Journal of Signal Processing, Vol.27, No.4, pp.87-91, July 2023

SELECTED PAPER AT NCSP'23

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

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

[1]Graduate School of Information, Production and Systems, Waseda University
2-7 Hibikino, Wakamatsu-ku, Kitakyushu, Fukuoka 808-0135, Japan
E-mail: {q.mohan@asagi., s.makino@}waseda.jp

[2]NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation
Atsugi 243-0198, Japan
E-mail: lili-0805@ieee.org

Abstract

We propose a deep complex-valued neural network-based beamforming framework for multichannel 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 needed for minimum variance distortionless response (MVDR) beamforming. We propose triple-path modeling for mask estimation, which takes both intrachannel and interchannel features into consideration. Our experimental results revealed that the proposed framework achieves better target speech separation performance than do 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 outstanding performance in target speech separation tasks. However, purely deep learning algorithms often import nonlinear distortion when recovering the target signal owing to nonlinear processing. On the other hand, the beamforming algorithms [3] import almost no distortion as a result of using a linear filter to recover the target speech. However, beamforming algorithms often need prior knowledge of the target source. In this case, whether we can combine the advantages of the two algorithms and benefit the ASR systems is drawing much research interest.

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

Phase reconstruction for the acoustic signal has been extensively studied because of 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 inference. As another mainstream method, the deep complex-valued neural network-based speech enhancement algorithm [8, 9] has been proposed and found to show outstanding results on speech enhancement tasks. The complex-valued neural network can handle the complex-valued spectrum directly and model the magnitude and phase information jointly. These research results motivate us to use complex-valued neural networks to benefit the DNN-based beamforming and improve speech enhancement performance.

For multichannel speech tasks, spatial information should be utilized as the cue for processing since it contains interchannel features. However, in [4], each channel was processed separately without utilizing spatial information.

To solve the above problems, we propose to use a complex-valued triple-path mask estimation model, which fully utilizes the phase and spatial information, and embed it into a complex-valued neural network-based beamforming framework. We evaluate our proposed method on the target speech separation task. The results show that the proposed method effectively 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 transform (STFT) domain as

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

where $\boldsymbol{a}(f) = [a_1(f), ..., a_M(f)]^T \in \mathbb{C}^M$ denotes the transfer function from the target source to microphones,

and 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 $\boldsymbol{w}(f)$ that 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 $(\cdot)^H$ denotes the Hermitian transpose.

3. Baseline Method: DNN Beamformer

We take DNN-based minimum variance distortionless response (MVDR) beamforming [4], which is shown in Fig. 1, as the baseline method, and it will be referred to as "DNN Beamformer" in the following sections.

The MVDR beamforming algorithm 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

$$\begin{aligned} \textbf{min} \;\; & \boldsymbol{w}^H(f)\boldsymbol{\Phi}_{nn}(f)\boldsymbol{w}(f) \\ \textbf{s.t.} \;\; & \boldsymbol{w}(f)^H\boldsymbol{a}(f) = 1 \end{aligned} \tag{3}$$

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, $\boldsymbol{a}(f)$ and $\boldsymbol{\Phi}_{nn}(f)$ are difficult to estimate without any prior knowledge. Therefore, 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

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

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

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

EVD denotes the eigenvalue decomposition.

The limitation of the DNN Beamformer is that the phase and spatial information is not fully utilized during mask estimation. Specifically, the bidirectional long short-term memory (BLSTM) network used in the DNN Beamformer only concatenates the spectrum of all channels in time dimension, and takes the magnitude spectra of one channel as the input in each step.

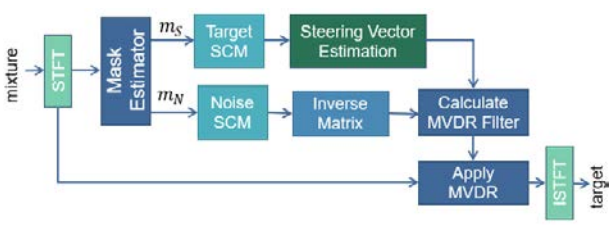

Figure 1: Process flow of DNN beamformer

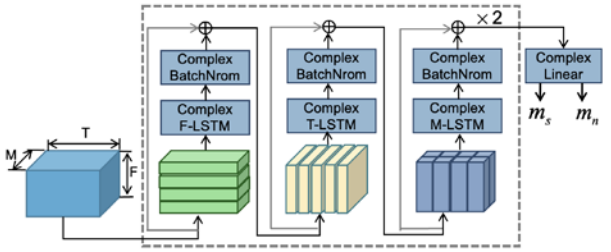

Figure 2: Triple-path mask estimation

4. Proposed Method

4.1 Triple-path mask estimation

To deal with the drawback of the DNN Beamformer, which does not utilize the spatial information in mask estimation, inspired by the dual-path recurrent neural network (RNN) [10], we propose triple-path modeling, whose composition is shown in Fig. 2. Each triple-path block consists of three sub-BLSTM networks, which are trained simultaneously in one end-to-end system. The multichannel mixture after STFT is denoted as $\boldsymbol{Y} \in \mathbb{C}^{M\times T\times F}$, where M, T, and F are the numbers of channels, time frames, and frequency bins, respectively. F-LSTM takes $\boldsymbol{Z}_1 \in \mathbb{C}^{M\times T}$ as the input for each step, which is the information of all channels at frequency bin f. T-LSTM takes $\boldsymbol{Z}_2 \in \mathbb{C}^{M\times F}$ as the input for each step, which is the information of all channels at time frame t. F-LSTM and T-LSTM perform sequential modeling between frequency bins and time frames utilizing spatial information, respectively. M-LSTM extracts the intrachannel features. It takes the complex-valued spectra of one channel $\boldsymbol{z}_3 \in \mathbb{C}^F$ as the input in each step. A residual connection is used to add the input and output of each sub-LSTM network, and then to transfer this information to the next sub-network. Such operation not only prevents information loss and unstable gradient updates during propagation but also enables M-LSTM to use spatial information for intrachannel modeling.

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 $\boldsymbol{X} = \boldsymbol{X}_{real} + j\boldsymbol{X}_{imag}$, where j denotes the imaginary unit, the output of the complex-valued BLSTM network $\boldsymbol{F}_{out}$ can

be obtained by

$$\begin{aligned}\boldsymbol{F}_{out} = &(\mathrm{BLSTM}_{\mathrm{real}}(\boldsymbol{X}_{real}) - \mathrm{BLSTM}_{\mathrm{imag}}(\boldsymbol{X}_{imag})) \\ &+j(\mathrm{BLSTM}_{\mathrm{real}}(\boldsymbol{X}_{imag}) + \mathrm{BLSTM}_{\mathrm{imag}}(\boldsymbol{X}_{real}))\end{aligned} \tag{7}$$

Such processing enables the network to extract information from both the real and imaginary parts in the input sequence.

The estimated multichannel complex ratio mask (CRM) $\boldsymbol{m}(t,f) = \boldsymbol{m}_{real}(t,f) + j\boldsymbol{m}_{imag}(t,f) \in \mathbb{C}^M$ can be expressed in polar coordinates as

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

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

$$\tilde{\boldsymbol{s}}(t,f) = \boldsymbol{y}_{mag}(t,f) \cdot \boldsymbol{m}^s{}_{mag}(t,f) \cdot e^{\boldsymbol{y}_{phase}(t,f) + \boldsymbol{m}^s{}_{phase}(t,f)} \tag{9}$$

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

In the conventional MVDR algorithm, the steering vector is obtained 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 a DNN and may not lead to the optimal solution. Therefore, all deep learning MVDR (ADL-MVDR) [5] is proposed to replace the eigenvalue decomposition with a real-valued GRU-Net. It concatenates the real and imaginary parts of the target speech SCM as the input, and predicts the steering vector. However, ADL-MVDR cannot fully explore the potential relationship between the real and imaginary parts because of the lack of information interaction between them.

To address this limitation, we propose using a complex-valued convolutional recurrent neural network (CRN) to predict the steering vector. We refer to the structure of DC-CRN [8] to design the module, 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. The decoder reduces the dimension of features to the original shape. Skip connections are used to prevent information loss and unstable gradient updates. The network output is fed into a complex-valued linear layer to obtain the steering vector.

Each block in the encoder or 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

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

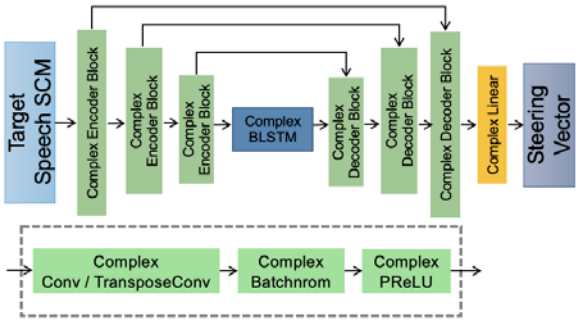

Figure 3: Steering vector estimation network

where * denotes the real-valued convolution operation and $\boldsymbol{W} = \boldsymbol{W}_{real} + j\boldsymbol{W}_{imag}$ is the complex-valued convolutional kernel.

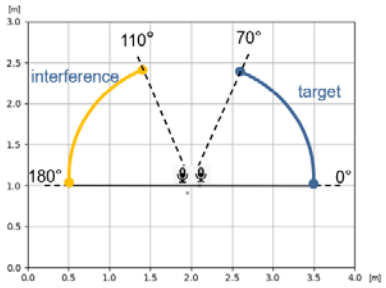

Figure 4: Layout of the experimental environment

5. Experiment

5.1 Dataset

The experimental dataset was simulated by an image method using the Pyroomacoustics library. There were 5,000, 1,000, and 1,000 4-second two-speaker mixtures for training, validation, and testing, respectively. The clean utterances of the source signals were selected from 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 target and interference speakers were randomly placed on the circle centered at the microphone array, from 0° to 70° and 110° to 180° with 1° increments, 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 shift of the window were set to 1,024 samples (64 ms) and 256 samples (16 ms), respectively.

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

Method	SI-SNR **train**	SI-SNR **test**	SDR
Single-IRM	13.22	10.76	12.58
Single-CRM	14.47	12.14	13.96
Oracle signal-MVDR	-	13.69	16.35
Triple-path CRM	**18.27**	**15.93**	**17.30**

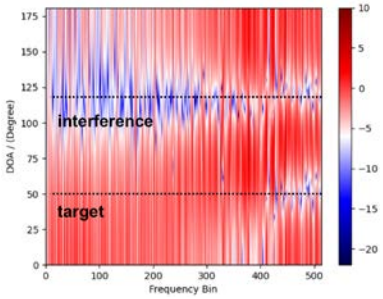

Figure 5: Example of the beam pattern obtained by the proposed triple-path CRM method, with target direction of 50° and interference direction of 120°

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

$$\begin{cases} \boldsymbol{s}_{target} = \dfrac{\langle \hat{\boldsymbol{s}}, \boldsymbol{s} \rangle \boldsymbol{s}}{\|\boldsymbol{s}\|^2} \\ \boldsymbol{e}_{noise} = \hat{\boldsymbol{s}} - \boldsymbol{s}_{target} \\ \text{SI-SNR} = 10 log_{10} \dfrac{\|\boldsymbol{s}_{target}\|^2}{\|\boldsymbol{e}_{noise}\|^2} \end{cases} \tag{11}$$

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

5.3 Evaluated systems

We compared several mask estimation methods, including single-IRM, single-CRM, oracle signal-MVDR, and triple-path CRM. Single-IRM and single-CRM estimate the masks for each channel separately by the method in [4]. In contrast, single-CRM takes 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. Oracle signal-MVDR calculated the SCMs directly from the reference clean speech.

The triple-path mask estimator had two triple-path blocks. The complex-valued BLSTM network in each block had two 1024-size hidden layers and one 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 symmetrical with that of the encoder blocks. 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 single-IRM and single-CRM, we found that the complex-valued neural network showed improved separation performance as a result of utilizing the phase information. The comparison of the results of triple-path CRM and single-CRM indicates that triple-path estimation was effective for promoting speech evaluation metrics because of fully utilizing spatial information. Moreover, we also found that the training loss converged faster when the spatial information was utilized.

Figure 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 that of oracle signal-MVDR. Since oracle signal-MVDR used clean speeches to calculate the SCM, the steering vector estimator improved the upper bound of the oracle MVDR algorithm.

6. Conclusions

We 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 spatial information when modeling the relationships between time frames and frequency bins. The complex-valued neural network models the magnitude and the phase information simultaneously. The results showed that our proposed triple-path mask estimation method achieved better performance than the baseline methods. The proposed steering vector estimator could predict the target speech direction correctly.

References

[1] R. Gu, L. Chen, S.X. Zhang, J. Zheng, Y. Xu, M. Yu, D. Su, Y. Zou and D. Yu: Neural spatial filter: Target speaker speech separation assisted with directional information, Interspeech, pp. 4290-4294, 2019.

[2] R. Gu, S.X. Zhang, Y. Xu, L. Chen, Y. Zou and D. Yu: Multi-modal multi-channel target speech separation, IEEE Journal of Selected Topics in Signal Processing, Vol. 14, No. 3, pp. 530-541, 2020.

[3] J. Benesty, J. Chen, Y. Huang and J. Dmochowski: On microphone-array beamforming from a mimo acoustic signal processing perspective, IEEE Trans. Audio, Speech, and Language Processing, Vol. 15, No. 3, pp. 1053-1065, 2007.

[4] J. Heymann, L. Drude and R. Haeb-Umbach: Neural network based spectral mask estimation for acoustic beamforming, 2016 IEEE Int. Conf. Acoust. Speech Signal Process., pp. 196-200, 2016.

[5] Z. Zhang, Y. Xu, M. Yu, S.X. Zhang, L. Chen and D. Yu: ADL-MVDR: All deep learning mvdr beamformer for target speech separation, 2021 IEEE Int. Conf. Acoust. Speech Signal Process., pp. 6089-6093, 2021.

[6] T. Ochiai, M. Delcroix, R. Ikeshita, K. Kinoshita, T. Nakatani and S. Araki: Beam-TasNet: Time-domain audio separation network meets frequency-domain beamformer, 2020 IEEE Int. Conf. Acoust. Speech Signal Process., pp. 6384-6388, 2020.

[7] D. Yin, C. Luo, Z. Xiong and W. Zeng: PHASEN: A phase-and-harmonics-aware speech enhancement network, Proceedings of the AAAI Conference on Artificial Intelligence, pp. 9458-9465, 2020.

[8] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang and L. Xie: DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement, Interspeech, pp. 2472-2476, 2020.

[9] Z.Q. Wang, P. Wang and D.L. Wang: Complex spectral mapping for single-and multi-channel speech enhancement and robust asr, IEEE/ACM Trans. Audio, Speech, and Language Processing, Vol. 28, pp. 1778-1787, 2020.

[10] Y. Luo, Z. Chen and T. Yoshioka: Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation, 2020 IEEE Int. Conf. Acoust. Speech Signal Process., pp. 46-50, 2020.

[11] V. Panayotov, G. Chen, D. Povey and S. Khudanpur: Librispeech: An asr corpus based on public domain audio books, 2015 IEEE Int. Conf. Acoust. Speech Signal Process., pp. 5206-5210, 2015.