Generalized Spatio-Temporal RNN Beamformer for Target Speech Separation

Yong Xu\textsuperscript{1}, Zhuohuang Zhang\textsuperscript{2*}, Meng Yu\textsuperscript{1}, Shi-Xiong Zhang\textsuperscript{1}, Dong Yu\textsuperscript{1}

\textsuperscript{1}Tencent AI Lab, Bellevue, USA
\textsuperscript{2}Indiana University, Bloomington, USA
{\{lucayongxu, raymondmyu, auszhang\}@tencent.com, zhuozhan@iu.edu}

Abstract
Although the conventional mask-based minimum variance distortionless response (MVDR) could reduce the non-linear distortion, the residual noise level of the MVDR separated speech is still high. In this paper, we propose a spatio-temporal recurrent neural network based beamformer (RNN-BF) for target speech separation. This new beamforming framework directly learns the beamforming weights from the estimated speech and noise spatial covariance matrices. Leveraging on the temporal modeling capability of RNNs, the RNN-BF could automatically accumulate the statistics of the speech and noise covariance matrices. Leveraging on the temporal modeling capability of RNNs, the RNN-BF could automatically accumulate the statistics of the speech and noise covariance matrices. The proposed GRNN-BF obtains better performance against prior arts in terms of speech quality (PESQ), speech-to-noise ratio (SNR) and word error rate (WER).

Index Terms: MVDR, Spatio-temporal RNN beamformer, ADL-MVDR, GEV, GRNN-BF, speech separation

1. Introduction

The mask-based MVDR \cite{1, 2, 3, 4, 5, 6} could obtain distortionless speech while existing purely “black box” neural network (NN) based speech separation methods \cite{7, 8, 9, 10, 11} always suffer from the non-linear distortion problem \cite{6}. However, the residual noise level of mask-based MVDR method is still high \cite{12, 6}. Most of mask-based beamformers are optimized in the chunk-level \cite{1, 3, 5, 6}. The calculated beamforming weights are hence chunk-level which is not optimal for each frame. Furthermore, the matrix inversion in the traditional beamformer (e.g., MVDR) comes with the numerical instability problem \cite{13, 14, 15, 16}, which is caused by the singularity in the matrix inversion \cite{13}. Although this issue could be alleviated by using some techniques, e.g., diagonal loading \cite{17, 14, 13}, it is not fully solved. This problem could be worse in the end-to-end joint training system \cite{13, 6}. Time-varying mask-based beamformers were investigated in \cite{18, 19}, however they also have the numerical instability problem \cite{13, 14}. The recurrent neural network (RNN) was once proved to have the ability to solve the matrix inversion \cite{20, 21} and the eigenvalue decomposition problems \cite{22, 23}, which are two main matrix operations in most of the beamformers’ solutions, e.g., MVDR \cite{3, 5, 24} and Generalized Eigenvalue (GEV) beamformer \cite{25, 26}. Although different types of beamformers \cite{24, 3, 25, 27} have different optimization and constraint conditions, most of their solutions are derived from the estimated speech and noise covariance matrices. These prior studies inspired us to use RNNs to directly learn the beamforming weights from the estimated speech and noise covariance matrices. Hence, we recently proposed an all-deep-learning MVDR (ADL-MVDR) method \cite{28} which was superior to the traditional MVDR beamformer \cite{3, 6}. In the ADL-MVDR \cite{28}, the matrix inversion and principal component analysis (PCA) operations of traditional MVDR are replaced by two RNNs with the estimated speech and noise covariance matrices as the input. In this work, we propose more advanced and generalized RNN-based beamformers (RNN-BFs). Note that there were also several other learning based beamforming methods \cite{29, 30} which yielded worse performance than the traditional mask-based MVDR \cite{3} approach due to the lack of explicitly using the speech and noise covariance matrices information \cite{29}.

In this work, three contributions are made to further improve the ADL-MVDR \cite{28} beamformer. First, we propose a RNN-based GEV (RNN-GEV) beamformer, where it achieves slightly better performance than the ADL-MVDR \cite{28}. It indicates that the RNNs could also be incorporated into other traditional beamforming algorithms. Second, a generalized RNN beamformer (GRNN-BF) is proposed and it is superior to the RNN-GEV and the ADL-MVDR \cite{28}. The GRNN-BF directly learns the frame-level beamforming weights from covariance matrices without following conventional beamformers’ solutions. It suggests that the GRNN-BF could learn a better beamforming solution by automatically accumulating the covariance matrices across history frames. Finally, the layer normalization \cite{31} is proposed to replace the commonly used mask normalization \cite{32, 29, 6, 13} on the covariance matrices. The layer normalization is more flexible than the mask normalization and it can achieve better performance. These improvements make our proposed GRNN-BF perform the best in terms of PESQ, SNR and word error rate (WER) comparing to the traditional MVDR beamformers \cite{6} and the ADL-MVDR beamformer \cite{28}.

The rest of this paper is organized as follows. In section 2, traditional mask-based beamformers are described. Section 3 presents the proposed generalized RNN-based beamformers (GRNN-BFs). The experimental setups and results are provided in Section 4 and 5, respectively. Section 6 concludes this paper.

2. Traditional mask-based beamformers

This section describes the traditional mask-based beamformers. Given the \textit{M}-channel speech mixture \textit{y} = \{\textit{y}\textsubscript{1}, \textit{y}\textsubscript{2}, ..., \textit{y}\textsubscript{N}\}, the corresponding \textit{M}-channel target speaker’s speech and the noise (the sum of interfering speakers’ speech and background noise) waveforms are denoted as \textit{s} and \textit{n}, respectively. After applying short-time Fourier transform (STFT), we have \textit{Y}, \textit{S}, \textit{N} in the time-frequency (T-F) domain.

\begin{equation}
\textit{Y}(t, f) = \textit{S}(t, f) + \textit{N}(t, f)
\end{equation}

where \((t, f)\) indicates the time and frequency indices of the T-F domain variables. In conventional mask-based beamforming \cite{1, 3, 5, 6}, a neural network is used to predict the real-valued...
speech mask $\mathbf{R}_\mathbf{M}$ and the real-valued noise mask $\mathbf{R}_\mathbf{N}$. Then the speech covariance matrix $\mathbf{\Phi}_\mathbf{S}$ is calculated with the predicted speech mask $\mathbf{R}_\mathbf{M}$,$$
abla \sum_{t=1}^{T} \mathbf{R}_\mathbf{M}(t, f) \mathbf{Y}(t, f) \mathbf{Y}^H(t, f)$$

Where $T$ stands for the total number of frames in a chunk. $\mathbf{H}$ is the Hermitian transpose. The noise covariance matrix $\mathbf{\Phi}_\mathbf{N}$ could be calculated in the same way with the noise mask $\mathbf{R}_\mathbf{N}$. The MVD solution [3] can be derived as,$$\mathbf{w}_{\text{MVD}}(f) = \frac{\mathbf{\Phi}_\mathbf{S}^H(f) \mathbf{v}(f)}{\mathbf{v}^H(f) \mathbf{\Phi}_\mathbf{S}^H(f) \mathbf{v}(f)}, \quad \mathbf{w}_{\text{MVD}}(f) \in \mathbb{C}^M$$

where $\mathbf{v}(f)$ represents the steering vector at $f$-th frequency bin. $\mathbf{v}(f)$ could be derived by applying PCA on $\mathbf{\Phi}_\mathbf{S}(f)$, namely $\mathbf{v}(f) = \mathcal{P}(\mathbf{\Phi}_\mathbf{S}(f))$. Another type of commonly used beamformer is the generalized eigenvalue (GEV) [1], where its optimal solution is the generalized principle component [1, 26].$$\mathbf{w}_{\text{GEV}}(f) = \mathcal{P}(\mathbf{\Phi}_\mathbf{S}^H(f) \mathbf{\Phi}_\mathbf{S}(f)), \quad \mathbf{w}_{\text{GEV}}(f) \in \mathbb{C}^M$$

However, the beamforming weights $\mathbf{w}$ above are usually chunk-level [1, 3, 5, 6], which is not optimal for each frame. Furthermore, the matrix inversion involved in Eq. (3) and Eq. (4) has the numerical instability problem [13, 14, 15, 16]. Note that we have already applied the diagonal loading technique [17, 14, 13] to alleviate this problem in our MVDR baselines.

On the other hand, although the MVDR and GEV are two different beamformers, their solutions are both derived from the speech and noise covariance matrices, namely $\mathbf{\Phi}_\mathbf{S}$ and $\mathbf{\Phi}_\mathbf{N}$. This is also our motivation to use RNNs to directly learn the beamforming weights from $\mathbf{\Phi}_\mathbf{S}$ and $\mathbf{\Phi}_\mathbf{N}$.

3. Proposed generalized RNN beamformer

We aim to extract the target speaker’s speech from the multi-channel multi-talker overlapped mixture. As shown in Fig. 1, the whole system consists of a complex-valued ratio filter (cRF) estimator and the proposed spatio-temporal RNN beamformer. The predicted cRFs are used to calculate the covariance matrices. Then the proposed GRNN-BF learns the beamforming weights from the covariance matrices. The details of the input features and the cRF estimator will be described in Sec. 4. Our proposed GRNN-BF will be first illustrated here.

Recently we proposed the ADL-MVDR method [28] which is superior to the traditional mask-based MVDR beamformers [3, 5, 6]. The ADL-MVDR uses two RNNs to replace the matrix inversion and PCA in MVDR solution (defined in Eq. (3)). Here we explore to use RNNs to implement the GEV beamformer (defined in Eq. (4)) and another more generalized RNN beamformer. Layer normalization [31] is also proposed to replace the commonly used mask normalization [32, 29, 13], which is applied on the covariance matrices.

3.1. Layer normalization on covariance matrices

Before we use RNNs to learn the beamforming weights, the speech and noise covariance matrices should be first estimated. Real-valued masks [32, 33], complex-valued ratio mask (cRM) [34, 6] or complex-valued ratio filter (cRF) [35, 28] could be used to estimate the speech and noise. In our previous ADL-MVDR work [28], the cRF [35] is demonstrated to be better than the cRM [34]. The cRF [35] is $K \times K$ size cRM [34] by using nearby $K \times K$ T-F bins around $(t, f)$. With the speech cRFs $(t, f)$, the multi-channel target speech is estimated as,$$
abla \mathbf{S}(t, f) = \sum_{\tau_1=K}^{t} \sum_{\tau_2=-K}^{K} \mathbf{cRF}(t, f, \tau_1, \tau_2) \mathbf{Y}(t + \tau_1, f + \tau_2)$$

Then the frame-wise speech covariance matrix is calculated as,$$
abla \mathbf{\Phi}_\mathbf{S}(t, f) = \mathbf{S}(t, f) \mathbf{S}^H(t, f) \sum_{t=1}^{T} \mathbf{cRM}(t, f) \mathbf{cRM}(t, f)$$

Where $\mathbf{cRM}(t, f)$ is the center unit of the speech cRFs $(t, f)$. Given the noise cRFs $(t, f)$, the noise $\mathbf{N}(t, f)$ and the frame-wise noise covariance matrix $\mathbf{\Phi}_\mathbf{N}(t, f)$ could be estimated in the same way. Different from Eq. (2) where the covariance matrix are averaged over a chunk of frames in the traditional mask-based MVDR, the covariance matrices here are frame-wise. This is because the covariance matrices are later fed into RNNs where the unidirectional RNN could automatically accumulate the statistics of covariance matrices across history frames in a recursive way. Note that the denominator in Eq (6) is the commonly used mask normalization [32, 29, 6, 13] to normalize the covariance matrix and equalize the training loss and gradients of different chunks in a mini-batch. In this work, we propose to use the layer normalization [31] to normalize the covariance matrices to achieve better performance.$$
abla \mathbf{\Phi}_\mathbf{S}(t, f) = \text{LayerNorm}(\mathbf{S}(t, f) \mathbf{S}^H(t, f))$$

Where the layer normalization [31] applies per-element scale and bias with learnable affine transform, which is more flexible than the mask normalization. Another layer normalization is also adopted for $\mathbf{\Phi}_\mathbf{N}(t, f)$.

Figure 1: The system framework includes the dilated Conv-1D blocks based complex-valued ratio filter (cRF) estimator and the proposed spatio-temporal RNN beamformer (RNN-BF). The cRF estimator is actually a Conv-TasNet variant [9] with a fixed STFT encoder [33]. $\circ$ indicates the complex-domain multiplication to estimate the multi-channel speech $\mathbf{S}$ and noise $\mathbf{N}$ through cRF (as shown in Eq. (5)). $\mathcal{M}$ is the matrix multiplication of beamforming (see Eq. (13)). The speech covariance matrix $\mathbf{\Phi}_\mathbf{S}$, noise covariance matrix $\mathbf{\Phi}_\mathbf{N}$ and beamforming weights $\mathbf{w}$ are complexed-valued variables, their real parts and imaginary parts are reshaped and concatenated together. The time-domain scale-invariant SNR (St-SNR) loss [9] is applied for end-to-end training.
3.2. Spatio-temporal RNN GEV beamformer

Similar to the ADL-MVDR [28], here the proposed spatio-temporal RNN GEV beamformer (RNN-GEV) also takes the estimated target speech covariance matrix $\mathbf{S}_n(t, f)$ and the noise covariance matrix $\mathbf{S}_n(t, f)$ as the input to predict the frame-wise beamforming weights. Following the solution of the traditional GEV beamformer defined in Eq. (4), we reformulate its form in the RNN-based beamforming framework as,

$$\hat{\mathbf{S}}_N^{-1}(t, f) = \text{RNN}(\mathbf{S}_n(t, f))$$ (8)

$$\hat{\mathbf{S}}_N(t, f) = \text{RNN}(\mathbf{S}_n(t, f))$$ (9)

$$\mathbf{w}_{\text{RNN-GEV}}(t, f) = \text{DNN}(\hat{\mathbf{S}}_N^{-1}(t, f), \hat{\mathbf{S}}_N(t, f))$$ (10)

$$\hat{\mathbf{W}}_{\text{RNN-GEV}}(t, f) = (\mathbf{w}_{\text{RNN-GEV}}(t, f))^\top \mathbf{Y}(t, f)$$ (11)

where $\mathbf{w}_{\text{RNN-GEV}}(t, f) \in \mathbb{C}^M$. $\hat{\mathbf{S}}_N(t, f)$ is the accumulated speech covariance matrix from the history frames by leveraging on the temporal modeling capability of RNNs. $\hat{\mathbf{S}}_N^{-1}(t, f)$ is assumed to be the matrix inversion of $\mathbf{S}_n(t, f)$. Instead of using the actual generalized PCA (as in Eq. (4)), a deep neural network (DNN) is utilized to calculate the beamforming weights for RNN-GEV. Hinton et al [36] shows that the DNN has the ability to conduct the non-linear generalized PCA.

3.3. Generalized spatio-temporal RNN beamformer

Finally, we propose a more generalized spatio-temporal RNN beamformer (GRNN-BF) without following any traditional beamformers’ solutions. This is motivated by that, different beamformers (e.g., MVDR and GEV) have different solutions but almost all solutions are derived from the speech and noise covariance matrices. The neural networks could be able to learn the beamformer’s solutions. This is motivated by that, different beamformer (GRNN-BF) without following any traditional GEV beamformer defined in Eq. (4), we reformulate it as:

$$\hat{\mathbf{S}}_N^{-1}(t, f) = \text{RNN}(\mathbf{S}_n(t, f), \mathbf{S}_s(t, f))$$ (12)

$$\hat{\mathbf{S}}_N(t, f) = (\mathbf{w}_{\text{GRNN-BF}}(t, f))^\top \mathbf{Y}(t, f)$$ (13)

where $\mathbf{w}_{\text{GRNN-BF}}(t, f) \in \mathbb{C}^M$. The input for the RNN-DNN is the concatenated tensor of $\mathbf{S}_n(t, f)$ and $\mathbf{S}_s(t, f)$. All of the covariance matrices and beamforming weights are complex-valued, and we concatenate the real and imaginary parts of any complex-valued tensors in the whole work.

4. Dataset and experimental setup

Dataset: The methods are evaluated on the mandarin audio-visual corpus [37, 33], which is collected from YouTube [38]. The dataset has 205500 clean speech segments (about 200 hours) over 1500 speakers. The audio sampling rate is 16 kHz. 512-point of STFT is used to extract features along 32ms Hann window with 50% overlap. There are one to three overlapped speakers in the simulated 15-channel mixture signal. The signal-to-interference ratio (SIR) is ranging from -6 to 6 dB. Noise with 18-30 dB SNR is added to all the 15-channel mixtures [37]. We use a 15-element non-uniform linear array. Based on the image-source simulation method [39], the simulated dataset contains 190000, 15000 and 500 multi-channel mixtures for training, validation and testing. The virtual acoustic room size is ranging from 4m×4m×2.5m to 10m×5m×6m. The reverberation time T60 is sampled in a range of 0.05s to 0.7s.

**cRF estimator:** We use the complex-valued ratio filter (cRF) [28, 35] to estimate the covariance matrices. As shown in Fig. 1, the input to the cRF estimator includes a 15-channel mixture audio and a target direction of arrival (DOA) (θ). From the multi-channel audio, log-power spectra (LPS) and interaural phase difference (IPD) [37] features are extracted. For the simulated data, the ground-truth target DOA is known. For the real-world scenario, we have the hardware where the 180° wide-angle camera and the 15-channel microphone array are aligned [6]. Hence the target DOA (θ) could be roughly estimated from the camera view by locating the target speaker’s face (see our actual hardware demo website: https://yongxuustc.github.io/grnnbf). Then the DOA guided directional feature (DF) [40], namely $d(θ)$, is estimated by calculating the cosine similarity between the target steering vector v and IPDs [40, 33]. Target DOA and $d(θ)$ are speaker-dependent features which can be used to extract the target speech. LPS, IPD and $d(θ)$ are merged and fed into a Conv-TasNet variant [9] with a fixed STFT encoder [37, 33]. A stack of eight successive dilated Conv-1D layers with 256 channels forms a network block and four blocks are piled together. The estimated cRF [35] size ($K \times K$) is empirically set to 3×3 [35, 28].

As for the RNN-BF module, the RNNs have 2-layer gated recurrent units (GRUs) with 500 hidden nodes. The non-linear DNN layer has 500 PreLU units. There are 30 linear units at the output DNN layer to predict the frame-wise beamforming weights. The model is trained in a chunk-wise mode with 4-second chunk size, using Adam optimizer. Initial learning rate is set to $1e^{-4}$. The objective is to maximize the time-domain scale-invariant source-to-noise ratio (Si-SNR) [9]. Pytorch 1.1.0 was used. Gradient norm is clipped with max norm 10. We evaluate the systems by using different metrics, including PEIQ, Si-SNR (dB), signal-to-distortion ratio (SDR) (dB). A commercial general-purpose mandarin speech recognition Tencent API [41] is used to test the ASR performance in WER. Note this work only focuses on speech separation and denoising without dereverberation. Hence the reverberant clean (without dereverberation) is used as the reference signal.

5. Results and discussions

We evaluate the target speech separation performance in the overlapped multi-talker scenario. The spatial angle between the target speaker and others (interfering speakers) lies within 0-180°. The more overlapped speakers and the smaller spatial angle would lead to more challenging separation tasks. The detailed PEIQ scores across different scenarios (i.e., angle between the target speaker and other speakers; number of overlapped speakers) are presented in Table 1. Other metrics are given with the average results.

**GRNN-BF vs. conventional MVDRs:** Two traditional MVDR systems, MVDR with mask normalization (i) [6, 28] and multi-tap (i.e., $\lfloor t - 1 \rfloor$) MVDR with mask normalization (ii) [6, 28] are compared here. They also use the cRF estimator to calculate covariance matrices but replace the RNN-BF module (as shown in Fig. 1) with conventional MVDR or multi-tap MVDR [6] solutions. They both work reasonably well, e.g., the multi-tap MVDR (iii) achieves 13.52% WER. However, the proposed GRNN-BF with mask normalization (vii) could obtain significantly better performance. The proposed GRNN-BF (vii) increases the average PEIQ to 3.52 from 3.08 of multi-tap MVDR (ii) and 2.92 of MVDR (i). The WER of the proposed GRNN-BF (vii) is better than the multi-tap MVDR (iii), i.e., 11.86 vs. 13.52. The corresponding Si-SNR and SDR are in-
Table 1: PESQ, Si-SNR(dB) [9], SDR(dB) and WER(%) results among Conv-TasNet with STFT [33], several MVDRs and proposed GRNN-BF systems. "MN" and "LN" denote mask normalization and layer normalization on the covariance matrices, respectively.

<table>
<thead>
<tr>
<th>systems/metrics</th>
<th>Angle between target &amp; others</th>
<th># of overlapped speakers</th>
<th>PESQ(-0.5, 4.5)</th>
<th>Si-SNR</th>
<th>SDR</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0-15</td>
<td></td>
<td>1SPK</td>
<td>2SPK</td>
<td>5SPK</td>
<td>Avg.</td>
</tr>
<tr>
<td>Reverberant clean reference</td>
<td>4.50</td>
<td></td>
<td>4.50</td>
<td>4.50</td>
<td>4.50</td>
<td>4.50</td>
</tr>
<tr>
<td>Mixture (overlapped speech+noise)</td>
<td>1.88</td>
<td></td>
<td>2.03</td>
<td>2.02</td>
<td>2.76</td>
<td>2.16</td>
</tr>
<tr>
<td>Conv-TasNet with STFT (i) [33]</td>
<td>2.75</td>
<td></td>
<td>3.09</td>
<td>3.06</td>
<td>3.10</td>
<td>3.10</td>
</tr>
<tr>
<td>Multi-tap MVDR w/MN (iii) [6]</td>
<td>2.67</td>
<td></td>
<td>3.10</td>
<td>3.06</td>
<td>3.06</td>
<td>3.06</td>
</tr>
<tr>
<td>ADL-MVDR w/ MN (iv) [28]</td>
<td>3.04</td>
<td></td>
<td>3.48</td>
<td>3.41</td>
<td>3.07</td>
<td>3.42</td>
</tr>
<tr>
<td>Prop. RNN-GEV w/ MN (v)</td>
<td>3.11</td>
<td></td>
<td>3.55</td>
<td>3.48</td>
<td>3.14</td>
<td>3.48</td>
</tr>
<tr>
<td>Prop. RNN-GEV w/ LN (vii)</td>
<td>3.15</td>
<td></td>
<td>3.57</td>
<td>3.41</td>
<td>3.17</td>
<td>3.51</td>
</tr>
<tr>
<td>Prop. GRNN-BF w/ MN (viii)</td>
<td>3.17</td>
<td></td>
<td>3.58</td>
<td>3.51</td>
<td>3.19</td>
<td>3.52</td>
</tr>
<tr>
<td>Prop. GRNN-BF w/ LN (viii)</td>
<td>3.23</td>
<td></td>
<td>3.62</td>
<td>3.57</td>
<td>3.24</td>
<td>3.56</td>
</tr>
</tbody>
</table>

6. Conclusions

In summary, we proposed a generalized RNN beamformer (GRNN-BF) that learns the frame-level beamforming weights directly from the estimated speech and noise covariance matrices. The layer normalization achieves better performance than the mask normalization for normalizing the speech and noise covariance matrices. The proposed GRNN-BF with layer normalization achieves the best objective scores (PESQ, Si-SNR, SDR) and the lowest WER among all evaluated systems. It achieves relative 10.8% and 16% WER reduction against the prior art methods, ADL-MVDR [28] and the conventional multi-tap MVDR [6], respectively. Although we only tested the proposed GRNN-BF on the DOA-guided target speech separation task, it could also be used for general scenarios without DOA information, e.g., the multi-channel speech enhancement task or the permutation invariant training (PIT) [10] based multi-channel speech separation task.
7. References


