INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Neural Spatio-Temporal Beamformer for Target Speech Separation

Yong Xu[1], *Meng Yu*[1], *Shi-Xiong Zhang*[1], *Lianwu Chen*[2], *Chao Weng*[1], *Jianming Liu*[1], *Dong Yu*[1]

[1]Tencent AI lab, Bellevue, WA, USA
[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,

$$\mathbf{w_{MVDR}} = \arg\min_{\mathbf{w}} \mathbf{w}^{\mathsf{H}} \mathbf{\Phi_{NN}} \mathbf{w} \quad \mathbf{s.t.} \quad \mathbf{w}^{\mathsf{H}} \mathbf{v} = \mathbf{1} \tag{1}$$

Where $\mathbf{\Phi_{NN}}(f) \in \mathbb{C}^{M \times M}$ is the covariance matrix of noise $\mathbf{N}$ at frequency bin f and $\mathbf{v}$ is the target steering vector. M is the number of the microphone. The constraint $\mathbf{w}^{\mathsf{H}}\mathbf{v} = \mathbf{1}$ is important to guarantee that the target source is distortionless. There are several solution variants for this optimization problem

http://dx.doi.org/10.21437/Interspeech.2020-1458

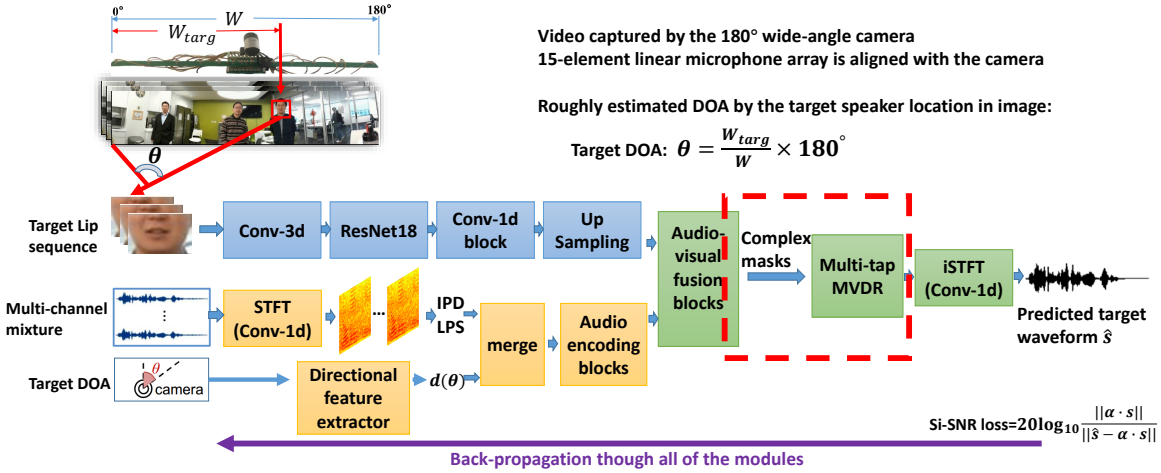

Figure 1: *Joint training of the multi-tap MVDR with complex-valued masks. The* ***complex masking*** *and* ***multi-tap MVDR****, highlighted in the dashed rectangle, which are the focus of this paper.* $\alpha = \hat{\mathbf{s}}^T\mathbf{s}/\mathbf{s}^T\mathbf{s}$ *is a scaling factor in the time-domain Si-SNR loss.*

[16, 30]. The solution based on the reference channel selection [31, 30, 16] is

$$\mathbf{w}_{\mathbf{MVDR}}(f) = \frac{\mathbf{\Phi}_{\mathbf{NN}}^{-1}(f)\mathbf{\Phi}_{\mathbf{SS}}(f)}{\text{Trace}(\mathbf{\Phi}_{\mathbf{NN}}^{-1}(f)\mathbf{\Phi}_{\mathbf{SS}}(f))}\mathbf{u} \quad , \quad \mathbf{w}(f) \in \mathbb{C}^M \tag{2}$$

where $\mathbf{u}$ is the one-hot vector representing a reference microphone channel and $\mathbf{\Phi}_{\mathbf{SS}}$ represents the covariance matrix of the target speech. The key step for the beamforming is to estimate the two covariance matrices, namely $\mathbf{\Phi}_{\mathbf{NN}}$ and $\mathbf{\Phi}_{\mathbf{SS}}$. For the traditional signal processing based techniques, the noise frames and speech frames are tracked to update $\mathbf{\Phi}_{\mathbf{NN}}$ and $\mathbf{\Phi}_{\mathbf{SS}}$ in a recursive way. Research indicates that better results can be achieved using the mask-based covariance matrix estimation method with neural networks [18].

2.2. Neural spatial filtering: Mask based MVDR

The idea behind the mask based approach for covariance matrix estimation is that we may more accurately estimate the target speech, and thus the covariance matrix, given a NN-based mask estimator (will discuss in Sec. 3.1). The most commonly used mask for the mask-based beamforming [18] is ideal ratio mask (IRM) [32] or sigmoid mask. In this work, we extend to use ReLU-Mask and linear uncompressed complex-valued mask for the covariance matrix calculation. The ReLU-mask (a.k.a. STFT magnitude mask) [32] is defined as,

$$\text{ReLU-Mask}(t,f) = \frac{|S(t,f)|}{|Y(t,f)|} \tag{3}$$

Where $|S|$ and $|Y|$ represents the target speech magnitude and noisy speech magnitude, respectively. The range of ReLU-Mask lies in [0,+∞]. 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 Fig. 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 $\mathbf{\Phi}_{\mathbf{SS}}$ of the beamformer can be computed as

$$\mathbf{\Phi}_{\mathbf{SS}}(f) = \frac{\sum_{t=1}^{T} \text{RM}_{\mathbf{S}}^2(t,f)\mathbf{Y}(t,f)\mathbf{Y}^{\mathsf{H}}(t,f)}{\sum_{t=1}^{T} \text{RM}_{\mathbf{S}}^2(t,f)} \tag{4}$$

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 $\text{CM}(t,f)$ was first proposed in [34] as,

$$S = S_r + jS_i = (\text{CM}_r + j\text{CM}_i) * (Y_r + jY_i) = \text{CM} * Y \tag{5}$$

where r and i denote the real part and the imaginary part of the complex spectrum, respectively. The theoretical range of CM lies in [$-\infty$,+∞]. 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, $\mathbf{\Phi}_{\mathbf{SS}}$ can be rewritten as

$$\mathbf{\Phi}_{\mathbf{SS}}(f) = \frac{\sum_{t=1}^{T} \hat{\mathbf{S}}(t,f)\hat{\mathbf{S}}^{\mathsf{H}}(t,f)}{\sum_{t=1}^{T} \text{CM}_{\mathbf{S}}^{\mathsf{H}}(t,f)\text{CM}_{\mathbf{S}}(t,f)} \tag{6}$$

$$= \frac{\sum_{t=1}^{T} (\text{CM}_{\mathbf{S}}(t,f)\mathbf{Y}(t,f))(\text{CM}_{\mathbf{S}}(t,f)\mathbf{Y}(t,f))^{\mathsf{H}}}{\sum_{t=1}^{T} \text{CM}_{\mathbf{S}}^{\mathsf{H}}(t,f)\text{CM}_{\mathbf{S}}(t,f)} \tag{7}$$

where M_S and CM_S 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 $\mathbf{\Phi}$. $\mathbf{\Phi}_{\mathbf{NN}}(f)$ 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,

$$\hat{\mathbf{S}}(t,f) = \mathbf{w}^{\mathsf{H}}(f)\mathbf{Y}(t,f) \tag{8}$$

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

[30]. Inspired by the single channel multi-frame MVDR [26, 27, 29] which utilizes the inter-frame correlation, we propose a multi-tap MVDR for the multi-channel neural beamforming to achieve distortionless speech and low residual noise simultaneously. We define the L-tap representation of the mixture speech as $\bar{\mathbf{Y}}(t,f) = [\mathbf{Y}^{\mathsf{T}}(t,f), \mathbf{Y}^{\mathsf{T}}(t-1,f), ..., \mathbf{Y}^{\mathsf{T}}(t-L+1,f)]^{\mathsf{T}} \in \mathbb{C}^{M\times L}$. The corresponding $\bar{\mathbf{S}}$, $\bar{\mathbf{N}}$, $\overline{\mathbf{CM}}$ can be defined in the same way. Then we can calculate the extended L-tap target speech covariance matrix $\mathbf{\Phi}_{\bar{\mathbf{S}}\bar{\mathbf{S}}}(f) \in \mathbb{C}^{ML\times ML}$ as

$$\mathbf{\Phi}_{\bar{\mathbf{S}}\bar{\mathbf{S}}}(f) = \frac{\sum_{t=1}^{T}(\overline{\mathrm{CM}}_{\mathbf{S}}(t,f)\bar{\mathbf{Y}}(t,f))(\overline{\mathrm{CM}}_{\mathbf{S}}(t,f)\bar{\mathbf{Y}}(t,f))^{\mathsf{H}}}{\sum_{t=1}^{T}\overline{\mathrm{CM}}_{\mathbf{S}}^{\mathsf{H}}(t,f)\overline{\mathrm{CM}}_{\mathbf{S}}(t,f)} \tag{9}$$

Benesty et al. [28] proposed the multi-channel speech enhancement 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

$$\bar{\mathbf{w}}(f) = \frac{\mathbf{\Phi}^{-1}_{\bar{\mathbf{N}}\bar{\mathbf{N}}}(f)\mathbf{\Phi}_{\bar{\mathbf{S}}\bar{\mathbf{S}}}(f)}{\mathrm{Trace}(\mathbf{\Phi}^{-1}_{\bar{\mathbf{N}}\bar{\mathbf{N}}}(f)\mathbf{\Phi}_{\bar{\mathbf{S}}\bar{\mathbf{S}}}(f))}\bar{\mathbf{u}} \quad , \quad \bar{\mathbf{w}}(f) \in \mathbb{C}^{M\times L} \tag{10}$$

where $\bar{\mathbf{u}}$ is an expanded one-hot vector of $\mathbf{u}$ with padding zeros in the tail. Note that the multi-tap MVDR follows the optimization 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,

$$\hat{\mathbf{S}}(t,f) = \bar{\mathbf{w}}^{\mathsf{H}}(f)\bar{\mathbf{Y}}(t,f) \tag{11}$$

The beamformed spectrum is converted to the time-domain waveform via iSTFT. Finally, the Si-SNR loss [11] calculated on the waveform is back-propagated through all of the modules (including the multi-tap MVDR module and the networks) as shown in Fig. 1. Different from the weighted prediction error (WPE) [35] for dereverberation, multi-tap MVDR utilizes the correlation of nearest frames (mainly the early reflection area) and aims only at recovering the reverberant clean speech. However, WPE keeps away from the early reflection area to avoid hurting the dry clean speech for the dereverberation [35].

In summary, the solution Eq. (5) is a complex-valued masking on the single channel. MVDR provides a solution Eq. (8) of complex masking on multiple channels, and our proposed multi-tap MVDR (Eq. (11)) conducts spatio-temporal filtering across frames and channels.

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 important speaker-dependent information, lip movement sequence and the DOA of the target speaker (denoted as θ in Fig. 1). The lip movement has been proven effective for the speech separation in [36, 37, 38, 12, 15]. In this work, we utilize the mouth region RGB pixels to represent the target speaker's lip feature. As shown in Fig. 1, a 3-D residual network [37, 39, 40] is adopted 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 feature (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° 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 calculating the cosine similarity between the target steering vector $v(\theta)$ 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 embeddings [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 ($\leqslant$ 17dB) and multi-face videos [13], resulting in 205500 clean video segments with single face (about 200 hours) over 1500 speakers. The sampling rate for audio and video are 16 kHz and 25 fps respectively. 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 transcript 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] (uncorrelated 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 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 dereverberation in this work. Our systems work well on different scenarios, e.g., different angles between the target speaker and other speakers, various number of overlapped speakers. The scenarios, e.g., small angles ($\leqslant 45^{\circ}$) 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

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

Systems/Metrics	PESQ $\in [-0.5, 4.5]$							PESQ	WER (%)
	Angle between target & others				# of overlapped speakers				
	0-15°	15-45°	45-90°	90-180°	1 SPK	2 SPK	3 SPK	Ave	Ave
Reverberant Clean (reference)	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	6.97%
Mixture (interfering speech + noise)	1.88	1.88	1.98	2.03	3.55	2.02	1.77	2.16	51.30%
ReLU mask (Audio only) on Channel 0 (i)	2.50	2.68	2.88	2.86	3.88	2.81	2.50	2.87	17.89%
ReLU mask (Lip only) on Channel 0 (ii)	2.44	2.52	2.74	2.68	3.86	2.76	2.34	2.76	23.25%
ReLU mask (Baseline) on Channel 0 (iii)	2.56	2.74	2.93	2.89	3.88	2.85	2.56	2.92	17.44%
Complex mask (CM) on Channel 0 (iv)	2.64	2.84	3.00	3.00	**3.89**	2.94	2.66	3.00	16.90%
Sigmoid mask MVDR joint train (JT) (v)	2.27	2.59	2.82	2.73	3.67	2.67	2.37	2.73	15.11%
ReLU mask MVDR JT (vi)	2.52	2.74	2.94	2.85	3.68	2.86	2.54	2.88	12.61%
CM MVDR JT (vii)	2.55	2.77	2.97	2.89	3.73	2.89	2.57	2.91	11.84%
Prop. CM multi-tap MVDR JT (viii)	**2.70**	**3.00**	**3.20**	**3.13**	3.83	**3.10**	**2.76**	**3.10**	**9.96%**

vs 2.92) and lower WER (16.90% vs 17.44%) compared to the ReLU mask baseline (iii). The difference between the ReLU mask and CM is also shown in Fig. 2. There are some spectral "black holes" distortion in the enhanced spectrogram of the ReLU mask baseline (iii). The problem is more severe in the sigmoid mask according to our observations. This type of non-linear spectral distortion is harmful to speech recognition. However, the CM can reduce the distortion and recover the phase simultaneously. Better mask can also help to estimate more accurate covariance matrix in the MVDR beamformer.

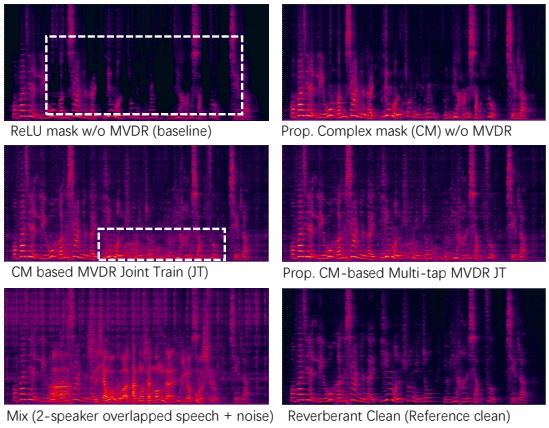

Figure 2: *Separated spectrogram demos of different systems.*

Mask-based MVDR: Although the CM-based network (iv) can reduce the distortion and achieve 3.00 PESQ on average, the ASR performance of 16.9% WER does not match the gain in PESQ. This phenomenon is widely observed in purely neural network based speech enhancement front-ends [7] because the non-linear distortion in the enhanced speech is not ASR friendly. Even for the commercial general-purpose ASR engine, non-distortion is more important than no residual noise, considering that the ASR engine is already robust to the mild level noise (but may not robust to the non-linear distortion). With the distortionless constraint in the MVDR, the beamformed speech can achieve much lower WER. For example, the jointly trained CM-based MVDR (vii) can reduce the WER from 16.9% to 11.84% when compared to the CM-based network w/o MVDR (iv). CM is superior to other real-valued masks (ReLU mask or sigmoid mask) in estimating the target speech and noise covariance matrix. Nonetheless, MVDR beamformer obtains this 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/mtmvdr.

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

[1] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[2] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.

[3] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *Interspeech*, 2013, pp. 436–440.

[4] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *ICASSP*, 2016.

[5] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *ICASSP*, 2017.

[6] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *ICASSP*, 2017.

[7] J. Du, Q. Wang, and et al., "Robust speech recognition with speech enhanced deep neural networks," in *Interspeech*, 2014.

[8] Z.-Q. Wang and D. Wang, "A joint training framework for robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 796–806, 2016.

[9] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "Joint training of front-end and back-end deep neural networks for robust speech recognition," in *ICASSP*, 2015, pp. 4375–4379.

[10] Y. Qian, X. Chang, and D. Yu, "Single-channel multi-talker speech recognition with permutation invariant training," *Speech Communication*, vol. 104, pp. 1–11, 2018.

[11] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[12] J. Wu, Y. Xu, S.-X. Zhang, L.-W. Chen, M. Yu, L. Xie, and D. Yu, "Time domain audio visual speech separation," *ASRU*, 2019.

[13] K. Tan, Y. Xu, S.-X. Zhang, M. Yu, and D. Yu, "Audio-visual speech separation and dereverberation with a two-stage multimodal network," *IEEE Journal of Selected Topics in Signal Processing*, 2020.

[14] F. Bahmaninezhad, J. Wu, R. Gu, S.-X. Zhang, Y. Xu, M. Yu, and D. Yu, "A comprehensive study of speech separation: spectrogram vs waveform separation," *Interspeech*, 2019.

[15] 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*, 2020.

[16] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on audio, speech, and language processing*, vol. 18, no. 2, pp. 260–276, 2009.

[17] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer Science & Business Media, 2008.

[18] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *ICASSP*, 2016.

[19] H. Erdogan, J. R. Hershey, and et al., "Improved MVDR beamforming using single-channel mask prediction networks." in *Interspeech*, 2016.

[20] Y. Xu, C. Weng, L. Hui, J. Liu, M. Yu, D. Su, and D. Yu, "Joint training of complex ratio mask based beamformer and acoustic model for noise robust ASR," in *ICASSP*, 2019.

[21] A. S. Subramanian, C. Weng, M. Yu, S.-X. Zhang, Y. Xu, S. Watanabe, and D. Yu, "Far-field location guided target speech extraction using end-to-end speech recognition objectives," in *ICASSP*, 2020.

[22] Z.-Q. Wang and D. Wang, "Mask weighted STFT ratios for relative transfer function estimation and its application to robust ASR," in *ICASSP*, 2018.

[23] C. Boeddeker, H. Erdogan, T. Yoshioka, and R. Haeb-Umbach, "Exploring practical aspects of neural mask-based beamforming for far-field speech recognition," in *ICASSP*, 2018.

[24] X. Xiao, S. Zhao, D. L. Jones, E. S. Chng, and H. Li, "On time-frequency mask estimation for MVDR beamforming with application in robust speech recognition," in *ICASSP*, 2017.

[25] A. Schasse and R. Martin, "Estimation of subband speech correlations for noise reduction via MVDR processing," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 9, pp. 1355–1365, 2014.

[26] Y. A. Huang and J. Benesty, "A multi-frame approach to the frequency-domain single-channel noise reduction problem," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1256–1269, 2011.

[27] M. Tammen, D. Fischer, and S. Doclo, "DNN-based multi-frame MVDR filtering for single-microphone speech enhancement," *arXiv preprint arXiv:1905.08492*, 2019.

[28] J. Benesty, J. Chen, and E. A. Habets, *Speech enhancement in the STFT domain*. Springer Science & Business Media, 2011.

[29] D. Fischer and S. Doclo, "Sensitivity analysis of the multi-frame MVDR filter for single-microphone speech enhancement," in *EUSIPCO*, 2017, pp. 603–607.

[30] E. A. Habets and J. Benesty, "A two-stage beamforming approach for noise reduction and dereverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 945–958, 2013.

[31] A. S. Subramanian, X. Wang, S. Watanabe, T. Taniguchi, D. Tran, and Y. Fujita, "An investigation of end-to-end multichannel speech recognition for reverberant and mismatch conditions," *arXiv preprint arXiv:1904.09049*, 2019.

[32] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[33] J. Yu, B. Wu, and et al., "Audio-visual multi-channel recognition of overlapped speech," *arXiv preprint arXiv:2005.08571*, 2020.

[34] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, pp. 483–492, 2015.

[35] K. Kinoshita, M. Delcroix, H. Kwon, T. Mori, and T. Nakatani, "Neural network-based spectrum estimation for online WPE dereverberation." in *Interspeech*, 2017.

[36] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *SIGGRAPH*, 2018.

[37] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," *Interspeech*, 2018.

[38] G. Morrone, S. Bergamaschi, L. Pasa, L. Fadiga, V. Tikhanoff, and L. Badino, "Face landmark-based speaker-independent audio-visual speech enhancement in multi-talker environments," in *ICASSP*, 2019.

[39] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Asian Conference on Computer Vision*, 2016.

[40] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with LSTMs for lipreading," *arXiv preprint arXiv:1703.04105*, 2017.

[41] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li, and Y. Gong, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *2018 IEEE SLT*, 2018.

[42] E. A. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, 2006.

[43] https://speech.yitutech.com.