}(X_{imag}) + \text{BLSTM}_{\text{imag}}(X_{real})). \end{aligned} \tag{7}$$

Such processing enables the network to extract information from both the real part and the imaginary part.

The estimated multi-channel complex ratio mask (CRM) $\boldsymbol{m} = \boldsymbol{m}_{real} + j\boldsymbol{m}_{imag} \in \mathbb{C}^{M\times T\times F}$ can be expressed in polar coordinates as

$$\begin{cases} \boldsymbol{m}_{mag} = \sqrt{\boldsymbol{m}_{real}^2 + \boldsymbol{m}_{imag}^2}, \\ \boldsymbol{m}_{phase} = arctan(\dfrac{\boldsymbol{m}_{imag}}{\boldsymbol{m}_{real}}) . \end{cases} \tag{8}$$

The observed signal can be similarly expressed in polar coordinates. The estimated target speech $\hat{\boldsymbol{s}}$ is obtained as

$$\hat{\boldsymbol{s}} = \boldsymbol{y}_{mag} \cdot \boldsymbol{m}^s{}_{mag} \cdot e^{\boldsymbol{y}_{phase} + \boldsymbol{m}^s{}_{phase}} . \tag{9}$$

4.2 Deep complex-valued convolutional recurrent network for steering vector estimation

In the conventional MVDR algorithm, the steering vector is obatined as the eigenvector corresponding to the maximum eigenvalue of the target speech SCM. However, the eigenvalue decomposition involved in the MVDR is sometimes numerically unstable when jointly trained with DNN and may not lead to the optimal solution. Therefore, All Deep Learning MVDR (ADL-MVDR) [5] proposed to replace the eigenvalue decomposition with a real-valued GRU-Net, which concatenates the real and the imaginary parts of the target speech SCM as the input and predicts the steering vector. However, ADL-MVDR can not fully explore the potential relationship between the real and the imaginary parts because of the lack of information interaction between them.

To address this limitation, we propose using complex-valued convolutional recurrent network to predict the steering vector. We refer to the structure of DCCRN [8] to design our model, which is shown in Fig. 3. The complex-valued encoder extracts high-dimensional features from the target SCM. Complex-valued BLSTM performs sequence modeling between frequency bins in high-dimensional feature space. Decoder reduces the dimension of features to the original shape. Skip connections are used to prevent information loss and unstable gradient update. The network output is fed into a complex-valued linear layer to obtain the steering vector.

Each block in the encoder or the decoder consists of a complex-valued convolutional layer, a complex-valued batch normalization, and a complex-valued PReLU. The complex-valued convolution operation $\circledast$ is defined as:

$$\begin{aligned} X \circledast W = &(X_{real} * W_{real} - X_{imag} * W_{imag}) \\ &+j(X_{real} * W_{imag} + X_{imag} * W_{real}), \end{aligned} \tag{10}$$

where * denotes the real-valued convolution operation, $W = W_{real} + jW_{imag}$ is the complex-valued convolutional kernel.

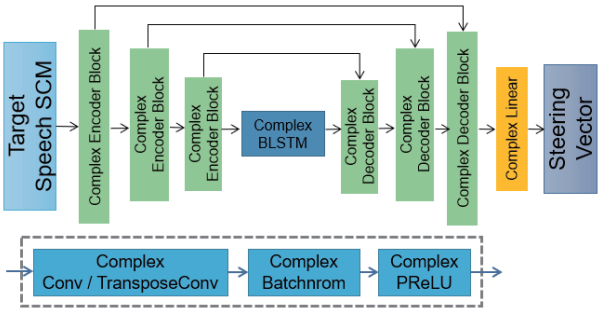

Figure 3: Steering Vector Estimation Network

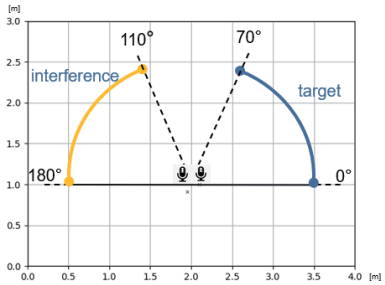

Figure 4: Layout of the experimental environment

5. Experiment

5.1 Dataset

The experimental dataset was simulated by image method using the Pyroomacoustics library. There were 5,000, 1,000 and 1,000 4-second two-speaker reverberant speech for training, validation and testing, respectively. The clean utterances of the source signals were selected form the LibriSpeech dataset [11]. The layout of the experimental environment is shown in Fig. 4. The position of the microphone array was fixed. The distance between the two microphones was 4 cm. The distance between each speaker and the microphone array was 1.5 m. The positions of the target speaker and the interference speaker were randomly placed on the circle centered at the microphone array, from 0° to 70° and 110° to 180° with 1° increment, respectively. The reverberation time was 100 ms. All the speech signals were sampled at 16 kHz.

5.2 Experimental configurations

The STFT was computed using a Hanning window. The length and the shift of the window were set to 1,024 samples (64 ms) and 256 samples (16 ms), respectively.

Adam was used as the optimizer. The learning rate was initialized as $1e^{-3}$ and was halved when the loss stopped decreasing for 2 epochs. The training objective was to maximize the scale-invariant signal-to-noise ratio (SI-SNR).

SI-SNR and signal-to-distortion ratio (SDR) were used as the evaluation metrics.

Table 1: SI-SNR [dB] and SDR [dB] scores on the testset.

Method	SI-SNR	SDR
Single-IRM	10.76	12.58
Single-CRM (proposed)	12.14	13.96
oracle Sig-MVDR	13.69	16.35
Triple-path CRM (proposed)	**15.93**	**17.30**

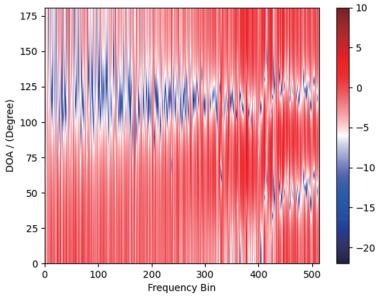

Figure 5: Example of the beam pattern obtained by the proposed Triple-path CRM method, the target direction was 50°, the interference direction was 120°

5.3 Evaluated systems

We compared several mask estimation methods including Single-IRM, Single-CRM, oracle Sig-MVDR, and Triple-path CRM. Single-IRM and Single-CRM estimate the masks for each channel separately using the method in [4]. Differently, the Single-CRM took the complex-valued spectra of one channel as the input and predicted CRMs using a complex-valued BLSTM network. For a fair comparison, the Single-IRM and Single-CRM estimation networks were designed to have the same parameter configuration. Specifically, they both had three 1024-size hidden layers and one 320-size linear projection layer. The oracle Sig-MVDR calculated the SCMs directly from the reference clean speech.

The Triple-path mask estimator had two Triple-path blocks. Complex-valued BLSTM network in each block had two 1024-size hidden layers and a 320-size complex-valued linear layer. Both the encoder and the decoder in the steering vector estimator had five blocks. The output dimension of each block in the encoder was {32, 64, 128, 256, 256}. The output dimension of the decoder blocks was the same as the encoder in reversed order. The complex-valued BLSTM network in the steering vector estimator had two 1024-size hidden layers.

5.4 Experimental results

Table 1 shows the SI-SNR and SDR scores of the evaluated target speech separation systems. Comparing the results of the Single-IRM with the Single-CRM, we could find that the complex-valued neural network improved separation performance by utilizing the phase information when estimation. Comparing the result of the Triple-path CRM with the Single-CRM, it indicates that Triple-path estimation was effective for promoting speech evaluation metrics by fully utilizing spatial information. Moreover, we also found that the training loss converged faster when the spatial information was utilized.

Fig. 5 shows the beam pattern obtained from the separation filter of the proposed method. It shows that the model can preserve the signal in the target direction and suppress the noise in the interference direction. The result of the proposed method was also better than the oracle Sig-MVDR. Since the oracle Sig-MVDR used clean speech to calculate the SCM, it indicates that the Steering Vector Estimator improved the upper bound of the oracle MVDR algorithm.

6. Conclusions

This paper proposed a deep complex-valued neural network-based MVDR beamforming framework for target speech separation and demonstrated its effectiveness. The Triple-path mask estimator fully utilizes the spatial information when modeling the relationships between time frames and frequency bins. The complex-valued neural network models the magnitude and the phase information effectively at once. The results showed that our proposed Triple-path mask estimation method achieved higher performance than the baseline methods. The proposed steering vector estimator could predict the target speech direction correctly.

References

[1] R. Gu, *et al.*, "Neural Spatial Filter: Target Speaker Speech Separation Assisted with Directional Information," in *Proc. Interspeech*, 2019, pp. 4290-4294.

[2] R. Gu, *et al.*, "Multi-Modal Multi-Channel Target Speech Separation," *IEEE JSTSP*, 2020, pp. 530-541.

[3] J. Benesty, *et al.*, "On Microphone-Array Beamforming From a MIMO Acoustic Signal Processing Perspective," *IEEE Trans. ASLP*, 2007, pp. 1053-1065.

[4] J. Heymann, *et al.*, "Neural Network Based Spectral Mask Estimation for Acoustic beamforming," in *Proc. ICASSP*, 2016, pp. 196–200.

[5] Z. Zhang, *et al.*, "ADL-MVDR: All Deep Learning MVDR Beamformer for Target Speech Separation," in *Proc. ICASSP*, 2021, pp. 6089-6093.

[6] T. Ochiai, *et al.*, "Beam-TasNet: Time-domain Audio Separation Network Meets Frequency-domain Beamformer," in *Proc. ICASSP*, 2020, pp. 6384-6388.

[7] D. Yin, *et al.*, "PHASEN: A Phase-and-Harmonics-Aware Speech Enhancement Network," in *Proc. AAAI*, 2020, pp. 9458-9465.

[8] Y. Hu, *et al.*, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in *Proc. Interspeech*, 2020, pp. 2472-2476.

[9] Z. Q. Wang, *et al.*, "Complex Spectral Mapping for Single-and Multi-channel Speech Enhancement and Robust ASR," *IEEE/ACM Trans. ASLP*, 2020, pp. 1778-1787.

[10] Y. Luo, *et al.*, "Dual-Path RNN: Efficient Long Sequence Modeling for Time-Domain Single-Channel Speech Separation," in *Proc. ICASSP*, 2020, pp. 46-50.

[11] V. Panayotov, *et al.*, "Librispeech: An ASR Corpus Based on Public Domain Audio Books," in *Proc. ICASSP*, 2015, pp. 5206-5210.