Neural Spatio-Temporal Beamformer for Target Speech Separation

Yong Xu\textsuperscript{1}, Meng Yu\textsuperscript{1}, Shi-Xiong Zhang\textsuperscript{1}, Lianwu Chen\textsuperscript{2}, Chao Weng\textsuperscript{1}, Jianming Liu\textsuperscript{1}, Dong Yu\textsuperscript{1}

\textsuperscript{1}Tencent AI lab, Bellevue, WA, USA
\textsuperscript{2}Tencent AI lab, Shenzhen, China

\{lucayongxu, raymondmyu, auszhang, lianwuchen, cweng, jianmingliu, dyu\}@tencent.com

Abstract

Purely neural network (NN) based speech separation and enhancement methods, although can achieve good objective scores, inevitably cause nonlinear speech distortions that are harmful for the automatic speech recognition (ASR). On the other hand, the minimum variance distortionless response (MVDR) beamformer with NN-predicted masks, although can significantly reduce speech distortions, has limited noise reduction capability. In this paper, we propose a multi-tap MVDR beamformer with complex-valued masks for speech separation and enhancement. Compared to the state-of-the-art NN-mask based MVDR beamformer, the multi-tap MVDR beamformer exploits the inter-frame correlation in addition to the inter-microphone correlation that is already utilized in prior arts. Further improvements include the replacement of the real-valued masks with the complex-valued masks and the joint training of the complex-mask NN. The evaluation on our multi-modal multi-channel target speech separation and enhancement platform demonstrates that our proposed multi-tap MVDR beamformer improves both the ASR accuracy and the perceptual speech quality against prior arts.

Index Terms: target speech separation, multi-tap MVDR, mask-based MVDR, spatio-temporal beamformer

1. Introduction

The deep learning based speech enhancement [1, 2, 3] and speech separation [4, 5, 6] methods have attracted lots of research attention since the renaissance of the neural network. However, the purely neural network based front-end approaches inevitably cause nonlinear speech distortions [7]. The speech distortion can degrade the performance of the speech recognition system [7], even for the commercial general-purpose ASR engine which is already robust enough to the background noise. The refinement [7] or joint training [8, 9, 10] on the enhanced speech can make the front-end output and the back-end acoustic model match better. Nevertheless, these approaches cannot explicitly reduce the speech distortion. Furthermore, the joint training with the commercial general-purpose ASR engine is usually not feasible either because the training data is too large and noisy or because the ASR engine is third-party.

For example, the fully-convolutional time-domain audio separation network (Conv-TasNet) [11] has shown significant improvement in the speech separation task. We further proposed several audio-visual [12] or multi-channel [13, 14, 15] speech separation techniques based on the Conv-TasNet. Although these models can obtain substantial gain according to the objective measures [11, 12, 14], they cause some nonlinear distortions in the separated speech because such distortion is not considered for attenuation in the model.

On the other hand, the minimum variance distortionless response (MVDR) beamformer [16], as its name suggests, explicitly requires distortionless filtering on the target direction [17] and thus has significantly less speech distortions in the separated speech. Recently, MVDR have been improved by exploiting better covariance matrix computation through NN estimated ideal ratio masks (IRMs) [18, 19, 20, 21, 22]. Although NN-mask based MVDR [23, 24] can achieve better ASR accuracy than purely NN-based approaches due to less distortions, the residual noise level of the enhanced speech is high.

In this work, we propose a neural spatio-temporal beamforming approach, named multi-tap MVDR beamformer with complex-valued masks, for speech separation and enhancement to simultaneously obtain high ASR accuracy and PESQ score. The multi-tap MVDR for the multi-channel scenario is inspired by the multi-frame MVDR on the single channel [25, 26, 27, 28, 29]. Similar to the MVDR, multi-tap MVDR enforces distortionless at the target direction. Different from the MVDR and multi-frame MVDR, which utilize the inter-microphone correlation and inter-frame correlation, respectively, the multi-tap MVDR exploits both correlations and thus has higher potential. Benesty et al. [28] proposed a similar idea for the multi-channel speech enhancement from the signal processing perspective. Our proposed approach differentiates with theirs in that ours is NN-mask based. Additional novelties in our approach include the replacement of the real-valued masks [4, 19, 15, 13] with the complex-valued masks (CMs), and the joint training of the CMs in the multi-tap MVDR framework. We evaluated our proposed approach on our multi-modal multi-channel target speech separation platform [13, 15] by replacing the speech separation component shown in Fig. 1. Our experiments indicate that the multi-tap MVDR beamformer with CMs improves both the ASR accuracy and the perceptual speech quality against prior arts.

The rest of the paper is organized as follows. In Section 2, we describe our proposed multi-tap MVDR beamformer with complex-valued masks. In Section 3 we present the baseline system and the experimental setup. The results are given in Section 4. We conclude the paper in Section 5.

2. Neural Spatio-Temporal Beamformer: Multi-tap MVDR with Complex Mask

2.1. Spatial filtering: MVDR beamformer

MVDR is a widely used beamformer for ASR [18]. It minimizes the power of the noise (interfering speech + additive noise) while ensuring that the signal at the desired direction is not distorted. Mathematically, this can be formulated as,

\[
\mathbf{w}^{\text{MVDR}} = \arg \min_{\mathbf{w}} \mathbf{w}^{H} \Phi_{\text{NN}} \mathbf{w} \quad \text{s.t.} \quad \mathbf{w}^{H} \mathbf{v} = 1 \tag{1}
\]

Where \( \Phi_{\text{NN}}(f) \in \mathbb{C}^{M \times M} \) is the covariance matrix of noise \( N \) at frequency bin \( f \) and \( \mathbf{v} \) is the target steering vector. \( M \) is the number of the microphone. The constraint \( \mathbf{w}^{H} \mathbf{v} = 1 \) is important to guarantee that the target source is distortionless. There are several solution variants for this optimization problem.
The solution based on the reference channel selection [31, 30, 16] is

\[ w_{\text{MVDR}}(f) = \frac{\Phi_{SS}^{-1}(f) \Phi_{SS}(f)}{\text{Trace}(\Phi_{SS}(f) \Phi_{SS}(f))} \cdot u(f) \in C^M \]  

(2)

where \( u \) is the one-hot vector representing a reference microphone channel and \( \Phi_{SS} \) represents the covariance matrix of the target speech. The key step for the beamforming is to estimate the two covariance matrices, namely \( \Phi_{NN} \) and \( \Phi_{SS} \). For the traditional signal processing-based techniques, the noise mask for the covariance matrix calculation. The ReLU-mask (a.k.a. STFT magnitude mask) [32] is defined as,

\[ \text{ReLU-Mask}(t, f) = \frac{|S(t, f)|}{|Y(t, f)|} \]  

(3)

where \( |S| \) and \( |Y| \) represents the target speech magnitude and noisy speech magnitude, respectively. The range of ReLU-Mask lies in \([0, +\infty]\). Note that no value clipping is needed in our implementation, which is different from the FFT-MASK in [32] where the value was clipped into \([0,10]\). This is because our scale-invariant source-to-noise ratio (Si-SNR) [11] loss function (shown in Eq. 1) is optimized on the recovered time-domain waveform rather than on the mask itself.

Given the real-valued mask (RM) (as the output of a sigmoid or ReLU function) defined on the magnitude, the covariance matrix \( \Phi_{SS} \) of the beamformer can be computed as

\[ \Phi_{SS}(f) = \frac{\sum_{t=1}^{T} \text{RM}_S^*(t, f) Y(t, f) Y^H(t, f)}{\sum_{t=1}^{T} \text{RM}_S^2(t, f)} \]  

(4)

where \( T \) is the chunk size. We argue that better covariance matrix estimation can be achieved with the complex-valued mask (CM) for speech separation in this work and ASR [20, 33]. The CM \((t, f)\) was first proposed in [34] as,

\[ S = \hat{S}_r + j \hat{S}_i = (CM_r + jCM_i) \ast (Y_r + j Y_i) = CM \ast Y \]  

(5)

where \( r \) and \( i \) denote the real and imaginary part of the complex spectrum, respectively. The theoretical range of CM lies in \([-\infty, +\infty]\). In [34], the CM was compressed into \([-10, 10]\) since their model was trained to estimate the CM itself. In our implementation, however, value compression is not necessary and can be harmful. We implicitly estimate the CM with a linear activation function and then multiply it with the complex spectrum of the mixture to obtain the estimated clean speech. The Si-SNR loss [11] function is optimized on the reconstructed time-domain waveform rather than on CM itself. With CM, \( \Phi_{SS} \) can be rewritten as

\[ \Phi_{SS}(f) = \frac{\sum_{t=1}^{T} \hat{S}_r(t, f) \hat{S}_r^H(t, f)}{\sum_{t=1}^{T} \text{CM}_S \ast \text{CM}_S^H(t, f)} \]  

(6)

\[ = \frac{\sum_{t=1}^{T} \text{CM}_S(t, f) Y(t, f) (CM_S(t, f) Y(t, f))^H}{\sum_{t=1}^{T} \text{CM}_S \ast \text{CM}_S(t, f)} \]  

(7)

where \( M_S \) and CMs are shared across channels. The mask normalization in the denominator is the key to success since the weighted mask is to attend on the most related frames to calculate \( \Phi \). \( \Phi_{NN}(f) \) can be computed in the similar way. According to the MVDR solution Eq. (2), the beamformed speech of the target speaker can be estimated by,

\[ \hat{S}(t, f) = w^H(t, f) Y(t, f) \]  

(8)

2.3. Neural spatio-temporal filtering: CM based Multi-tap MVDR

Although MVDR can improve the ASR performance, it keeps the speech distortion low at the cost of high residual noise.
networks to compute the covariance matrix. Similar to Eq. (2),
that we are using complex-valued masks estimated by neural
Benesty et al. [28] proposed the multi-channel speech enhance-
Φ

speech covariance matrix
same way. Then we can calculate the extended
L

¯

as

\( \mathbf{Y}(t, f) = \left[ \mathbf{Y}^T(t, f), \mathbf{Y}^T(t-1, f), \ldots, \mathbf{Y}^T(t-L+1, f) \right] \in \mathbb{C}^{M \times L} \)

The corresponding \( \mathbf{S}, \mathbf{N}, \mathbf{CM} \) can be defined in the
the same way. Then we can calculate the extended \( \mathbf{L} \)-tap target
speech covariance matrix \( \Phi_{SS}(f) \in \mathbb{C}^{M \times ML} \), as

\[
\Phi_{SS}(f) = \sum_{t=1}^{T} \left[ (\mathbf{CM}_t(f,t)\mathbf{Y}(t,f))\mathbf{CM}_t(f,t)^{H} \right] - \sum_{t=1}^{T} \mathbf{CM}_t(f,t)\mathbf{CM}_t(f,t)^{H}
\]

Benesty et al. [28] proposed the multi-channel speech enhance-
ment filter. However, our approach is different from theirs in
that we are using complex-valued masks estimated by neural
networks to compute the covariance matrix. Similar to Eq. (2),
the multi-tap MVDR solution is

\[
\hat{w}(f) = \frac{\Phi_{NN}(f)^{-1}\Phi_{SS}(f)}{\text{Trace}(\Phi_{NN}(f)^{-1}\Phi_{SS}(f))} \mathbf{u}, \quad \hat{w}(f) \in \mathbb{C}^{M \times L}
\]

where \( \mathbf{u} \) is an expanded one-hot vector of \( \mathbf{u} \) with padding zeros
in the tail. Note that the multi-tap MVDR follows the optimiza-
tion process of MVDR in Eq. (1) for the multi-channel scenario.
Hence, it is different from the multi-frame MVDR (MFMVDR)
[26, 27] defined on the single channel. The enhanced speech of
the multi-tap MVDR can be obtained as,

\[
\mathbf{S}(t, f) = \hat{w}^{H}(f)\mathbf{Y}(t, f)
\]

The beamformed spectrum is converted to the time-domain
waveform via iSTFT. Finally, the Si-SNR loss [11] calculated
along 32ms Hann window with 50% overlap. A mouth region
(size=112x112x3) detection program [15] is run on the target
speaker’s video to capture the the lip movements.

The new and larger multi-talker multi-channel far-field
dataset are simulated in the similar way with our previous work
[13, 15]. The simulated dataset contains 190000, 15000 and 500
multi-channel mixtures for training, validation and testing. The
speakers in the test set are unseen in the training set. The tran-
script of the speech for the ASR evaluation is manually labeled
by human in this work. The multi-channel signals are generated
by convolving speech with RIRs simulated by image-source
method [42]. The signal-to-interference ratio (SIR) is ranging
from -6 to 6 dB. Also, noise with 18-30 dB SNR is added to
all the multi-channel mixtures [13]. A commercial general-
purpose mandarin speech recognition Yitu API [43] (unorre-
ted to this work) is used to test the ASR performance.

The multi-modal network is trained in a chunk-wise mode
with chunk size 4 seconds, using Adam optimizer with early
stopping. Initial learning rate is set to 1e-3. The \( \mathbf{L} \)-tap in
the multi-tap MVDR is set to 3 empirically. Pytorch 1.1.0 was used.

4. Results and Discussions

The PESQ and ASR word error rate (WER) results are shown in
Table 1 to compare among purely network-based systems and
several jointly trained MVDR systems. Note that we only
conduct speech separation and denoising without dereverbera-
tion in this work. Our systems work well on different scena-
rios, e.g., different angles between the target speaker and other
speakers, various number of overlapped speakers. The scenar-
io, e.g., small angles (\( \leq 45^\circ \)) or more overlapped speakers, are
a bit more challenging.

Real-valued mask VS CM: The linear uncompressed complex
mask (CM) based system (iv) achieves higher PESQ (3.00

\[
\text{shown in Fig. 1, a 3-D residual network [37, 39, 40] is adopted}
\text{to extract the target speech related lip movement embeddings.}
\]

Audio encoder: The audio input includes the speaker-

independent features (e.g., log-power spectra (LPS) and inter-

aural phase difference (IPD) [13]) and speaker-dependent fea-
tures (e.g., directional feature \( d(\theta) \) [41, 15]). As shown in Fig.
1, the 15-element non-uniform linear microphone array [13]
is co-located with the 180\(^\circ\) wide-angle camera. The location of
the target speaker’s face in the whole camera view can provide
a rough DOA estimation of the target speaker. Chen et al. [41]
proposed a location guided directional feature (DF) \( d(\theta) \) to
extract the target speech from the specific DOA. DF aims at cal-
culating the cosine similarity between the target steering vector
\( \psi(\theta) \) and IPDs [41]. The LPS, IPDs and DF are merged and fed
into a bunch of dilated 1D-CNNs. The details can be found in
our previous work [13, 15].

Then the concatenated lip embeddings and audio embed-
dings [13] are used to predict the sigmoid mask (i.e., IRM) or
the ReLU-mask (Eq. (3)) used in our previous work [13], or
the complex-valued mask (as Eq. (5)) proposed in this study.

3.2. Dataset and experimental setup

The mandarin audio-visual corpus [13] used for experiments is
collected from Youtube. We use SNR estimation tool and
face detection tool to filter out low SNR (\( \leq 17dB \)) and multi-
face videos [13], resulting in 205500 clean video segments with
single face (about 200 hours) over 1500 speakers. The sam-
pling rate for audio and video are 16 kHz and 25 fps respec-
tively. 512-point of STFT is used to extract audio features
along 32ms Hann window with 50% overlap. A mouth region
(size=112x112x3) detection program [15] is run on the target
speaker’s video to capture the the lip movements.

The new and larger multi-talker multi-channel far-field
dataset are simulated in the similar way with our previous work
[13, 15]. The simulated dataset contains 190000, 15000 and 500
multi-channel mixtures for training, validation and testing. The
speakers in the test set are unseen in the training set. The tran-
script of the speech for the ASR evaluation is manually labeled
by human in this work. The multi-channel signals are generated
by convolving speech with RIRs simulated by image-source
method [42]. The signal-to-interference ratio (SIR) is ranging
from -6 to 6 dB. Also, noise with 18-30 dB SNR is added to
all the multi-channel mixtures [13]. A commercial general-
purpose mandarin speech recognition Yitu API [43] (unorre-
ted to this work) is used to test the ASR performance.

The multi-modal network is trained in a chunk-wise mode
with chunk size 4 seconds, using Adam optimizer with early
stopping. Initial learning rate is set to 1e-3. The \( \mathbf{L} \)-tap in
the multi-tap MVDR is set to 3 empirically. Pytorch 1.1.0 was used.

3. Experimental Setup and Baselines

We evaluate our proposed methods on our multi-modal multi-
channel target speech separation platform [13, 15]. The audio-
visual structure is shown in Fig. 1 and briefly overviewed below.

3.1. Multi-modal multi-channel mask estimator baseline

As shown in Fig. 1, we use the direction of arrival (DOA) of
the target speaker and the speaker-dependent lip sequence for
informing the dilated convolutional neural networks (CNNs) to
extract the target speech from the multi-talker mixture.

Video encoder: The captured video can provide two im-
portant speaker-dependent information, lip movement sequence
and the DOA of the target speaker (denoted as \( \theta \) in Fig. 1). The
lip movement has been proven effective for the speech separa-
tion in [36, 37, 38, 12, 15]. In this work, we utilize the mouth re-

region RGB pixels to represent the target speaker’s lip feature. As

58
Figure 2: Separated spectrogram demos of different systems.

Table 1: PESQ and WER results of some dilated CNN baselines and proposed jointly trained multi-tap MVDR system.

<table>
<thead>
<tr>
<th>Systems/Metrics</th>
<th>Angle between target &amp; others</th>
<th># of overlapped speakers</th>
<th>PESQ</th>
<th>WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0-15°</td>
<td>15-45°</td>
<td>45-90°</td>
<td>90-180°</td>
</tr>
<tr>
<td>Reverberant Clean (reference)</td>
<td>4.50 4.50 4.50 4.50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mixture (interfering speech + noise)</td>
<td>4.50 4.50 4.50 4.50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ReLU mask (Audio only) on Channel 0</td>
<td>2.50 2.68 2.88 2.86</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ReLU mask (Lip only) on Channel 0</td>
<td>2.44 2.52 2.74 2.68</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ReLU mask (Baseline) on Channel 0</td>
<td>2.56 2.74 2.93 2.89</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complex mask (CM) on Channel 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.64 2.84 3.00 3.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sigmoid mask MVDR joint train (JT) (vi)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.27 2.59 2.82 2.73</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ReLU mask MVDR JT (vii)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.52 2.74 2.94 2.85</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CM MVDR JT (vii)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.55 2.77 2.97 2.89</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prop. CM multi-tap MVDR JT (viii)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.70 3.00 3.20 3.13</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Distortionless advantage by sacrificing the strength of residual noise reduction [30], e.g., the jointly trained CM-based MVDR (vii) only achieves 2.91 PESQ on average and is lower than purely network-based system (iv) with 3.00 PESQ.

CM-based multi-tap MVDR: The proposed jointly trained complex mask based multi-tap MVDR (viii) can get the best average PESQ, i.e., 3.10 and lowest WER, i.e., 9.96%, surpassing the best purely network-based system (iv). Compared to the common MVDR (vii), the multi-tap MVDR (viii) can achieve about 0.2 PESQ improvement on the 2/3-speaker cases since the multi-tap MVDR can utilize the inter-frame correlation and reduce the uncorrelated noise. The difference is also shown in Fig. 2 where the proposed multi-tap MVDR can reduce more residual noise while ensuring the distortionless constraint. More demos (including real-world testing demos) can be found at our website: https://yongxuustc.github.io/mtmvdv.

Directional feature VS lip feature: As introduced in Sec. 3.1, two speaker dependent features are used in this work, namely lip features and the DF ($d(\theta)$). Although multi-modality evaluation is not the focus of this work, we compare the audio only (using $d(\theta)$ w/o lip) and the lip only (using lip w/o $d(\theta)$) setup for the ablation study. The audio only system (i) is better than the lip only system (ii) (WER 17.89% vs 23.25%). It indicates that the DF ($d(\theta)$) is more distinct than lip feature. But when the two modalities are concatenated together to form the system (iii), slightly better performance can be achieved with WER 17.44%. More analysis about the multi-modality fusion can be found in our previous work [13, 15].

5. Conclusions and Future Work

In this work, we proposed the multi-tap MVDR with complex-valued masks (CMs). We demonstrated that CM can achieve less distortion and better ASR performance for the purely neural network based systems, and can better estimate the covariance matrix in the mask-based beamformer, than the real-valued masks. With the proposed CM based multi-tap MVDR, we obtain both the best ASR performance and PESQ among all systems. Compared to the purely neural network baseline using ReLU-mask, multi-tap MVDR can significantly reduce the WER from 17.44 % to 9.96% and improve the PESQ from 2.92 to 3.10 on average. We want to emphasize that the results achieved with multi-tap MVDR indicates that using filter-based instead of mask-based models for speech separation is promising. We will further extend the spatio-temporal filtering to spatio-temporal-frequency filtering and conduct separation and dereverberation in an integrated framework.
6. References


