Multi-channel Multi-frame ADL-MVDR for Target Speech Separation

Zhuohuang Zhang, *Student Member*, *IEEE*, Yong Xu, Meng Yu, Shi-Xiong Zhang, Lianwu Chen, Donald S. Williamson, *Member*, *IEEE*, Dong Yu, *Fellow*, *IEEE*

***Abstract*—Many purely neural network based speech separation approaches have been proposed that greatly improve objective assessment scores, but they often introduce nonlinear distortions that are harmful to automatic speech recognition (ASR). Minimum variance distortionless response (MVDR) filters strive to remove nonlinear distortions, however, these approaches either are not optimal for removing residual (linear) noise, or they are unstable when used jointly with neural networks. In this study, we propose a multi-channel multi-frame (MCMF) all deep learning (ADL)-MVDR approach for target speech separation, which extends our preliminary multi-channel ADL-MVDR approach. The MCMF ADL-MVDR handles different numbers of microphone channels in one framework, where it addresses linear and nonlinear distortions. Spatio-temporal cross correlations are also fully utilized in the proposed approach. The proposed system is evaluated using a Mandarin audio-visual corpora and is compared with several state-of-the-art approaches. Experimental results demonstrate the superiority of our proposed framework under different scenarios and across several objective evaluation metrics, including ASR performance.**

***Index Terms*—Speech separation, deep learning, MVDR, ADL-MVDR.**

I. Introduction

SPEECH separation algorithms extract a target speech signal in adverse acoustic environments where multiple sources of interfering speakers and noise exist. These algorithms serve as important front-ends for many speech communication systems such as automatic speech recognition (ASR) [1], [2], [3], [4], speaker verification [5], and digital hearing-aid devices [6]. With the recent achievements in deep learning, many neural network (NN) based speech separation systems have been proposed. Many early approaches synthesize the separated speech after combining the time-frequency (T-F) masked spectrogram with the original noisy phase [7], [8], [9]. The use of the noisy phase sets a sub-optimal upper bound on the system's performance as phase plays an important role in the perceptual speech quality and intelligibility [10], [11], [12], [13]. Phase-aware T-F masks have later been proposed, including the phase-sensitive mask [14], [15], [9], and complex ratio mask [11]. Yet an accurate estimate of the phase component is still difficult for a NN to learn, due to the lack of structure in the phase response.

Besides T-F mask based systems, many recent speech separation systems have been proposed that operate directly on the time-domain speech signal in an end-to-end fashion [16], [17], [18], [19], [20], [21], to avoid directly estimating the magnitude and phase components. Some of these approaches (e.g., Wave-U-Net [20] and TasNet [18]) replace the conventional STFT and inverse STFT (iSTFT) signal processing procedures with a learnable NN-based encoder and decoder structure. The encoded features are then altered by a learned-latent mask, where they are later fed to the decoder. The recent time-domain fully-convolutional Conv-TasNet [21] has substantially improved performance according to many objective measures, where it features a TasNet-like encoder-decoder structure that extracts the target speech in a learned latent space [18]. Alternatively, other approaches implicitly combine the feature extraction and the separation steps as reported in [16], [19].

Purely NN-based speech separation systems have achieved impressive objective speech quality scores, since they greatly reduce the amount of noise or interfering speech. These approaches, however, often introduce unwanted nonlinear distortions into the separated signal, since these models focus on removing unwanted interferences without imposed constraints that limit the solution space. These resulting nonlinear distortions negatively affect the performance of ASR systems [22], [21], [23]. Many approaches have been developed to reduce nonlinear distortions, including the multi-channel Wiener filter (MWF) [24], [25], the linearly constrained minimum variance (LCMV) filter [26] and the minimum variance distortionless response (MVDR) filter. The MVDR filter can be viewed as a special case of MWF and LCMV, as it forces a distortionless response when oracle directional information is available [27], [28], [29], [30]. MVDR filters have been widely used in speech separation systems to reduce the amount of nonlinear distortions, which is helpful to ASR systems [31], [32], [22]. The distortionlessness of the separated speech is ensured as the MVDR filter is derived under constraints that preserve speech information at the target direction. On the contrary, other beamformers such as the Generalized Eigenvalue (GEV) beamformer [33], [34] aims to improve the signal-to-noise ratio (SNR) without controlling the amount of distortions in the separated speech signal. Additionally, multi-frame MVDR (MF-MVDR) filters [35], [36], [37] have been adopted in single-channel speech separation systems to remove the noise

This work was done while Z. Zhang was a research intern at Tencent AI Lab, Bellevue, USA.

Zhuohuang Zhang is with the Department of Computer Science and Department of Speech, Language and Hearing Sciences, Indiana University, Bloomington, IN, 47408 USA (e-mail: zhuozhan@iu.edu).

Yong Xu, Meng Yu, Shi-Xiong Zhang, and Dong Yu are with the Tencent AI Lab, Bellevue, WA 98004 USA (e-mail: lucayongxu@tencent.com; raymondmyu@tencent.com; auszhang@tencent.com; dyu@tencent.com).

Lianwu Chen is with Tencent AI Lab, Shenzhen, China (e-mail: lianwuchen@tencent.com).

Donald S. Williamson is with the Department of Computer Science, Indiana University, Bloomington, IN, 47408 USA (e-mail: williads@indiana.edu).

and ensure the distortionlessness of the separated speech. Prior studies have shown that when oracle information is available, the MF-MVDR filter can greatly diminish the noise while introducing few distortions [35], [38].

Recent MVDR approaches are often combined with a NN-based T-F mask estimator [32], [39], [40], [22] that leads to more accurate estimates of the speech and noise components, and better ASR performance due to fewer nonlinear distortions. However, many of these conventional mask-based MVDR systems result in high levels of residual noise (e.g., linear distortions), since segment- or utterance-level beamforming weights are not optimal for noise reduction [8], [32], [22]. More recently, frame-level beamforming weights estimation approaches have been proposed to address the residual noise issue. Souden et al. [41] proposed a recursive method with heuristic updating factors to estimate the time-varying speech and noise covariance matrices, but these heuristic updating factors are hard to determine and often limit the system's performance. The calculated beamforming weights are also numerically unstable when jointly trained with NNs. Our recent work further incorporates multi-frame (MF) information during beamforming weight derivation [22], where it leads to better ASR accuracy and higher PESQ scores when compared to conventional mask-based MVDR approaches. Unfortunately, the amount of residual noise in the separated signal is still high.

In the current study, we propose a generalized all deep learning MVDR (ADL-MVDR) framework that is capable of performing speech separation under different microphone settings, including multi-channel (single-frame), multi-frame (i.e., when only one channel is available), and multi-channel multi-frame (MCMF) scenarios. This study extends our preliminary work on the ADL-MVDR beamformer [42], which has proven to work well on multi-channel (MC) speech separation tasks. The ADL-MVDR beamformer incorporates a front-end complex filter estimator (i.e., a Conv-TasNet variant based on our prior work [23], [43]) that consists of dilated 1D-convolution blocks for speech and noise component estimation and another ADL-MVDR module for frame-level MVDR beamforming weights estimation. In contrast to conventional per T-F bin mask-based approaches, complex ratio filtering (denoted as cRF) [44] is used for more accurate estimates of the speech and noise components, while also addressing issues with handling phase. Earlier approaches have verified the idea of applying NNs for matrix inversion [45], [46], [47], [48] and principal component analysis (PCA) [49], [47]. The proposed ADL-MVDR module deploys two recurrent neural networks (RNNs) to separately perform matrix inversion and PCA on the noise and speech covariance matrices, respectively. Leveraging on the temporal properties of RNNs, the statistical variables (i.e., inverse of noise covariance matrix and steering vector) are estimated adaptively at the frame-level, enabling the derivation of time-varying beamforming weights, which is more suitable for diminishing the residual noise at each frame. The system also uses visual information (described in our prior work [23], [43]) to extract the direction of arrival (DOA) of the target speaker. Results from our prior study [42] indicate that for MC speech separation tasks, the ADL-MVDR system can greatly suppress the residual noise while also ensuring that fewer distortions are introduced into the separated speech signal when compared to conventional mask-based MVDR approaches.

The major contributions of this work consist of the following. Firstly, we verify the idea of applying the ADL-MVDR framework to MF speech separation tasks, when only one channel of the signal is available, to further evaluate generalization. Secondly, we further adapt the ADL-MVDR framework to a MCMF speech separation task for spatio-temporal speech separation, which has not been previously done, to determine if additional MF information leads to further improvements. Thirdly, we examine and quantify the influence of the cRF and MF sizes on the performance of different ADL-MVDR systems.

The rest of this paper is organized as follows. Section II describes the signal models for conventional mask-based MC and MF-MVDR systems. The proposed ADL-MVDR system is revealed in Section III. The experimental setup is given in Section IV. We present and discuss the results in Section V. Finally, we conclude our work in Section VI.

II. Conventional Mask-based MVDR Filter

In this section, we discuss the signal models of conventional mask-based MVDR filters under two different conditions, i.e., MC and MF-MVDR speech separation.

A. Conventional Mask-based Multi-channel MVDR

In the MC speech separation scenario, consider a time-domain noisy speech signal $\mathbf{y} = [y^{(0)}, y^{(1)}, ..., y^{(M-1)}]^T$ recorded by an M-channel microphone array, where $y^{(i)}$ is the signal recorded from the i-th channel. Let $\mathbf{Y}(t,f)$, $\mathbf{X}(t,f)$ and $\mathbf{N}(t,f)$ denote the T-F domain MC noisy-reverberant speech, reverberant speech and noise signals, respectively. We have

$$\mathbf{Y}(t,f) = \mathbf{X}(t,f) + \mathbf{N}(t,f), \tag{1}$$

where (t,f) represents the corresponding frame and frequency index. Note that we use reverberant speech as the learning target as we focus on separation only in this study. In the time-domain, we have $\mathbf{x}(t) = \mathbf{g}(t) * \mathbf{s}(t)$, where $\mathbf{x}(t)$ and $\mathbf{s}(t)$ are the time-domain MC reverberant and anechoic speech signals, $\mathbf{g}(t)$ represents the room impulse response and '$*$' denotes linear convolution.

The estimated single-channel of the reverberant speech, $\hat{\mathrm{X}}^{(0)}(t,f)$, can be obtained with the MC-MVDR filter as

$$\hat{\mathrm{X}}^{(0)}(t,f) = \mathbf{h}^{\mathrm{H}}_{\text{MC-MVDR}}(f)\mathbf{Y}(t,f), \tag{2}$$

where $\mathbf{h}_{\text{MC-MVDR}} \in \mathbb{C}^M$ denotes the MC-MVDR beamforming weights, and $^{\mathrm{H}}$ is the Hermitian operator. The objective of MC-MVDR filtering is to minimize the power of the noise without introducing distortions into the target speech signal. This can be formulated as

$$\mathbf{h}_{\text{MC-MVDR}} = \arg\min_{\mathbf{h}} \mathbf{h}^{\mathrm{H}}\mathbf{\Phi_{NN}}\mathbf{h} \quad \text{s.t.} \quad \mathbf{h}^{\mathrm{H}}\boldsymbol{v} = \mathbf{1}, \tag{3}$$

where $\boldsymbol{v}(f) \in \mathbb{C}^M$ stands for the target speech steering vector and can be derived by applying PCA on the speech covariance

matrix, i.e., $\boldsymbol{v}(f) = \text{PCA}(\mathbf{\Phi_{XX}})$. $\mathbf{\Phi_{NN}}$ represents the noise covariance matrix. Soving Eq. (3), the MC-MVDR filter can be derived as [50], [51]

$$\mathbf{h}_{\text{MC-MVDR}}(f) = \frac{\mathbf{\Phi}_{\mathbf{NN}}^{-1}(f)\boldsymbol{v}(f)}{\boldsymbol{v}^{\text{H}}(f)\mathbf{\Phi}_{\mathbf{NN}}^{-1}(f)\boldsymbol{v}(f)}. \tag{4}$$

Note that the matrix inversion and PCA process for steering vector derivation are not stable during NN joint training. For example, the estimated noise covariance matrix could be singular, which will cause numerical instability when computing the matrix inverse.

In a typical mask-based MVDR system, the covariance matrices are estimated utterance-wise with T-F masks [52], [8], [40], [53], [22]. A system that uses a real-valued T-F mask (RM) (e.g., with an ideal binary mask (IBM), ideal ratio mask (IRM), ...) performs covariance matrix estimation as follows

$$\hat{\mathbf{\Phi}}_{\mathbf{XX}}(f) = \frac{\sum_{t=1}^{T} \text{RM}_{\text{X}}^2(t,f)\mathbf{Y}(t,f)\mathbf{Y}^{\text{H}}(t,f)}{\sum_{t=1}^{T} \text{RM}_{\text{X}}^2(t,f)}, \tag{5}$$

where RM_{X} stands for the RM for estimating the speech component and T is the total number of frames. The noise covariance matrix $\hat{\mathbf{\Phi}}_{\mathbf{NN}}$ can be computed in a similar manner. Nevertheless, we want to point out that the utterance-wise covariance matrices are not optimal for noise reduction on each frame and therefore the high level of residual noise becomes a hand-in-hand problem for conventional mask-based MC-MVDR systems.

B. Conventional Multi-frame MVDR

The MF-MVDR filter can be viewed as a special extension of the MC-MVDR beamformer when only one channel of the signal is available. In this case, the spatial information is lost and therefore, the MF-MVDR filter tries to explore the inter-frame correlations instead of performing spatial beamforming. Many MF-MVDR systems have been proposed recently [35], [54], [40]. Analogous to the MC-MVDR speech separation scenario, the process of obtaining the MF-MVDR enhanced speech can be described as

$$\hat{\text{X}}^{(0)}(t,f) = \mathbf{h}_{\text{MF-MVDR}}^{\text{H}}(t,f)\overline{\mathbf{Y}}^{(\mathbf{0})}(t,f), \tag{6}$$

where $\mathbf{h}_{\text{MF-MVDR}} \in \mathbb{C}^L$ and $\overline{\mathbf{Y}}^{(\mathbf{0})}$ represent the L-dimensional MF-MVDR filter coefficients and L consecutive STFT frames of the single-channel noisy speech signal, respectively,

$$\begin{aligned} \mathbf{h}_{\text{MF-MVDR}}(t,f) &= [\text{h}_0(t,f), \text{h}_1(t,f), ..., \text{h}_{L-1}(t,f)]^T, \\ \overline{\mathbf{Y}}^{(\mathbf{0})}(t,f) &= [\text{Y}^{(0)}(t,f), \text{Y}^{(0)}(t-1,f), ..., \\ &\quad \text{Y}^{(0)}(t-L+1,f)]^T, \end{aligned} \tag{7}$$

where $h_l(t,f)$ represents the l-th filter coefficient and $\text{Y}^{(0)}(t,f)$ is the single-channel noisy speech STFT. This is similar to MC beamforming methods by viewing different frames as microphone inputs of different channels. The MF speech and noise components are defined in a similar way. The objective of the MF-MVDR filter is also to minimize the power of the interfering sources while preserving the components from the target speech, which can be computed as

$$\mathbf{h}_{\text{MF-MVDR}} = \arg\min_{\overline{\mathbf{h}}} \overline{\mathbf{h}}^{\text{H}}\mathbf{\Phi}_{\mathbf{VV}}\overline{\mathbf{h}} \quad \text{s.t.} \quad \overline{\mathbf{h}}^{\text{H}}\boldsymbol{\gamma}_{\mathbf{x}} = \mathbf{1}, \tag{8}$$

where $\mathbf{\Phi_{VV}}$ denotes the covariance matrix of the MF undesired signal component [35], [36], [40] which consists of the noise and the uncorrelated speech components. $\boldsymbol{\gamma}_{\mathbf{x}}$ is the speech IFC vector that describes the correlation between the previous and current frame. According to [54], [40], the speech IFC vector $\boldsymbol{\gamma}_{\mathbf{x}}$ can be formulated as

$$\boldsymbol{\gamma}_{\mathbf{x}}(t,f) = \frac{\mathbf{\Phi_{XX}}(t,f)\mathbf{e}}{E[|\text{X}^{(0)}(t,f)|^2]}, \tag{9}$$

where $\mathbf{\Phi_{XX}}$ in this context stands for the covariance matrix of the MF speech and $\text{X}^{(0)}(t,f)$ represents the single-channel speech. $\mathbf{e}$ is a vector selecting the first column of the speech covariance matrix and $E[\cdot]$ denotes mathematical expectation.

Solving Eq. (8), the MF-MVDR filter vector can be obtained as [35], [54], [40]

$$\mathbf{h}_{\text{MF-MVDR}}(t,f) = \frac{\mathbf{\Phi}_{\mathbf{VV}}^{-1}(t,f)\boldsymbol{\gamma}_{\mathbf{x}}(t,f)}{\boldsymbol{\gamma}_{\mathbf{x}}^{\text{H}}(t,f)\mathbf{\Phi}_{\mathbf{VV}}^{-1}(t,f)\boldsymbol{\gamma}_{\mathbf{x}}(t,f)}. \tag{10}$$

Note that in [40], $\mathbf{\Phi_{VV}}$ was replaced by the MF noise covariance matrix $\mathbf{\Phi_{NN}}$ under the assumption that the uncorrelated speech component is negligible, which imposes an upper bound on the system's performance.

III. Proposed ADL-MVDR Beamformer

As mentioned in previous sections, the covariance matrices and steering vector for most of the conventional mask-based MC-MVDR systems are computed at the utterance-level that discards the temporal information. This results in utterance-level beamforming weights which are not optimal for noise reduction at each frame. Additionally, the matrix inversion and PCA steps involved in conventional mask-based MVDR systems are not numerically stable when jointly trained with NNs. In some classical MVDR approaches, the derivations of the steering or IFC vectors and covariance matrices are based on recursive methods which requires heuristic updating factors between consecutive frames [41], [55], [40]. However, these factors are usually hard to determine and could easily influence the accuracy of the estimated terms.

In this study, we propose a novel ADL-MVDR framework that can be generalized to many different settings, including MC beamforming, MF filtering and MCMF beamforming. The key point of our proposed ADL-MVDR framework is using two separate gated recurrent unit (GRU) based networks (denoted as GRU-Nets) to estimate the inverse of the noise/undesired signal covariance matrix and steering/IFC vector at frame-level. Leveraging on the temporal properties of RNNs, the GRU-Nets can better explore and utilize the temporal information from previous frames without any needs of heuristic updating factors. Estimating MVDR coefficients via GRU-Nets also resolves the instability issue when the MVDR beamformer is jointly trained with NNs. Note that previous approaches that adopt NNs to directly learn the beamforming filtering weights [56], [57] are not successful since noise information is not explicitly considered, whereas our proposed ADL-MVDR beamformer explicitly utilizes the cross-channel information from both estimated speech and noise covariance matrices.

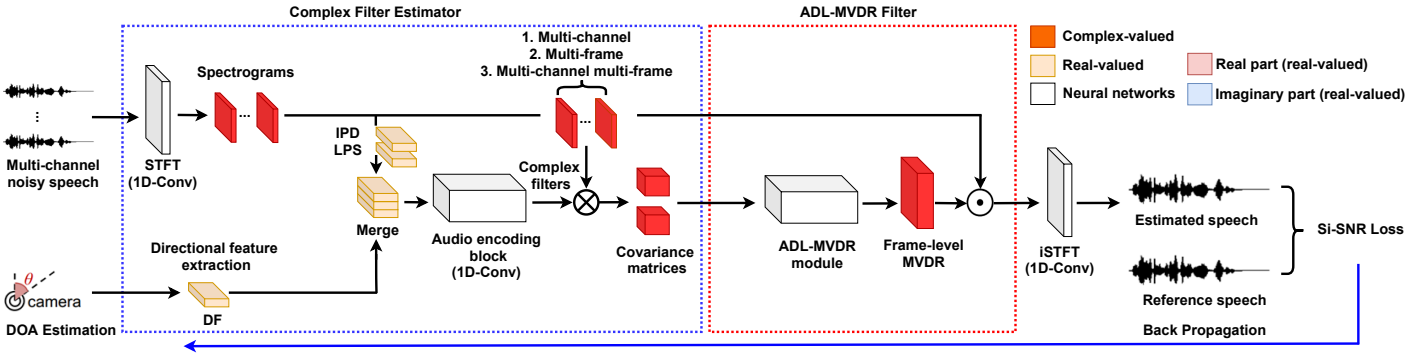

Fig. 1. (Color Online). Network architecture of our proposed ADL-MVDR framework. It consists of a complex filter estimator (i.e., highlighted in blue dashed box) for components estimation (depending on the case), and an ADL-MVDR beamformer (i.e., highlighted in red dashed box) that consists of two GRU-Nets for frame-wise MVDR coefficients estimation. $\otimes$ and $\odot$ denote the operations expressed in Eq. (12) and (16), (19) or (21), respectively, depending on the situation. The entire system is jointly trained with time-domain scale-invariant source-to-noise ratio (Si-SNR) loss.

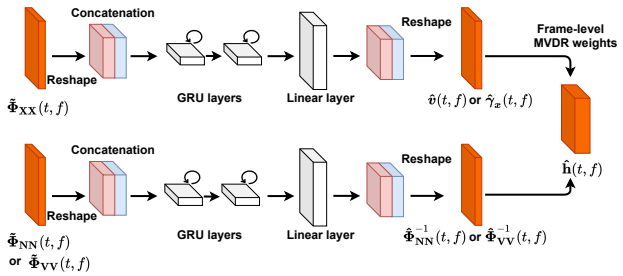

Fig. 2. (Color Online). Detailed network architecture of the ADL-MVDR module. The real and imaginary parts of the covariance matrices are concatenated before fed into the GRU-Nets. The estimated MVDR coefficients are reshaped back to their original forms before computing the frame-level MVDR weights.

A. System Overview

The general framework of our proposed ADL-MVDR beamformer is depicted in Fig. 1. The system consists of two parts, a complex filter estimator that is based on our previously proposed multi-modal MC speech separation platform [23], [43] (i.e., a Conv-TasNet [21] variant) for frame-level covariance matrices estimation, followed by another ADL-MVDR module (depicted in Fig. 2) for frame-level MVDR weights derivation. As described in our previous works [43], [23], inside the complex filter estimator, the interaural phase difference (IPD) and log-power spectra (LPS) features are extracted from the MC noisy speech and further merged with the directional feature (DF) as input to the audio encoding block that comprises a set of dilated 1D-convolution blocks. A 180° wide view camera is used to extract the DOA according to the location of the target speaker's face, which is further used to extract the DF following the method developed in [58].

Our recent work [22] suggests that the complex ratio mask (denoted as cRM) can lead to better estimates of the covariance matrices, the procedure of using cRM to derive the covariance matrix is described below as

$$\begin{aligned}\hat{\mathbf{X}}_{\mathbf{cRM}}(t,f) &= (\mathrm{cRM}_{\mathrm{r}} + j\mathrm{cRM}_{\mathrm{i}}) \cdot (\mathbf{Y}_{\mathrm{r}} + j\mathbf{Y}_{\mathrm{i}}) \\ &= \mathrm{cRM}_{\mathrm{X}}(t,f) \cdot \mathbf{Y}(t,f), \\ \hat{\mathbf{\Phi}}_{\mathbf{XX}}(f) &= \frac{\sum_{t=1}^{T} \hat{\mathbf{X}}_{\mathbf{cRM}}(t,f)\hat{\mathbf{X}}_{\mathbf{cRM}}^{\mathrm{H}}(t,f)}{\sum_{t=1}^{T} \mathrm{cRM}_{\mathrm{X}}^{\mathrm{H}}(t,f)\mathrm{cRM}_{\mathrm{X}}(t,f)},\end{aligned} \tag{11}$$

where $\hat{\mathbf{X}}_{\mathbf{cRM}}(t,f)$ represents the estimated MC speech component via the complex speech mask $\mathrm{cRM}_{\mathrm{X}}$. r and i denote the real and imaginary parts, respectively. '$\cdot$' is the complex multiplier and j is the complex number.

In this study, different from prior mask-based MVDR approaches that use T-F masks (e.g., ideal ratio mask (IRM), cRM, etc.) to estimate the speech and noise/undesired signal components, we adopt the cRF [44] method for estimation. As depicted in Fig. 3, the cRF differs from the cRM that instead of using one-to-one mapping, it utilizes the nearby T-F bins to estimate each target T-F bin. The example shown in Fig. 3 can be formulated as

$$\begin{aligned}\hat{\mathbf{X}}_{\mathbf{cRF}}(t,f) = \sum_{\tau_1=-J_1}^{J_2} \sum_{\tau_2=-K_1}^{K_2} \mathrm{cRF}(t+\tau_1, f+\tau_2) \\ \cdot \mathbf{Y}(t+\tau_1, f+\tau_2),\end{aligned} \tag{12}$$

where $\hat{\mathbf{X}}_{\mathbf{cRF}}$ is the estimated MC speech component using the cRF method, the cRF has the size of $(J_2 + J_1 + 1) \times (K_2 + K_1 + 1)$. J_1, J_2 and K_1, K_2 represent the number of previous and future frames and frequency bins used for filtering, respectively. The noise or the undesired signal components can be obtained in a similar manner. Different from [44], where the authors directly apply cRF for speech separation, we adopt cRF to estimate the speech and noise covariance matrices which are further used as inputs to the ADL-MVDR module. The cRF also plays an important role in the success of our ADL-MVDR beamformer, which will be illustrated afterwards in the ablation study.

B. Multi-channel ADL-MVDR

An accurate estimate of the steering vector is very important for a MC-MVDR system as it contains information about which direction the signal should be preserved [59], [60]. Yet previous approaches involving PCA on the speech covariance matrix are often numerically unstable when jointly trained with NNs. A similar issue exists for the inversion process of the noise covariance matrix. In order to estimate the frame-wise steering vector and inverse of the noise covariance matrix accurately and stably, we deploy two GRU-Nets to estimate the steering vector and the inverse of the noise covariance

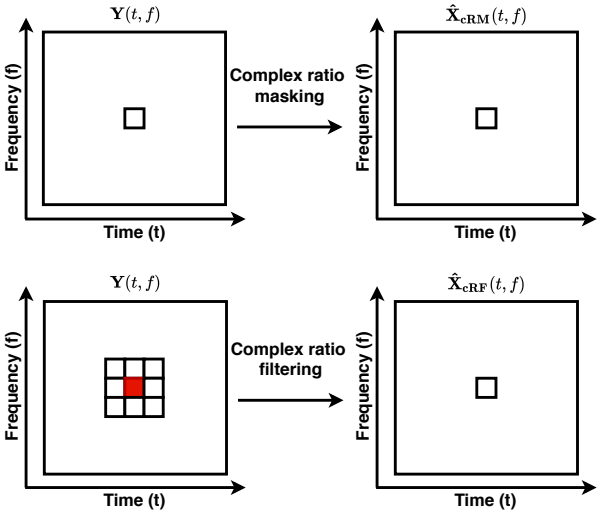

Fig. 3. (Color Online). Schematic diagrams for complex ratio masking and complex ratio filtering. The cRM is a one-to-one mapping, whereas the cRF is a many-to-one mapping. In this example, the cRF has a size of 3×3, where each T-F bin is estimated using 9 T-F bins in its neighboring filter range. The center mask is marked with a red fill.

matrix. Note that we concatenate the real and imaginary parts of the NN estimated time-varying covariance matrices (i.e., $\tilde{\mathbf{\Phi}}_{\mathbf{XX}}$ and $\tilde{\mathbf{\Phi}}_{\mathbf{NN}}$) before feeding them into the GRU-Nets, as shown in Fig. 2. The frequency index f is omitted for the rest of this paper for simplicity since all frequency bins are treated individually.

$$\begin{aligned}\hat{\boldsymbol{v}}(t) &= \mathbf{GRU\text{-}Net}_{\boldsymbol{v}}(\tilde{\mathbf{\Phi}}_{\mathbf{XX}}(t)),\\ \hat{\mathbf{\Phi}}^{-1}_{\mathbf{NN}}(t) &= \mathbf{GRU\text{-}Net}_{\mathbf{NN}}(\tilde{\mathbf{\Phi}}_{\mathbf{NN}}(t)).\end{aligned} \tag{13}$$

Leveraging on the temporal properties of RNNs, the frame-wise covariance matrices are fed into the GRU-Nets for MVDR coefficient estimation. Each frame of the estimated MVDR coefficients (i.e., steering vector and inverse of the noise covariance matrix) is based on its weighted previous frames. Without using any arbitrary inter-frame updating factors, the GRU-Nets can learn the temporal dependencies through the NN training process. The input time-varying speech and noise covariance matrices can be obtained from the cRF estimated speech and noise components as

$$\begin{aligned}\tilde{\mathbf{\Phi}}_{\mathbf{XX}}(t) &= \frac{\hat{\mathbf{X}}_{\mathbf{cRF}}(t)\hat{\mathbf{X}}^{\mathrm{H}}_{\mathbf{cRF}}(t)}{\sum_{t=1}^{T}\mathrm{cRM}^{\mathrm{H}}_{\mathrm{X}}(t)\mathrm{cRM}_{\mathrm{X}}(t)},\\ \tilde{\mathbf{\Phi}}_{\mathbf{NN}}(t) &= \frac{\hat{\mathbf{N}}_{\mathbf{cRF}}(t)\hat{\mathbf{N}}^{\mathrm{H}}_{\mathbf{cRF}}(t)}{\sum_{t=1}^{T}\mathrm{cRM}^{\mathrm{H}}_{\mathrm{N}}(t)\mathrm{cRM}_{\mathrm{N}}(t)},\end{aligned} \tag{14}$$

where $\hat{\mathbf{N}}_{\mathbf{cRF}}$ and $\mathrm{cRM}_{\mathrm{N}}$ denote the estimated MC noise component using the cRF method and the center complex mask of the noise cRF (as depicted in Fig. 3) that is used for normalization, respectively. The same notation also applies to $\mathrm{cRM}_{\mathrm{X}}$. Different from Eq. (11), we do not sum over the temporal dimension to preserve the frame-level information in the covariance matrices.

Based on these RNN-derived frame-wise MVDR coefficients, the MC MVDR beamforming weights can be derived at the frame-level as

$$\hat{\mathbf{h}}_{\text{MC ADL-MVDR}}(t) = \frac{\hat{\mathbf{\Phi}}^{-1}_{\mathbf{NN}}(t)\hat{\boldsymbol{v}}(t)}{\hat{\boldsymbol{v}}(t)^{\mathrm{H}}\hat{\mathbf{\Phi}}^{-1}_{\mathbf{NN}}(t)\hat{\boldsymbol{v}}(t)}, \tag{15}$$

where $\hat{\mathbf{h}}_{\text{MC ADL-MVDR}}(t) \in \mathbb{C}^M$ is the frame-wise MC ADL-MVDR beamforming weights which are different from the utterance-level weights derived in many conventional mask-based MC-MVDR systems. Finally, the MC ADL-MVDR enhanced speech is obtained as

$$\hat{\mathrm{X}}^{(0)}_{\text{MC ADL-MVDR}}(t) = \hat{\mathbf{h}}^{\mathrm{H}}_{\text{MC ADL-MVDR}}(t)\mathbf{Y}(t). \tag{16}$$

C. Multi-frame ADL-MVDR

We also introduce a MF setup for our ADL-MVDR framework to simulate an extreme condition when only one channel of the signal is available from the front-end complex filter estimator. Note that the MF ADL-MVDR system still uses the MC noisy speech as inputs to estimate the cRFs, however, we only use one channel of the signal as the inputs to the ADL-MVDR module. Without a loss of generality, the front-end complex filter estimator could be replaced by any other speech separation systems. The network architecture of the MF ADL-MVDR system is analogous to the MC ADL-MVDR, however, the purpose for each step in this case is very different. Since the spatial information from the microphone array is no longer available, the MF-MVDR explores the correlation information between consecutive frames instead. An accurate estimate on the speech IFC vector dominates the final performance of the system. Previous recursive estimation approaches based on heuristic updating factors are not optimal nor stable when jointly trained with NNs.

Similar to the MC case, the time-varying MF speech covariance matrix can be derived based on the estimated MF speech component. They are defined as

$$\begin{aligned}\hat{\overline{\mathbf{X}}}^{(\mathbf{0})}_{\mathbf{cRF}}(t) &= [\hat{\mathrm{X}}^{(0)}_{\mathrm{cRF}}(t), \hat{\mathrm{X}}^{(0)}_{\mathrm{cRF}}(t-1), ...,\\ &\quad \hat{\mathrm{X}}^{(0)}_{\mathrm{cRF}}(t-L+1)]^{T},\\ \tilde{\Phi}_{\mathrm{XX}}(t) &= \frac{\hat{\mathrm{X}}^{(0)}_{\mathrm{cRF}}(t)\hat{\mathrm{X}}^{(0)\mathrm{H}}_{\mathrm{cRF}}(t)}{\sum_{t=1}^{T}\mathrm{cRM}^{\mathrm{H}}_{\mathrm{X}}(t)\,\mathrm{cRM}_{\mathrm{X}}(t)},\end{aligned} \tag{17}$$

where $\hat{\overline{\mathbf{X}}}^{(\mathbf{0})}_{\mathbf{cRF}}$ is the estimated L-frame single-channel speech component in T-F domain using the cRF method and $\hat{\mathrm{X}}^{(0)}_{\mathrm{cRF}}$ is the estimated single-channel speech component via cRF. The time-varying covariance matrix of the undesired signal $\tilde{\mathbf{\Phi}}_{\mathbf{VV}}$ can be estimated in a similar way using cRF method. Note that the undesired signal component is estimated implicitly as the neural network can gradually learn the mapping during training.

In the MF ADL-MVDR system, two GRU-Nets are implemented to estimate the speech IFC vector and the inverse of the undesired signal covariance matrix. The inputs to these two networks are the time-varying MF speech and undesired

signal covariance matrices estimated via the cRF method. This is formulated as

$$\begin{aligned}\hat{\boldsymbol{\gamma}}_{\mathbf{x}}(t) &= \textbf{GRU-Net}_{\boldsymbol{\gamma}}(\tilde{\boldsymbol{\Phi}}_{\mathbf{XX}}(t)),\\ \hat{\boldsymbol{\Phi}}_{\mathbf{VV}}^{-1}(t) &= \textbf{GRU-Net}_{\mathbf{VV}}(\tilde{\boldsymbol{\Phi}}_{\mathbf{VV}}(t)).\end{aligned} \tag{18}$$

Once these variables are obtained, the MF ADL-MVDR filter weights $\hat{\mathbf{h}}_{\text{MF ADL-MVDR}}(t) \in \mathbb{C}^{L}$ and estimated speech are obtained as

$$\begin{aligned}\hat{\mathbf{h}}_{\text{MF ADL-MVDR}}(t) &= \frac{\hat{\boldsymbol{\Phi}}_{\mathbf{VV}}^{-1}(t)\hat{\boldsymbol{\gamma}}_{\mathbf{x}}(t)}{\hat{\boldsymbol{\gamma}}_{x}^{\mathrm{H}}(t)\hat{\boldsymbol{\Phi}}_{\mathbf{VV}}^{-1}(t)\hat{\boldsymbol{\gamma}}_{\mathbf{x}}(t)},\\ \hat{\mathrm{X}}_{\text{MF ADL-MVDR}}^{(0)}(t) &= \hat{\mathbf{h}}_{\text{MF ADL-MVDR}}^{\mathrm{H}}(t)\overline{\mathbf{Y}}^{(\mathbf{0})}(t).\end{aligned} \tag{19}$$

D. Multi-channel Multi-frame ADL-MVDR

The MCMF ADL-MVDR combines both the MC and MF information and uses them as inputs to the ADL-MVDR module. Let $\hat{\mathbf{X}}_{\mathbf{cRF}}$ denote the MC speech estimated using the cRF method, then the MCMF speech and its time-varying covariance matrix are obtained as

$$\begin{aligned}\overline{\hat{\mathbf{X}}}_{\mathbf{cRF}}(t) &= [\hat{\mathbf{X}}_{\mathbf{cRF}}(t), \hat{\mathbf{X}}_{\mathbf{cRF}}(t-1), ..., \hat{\mathbf{X}}_{\mathbf{cRF}}(t-L+1)]^{T},\\ \tilde{\boldsymbol{\Phi}}_{\mathbf{XX}}(t) &= \frac{\overline{\hat{\mathbf{X}}}_{\mathbf{cRF}}(t)\overline{\hat{\mathbf{X}}}_{\mathbf{cRF}}^{\mathrm{H}}(t)}{\sum_{t=1}^{T}\mathrm{cRM}_{\mathrm{X}}^{\mathrm{H}}(t)\mathrm{cRM}_{\mathrm{X}}(t)}.\end{aligned} \tag{20}$$

The MCMF estimated noise $\hat{\overline{\mathbf{N}}}_{\mathbf{cRF}}$ component and its time-varying covariance matrix $\tilde{\boldsymbol{\Phi}}_{\mathbf{NN}}$ can be estimated using the same method. Once the time-varying speech and noise covariance matrices are obtained, we can follow similar steps as the MC ADL-MVDR described in Eq. (13) to estimate the MCMF steering vector and inverse of the MCMF noise covariance matrix. After that, the MCMF ADL-MVDR beamforming weights $\hat{\mathbf{h}}_{\text{MCMF ADL-MVDR}}(t) \in \mathbb{C}^{M\times L}$ are derived in a similar manner as described in Eq. (15). Finally, the MCMF ADL-MVDR enhanced speech $\hat{\mathrm{X}}_{\text{MCMF ADL-MVDR}}^{(0)}$ is obtained

$$\hat{\mathrm{X}}_{\text{MCMF ADL-MVDR}}^{(0)}(t) = \hat{\mathbf{h}}_{\text{MCMF ADL-MVDR}}^{\mathrm{H}}(t)\overline{\mathbf{Y}}(t), \tag{21}$$

where $\overline{\mathbf{Y}}$ is the MCMF noisy speech.

IV. Experimental Setup

A. Speech Materials

We adopt a Mandarin audio-visual speech corpus (will be released soon [61]) collected from Youtube, which has been reported in our prior works [22], [23], [43]. An SNR estimator together with a face detector is used to filter out the low-quality ones [23], [43]. A total number of 205,500 clean video segments from around 1500 speakers are gathered. In contrast to our prior works [23], [43], we do not use the lip movement features for our system since we focus on beamforming in this study. There are 190,000 speech utterances in the training set, 15,000 utterances in the validation set, and another 500 utterances in the testing set. Speakers in the testing set are different from those in the training set. The sampling rate is set to 16 kHz, random clips from 255 noises recorded indoors are added to the clean utterances at SNRs from 18-30 dB [23]. The signal-to-interference ratio (SIR) is between -6 to 6 dB.

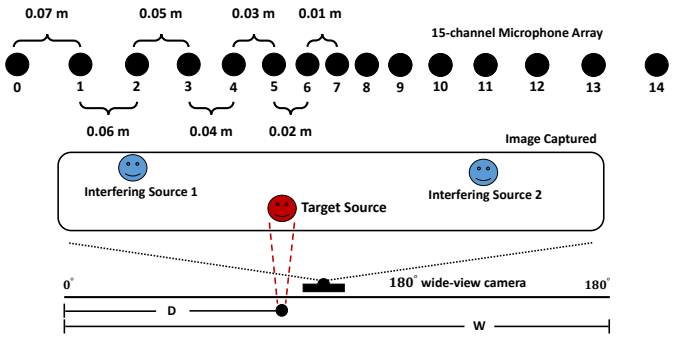

Fig. 4. (Color Online). Schematic diagrams for the 15-channel linear microphone array and the DOA estimation process with a wide view camera.

Reverberation effects are also applied following the image-source method [62], with T60 ranging from 0.05 s to 0.7 s. The 15-channel noisy reverberant mixtures are simulated according to the steps mentioned in [23], [43]. An illustration of the 15-channel microphone array that is calibrated with a 180° camera is depicted in Fig. 4, where the microphone array is a linear array and is symmetric to the center microphone (i.e., the 7th microphone in Fig. 4). The DOA is roughly estimated with the target source location D on the image with width W that is captured by the camera, where DOA $= \frac{\mathrm{D}}{\mathrm{W}} \times 180°$.

B. Audio Features and Training Procedure

In order to extract the audio features, we use a 512-point STFT together with a 32 ms Hann window and 16 ms step size. During the training stage, the batch size and audio chunk size are set to 12 and 4 s, respectively. We adopt Adam optimizer with the initial learning rate set to $1e^{-3}$, Pytorch 1.1.0 is used. All models are trained with 60 epochs and early stopping is applied. The system is trained to minimize the time-domain Si-SNR loss [63], which is the negative of Si-SNR, i.e.,

$$\begin{aligned}\mathcal{L}_{\text{Si-SNR}} &= -20\log_{10}\frac{\|\alpha\cdot\mathbf{x}\|}{\|\hat{\mathbf{x}}-\alpha\cdot\mathbf{x}\|},\\ \alpha &= \frac{\hat{\mathbf{x}}^{\mathrm{T}}\mathbf{x}}{\mathbf{x}^{\mathrm{T}}\mathbf{x}},\end{aligned} \tag{22}$$

where α is a scaling factor that ensures the scaling invariance, $\hat{\mathbf{x}}$ denotes the time-domain estimated speech. This time-domain loss fits our end-to-end training paradigm. Note that we use the reverberant clean speech as the learning target as we mainly focus on separation in this study, the systems are not trained for dereverberation in the present study.

C. System Setups

We adopt a Conv-TasNet variant [21] for complex filter estimation, which contains bunch of dilated 1-D convolution networks together with a pair of fixed STFT/iSTFT encoder/decoder [23], [43]. The details can be found in our previous works [43], [23]. In this study, we focus on the ADL-MVDR framework which is feasible for MC, MF and MCMF-MVDR filtering. Three setups corresponding to these three scenarios are described below.

For the MC ADL-MVDR system, both GRU-Nets consist of two layers of GRU and another fully connected layer. The $\mathbf{GRU\text{-}Net_{NN}}$ uses 500 units for both GRU layers and 450 units (i.e., # of channel $\times$ # of channel $\times$ real and imaginary parts = $15\times15\times2$) for the fully connected layer. The $\mathbf{GRU\text{-}Net}_{\boldsymbol{v}}$ contains 500 and 250 units for each GRU layer, respectively, followed by a fully connected layer with 30 units (i.e., 15×2). Tanh activation function is used for all GRU layers and linear activation function is used for the fully connected layers. The cRF size is empirically set to 3×3 (i.e., a T-F bin with its surrounding 8 T-F bins as depicted in Fig. 3) where it can utilize temporal information from 1 previous frame to 1 future frame and nearby frequency information from 1 frequency bin below to 1 frequency bin above.

In terms of the MF ADL-MVDR system, the GRU-Nets feature the same structure to the MC setup, but with different hidden sizes. We use a MF size of 5, i.e., from 2 previous frames to 2 future frames. The size of the cRF is 3×3, identical to the MC setup. Unit size of all GRU layers is set to 128. The fully connected layer contains 10 units for $\mathbf{GRU\text{-}Net}_{\boldsymbol{\gamma}}$ and 50 units for $\mathbf{GRU\text{-}Net_{VV}}$.

To investigate the influence of incorporating additional MF information on top of the MC spatial information, a 9-channel (i.e., the 0th, 2th, 3th, 5th, 7th, 9th, 11th, 12th and 14th channels) 3-frame (i.e., from 1 previous frame to 1 future frame) MCMF ADL-MVDR system is included. Here the $\mathbf{GRU\text{-}Net_{NN}}$ consists of two GRU layer with 500 units each, followed by another 1458-unit fully connected layer. The $\mathbf{GRU\text{-}Net}_{\boldsymbol{v}}$ contains two GRU layers with 500 and 250 units, respectively, with another fully connected layer of 54 units. The cRF size is also set to 3×3.

Meanwhile, we investigate several microphone and MF setups in the ablation study on MCMF ADL-MVDR systems, including when only 3 (i.e., the 0th, 7th and 14th channels), 7 (i.e., the 0th, 3th, 5th, 7th, 9th, 11th and 14th channels), 9 (identical to the one mentioned above) or all 15 channels are available to the ADL-MVDR module. The cRF sizes are all set to 3×3, the MF sizes are set to 3 (i.e., from 1 previous frame to 1 future frame) and 2 (i.e., 1 previous frame to current frame) for different MCMF ADL-MVDR systems as presented in Table II. We also include their corresponding MC ADL-MVDR systems (i.e., without additional MF information) with the same selected microphone channels for comparison approaches.

In other ablation studies, we further examine the influence of the cRF size on the performance of MC/MF ADL-MVDR systems. The effects of different MF sizes are also investigated for MF ADL-MVDR systems. These results are reported in the following section.

D. Evaluation Metrics

A set of objective evaluation metrics are used to evaluate the systems' performance from different perspectives. These metrics include PESQ [64] and source-to-distortion ratio (SDR) [65] for speech quality assessment. The Si-SNR score is also included as it has been utilized for many recent speech separation systems [21], [66], [22]. Moreover, we use a Tencent speech recognition API [67] to measure the ASR accuracy. The transcript of the speech is manually labelled by human annotators.

V. Results and Analysis

The general experimental results are provided in Table I, where we compare the performance of our proposed ADL-MVDR systems in MF, MC and MCMF conditions with its peers. Demos can be found on our website[1]. The performance of purely NN systems (i.e., a Conv-TasNet variant proposed in our prior work [23], [43], denoted as NN with cRM/cRF) are also included as baselines. We also include conventional mask-based MC-MVDR systems [22] and multi-tap MVDR systems [22] that incorporate this front-end NN for speech and noise components estimation with additional MVDR beamforming modules for comparison approaches in Table I. In terms of the results, the PESQ scores are further split up into detailed conditions, including the angle between the closest interfering source and total number of speakers. We present average results of other metrics (i.e., Si-SNR and SDR) for brevity.

In terms of the ablation studies, the simulation results for different MCMF ADL-MVDR systems are provided in Table II. Table III illustrates the results on the effects of different cRF sizes on both MF and MC ADL-MVDR systems, where the filtering region for each T-F pixel is described by its relative boundaries of frames and frequency bins. The effects of MF sizes on the performance of MF ADL-MVDR systems are also revealed in Table IV. We want to point out that there are infinitely many combinations of different cRF sizes and MF sizes, we only investigate a limited number of them which we consider to be representative.

A. Overview Results on ADL-MVDR Systems

MC ADL-MVDR vs. Conventional mask-based MVDR: we first investigate the performance of our proposed ADL-MVDR framework in the MC scenario. As provided in the third block of Table I, the MC ADL-MVDR system outperforms conventional MVDR systems by a large margin across all objective scores. For instance, in terms of the speech quality, our proposed MC ADL-MVDR system outperforms the MVDR system with cRF for more than 17% (i.e., PESQ: 3.42 vs. 2.92). Even under extreme conditions when the interfering sources are very close to the target speaker (i.e., angles less than 15°), the MC ADL-MVDR system can still restore the separated speech quality to a high level (i.e., PESQ: 3.04). In terms of the Si-SNR and SDR performance, our proposed MC ADL-MVDR system also achieves nearly 31% and 23% improvements over the baseline MVDR system with cRF (i.e., Si-SNR: 14.80 dB vs. 11.31 dB, SDR: 15.45 dB vs. 12.58 dB). Note that the residual noise in both the conventional MVDR and multi-tap MVDR systems is still at relatively high level according to their PESQ scores (i.e., average scores of 2.92 and 3.08, respectively) when compared to the purely

[1]Samples of separated speech (including real-world scenarios) are available at https://zzhang68.github.io/mcmf-adl-mvdr/

TABLE I

EVALUATION RESULTS FOR OUR PROPOSED ADL-MVDR SYSTEMS. THE PESQ SCORES ARE PRESENTED IN DETAILED CONDITIONS INCLUDING ANGLE BETWEEN THE CLOSEST INTERFERING SOURCE AND TOTAL NUMBER OF SPEAKERS. THE AVERAGE SCORES OF SI-SNR AND SDR ARE GIVEN FOR BREVITY. THE ASR ACCURACY IS MEASURED WITH THE WER AND THE BEST SCORES ARE HIGHLIGHTED IN **BOLD** FONTS.

Systems/Metrics	PESQ $\in [-0.5, 4.5]$								Si-SNR (dB)	SDR (dB)	WER (%)
	0-15°	15-45°	45-90°	90-180°	1spk	2spk	3spk	Avg.	Avg.	Avg.	
Reverberant clean (reference)	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	∞	∞	8.26
Noisy Mixture	1.88	1.88	1.98	2.03	3.55	2.02	1.77	2.16	3.39	3.50	55.14
Purely NNs and our proposed MF ADL-MVDR system											
NN with cRM	2.72	2.92	3.09	3.07	3.96	3.02	2.74	3.07	12.23	12.73	22.49
NN with cRF	2.75	2.95	3.12	3.09	3.98	3.06	2.76	3.10	12.50	13.01	22.07
MF ADL-MVDR with cRF	**2.80**	**2.99**	**3.16**	**3.11**	**4.01**	**3.10**	**2.80**	**3.14**	**12.60**	**13.17**	**19.57**
Conventional mask-based MVDR systems and our proposed MC ADL-MVDR systems											
MVDR with cRM [22]	2.55	2.76	2.96	2.84	3.73	2.88	2.56	2.90	10.62	12.04	16.85
MVDR with cRF	2.55	2.77	2.96	2.89	3.82	2.90	2.55	2.92	11.31	12.58	15.91
Multi-tap MVDR with cRM [22]	2.70	2.96	3.18	3.09	3.80	3.07	2.74	3.08	12.56	14.11	13.67
Multi-tap MVDR with cRF	2.67	2.95	3.15	3.10	3.92	3.06	2.72	3.08	12.66	14.04	13.52
MC ADL-MVDR with cRF	**3.04**	**3.30**	**3.48**	**3.48**	**4.17**	**3.41**	**3.07**	**3.42**	**14.80**	**15.45**	**12.73**
Our proposed MCMF ADL-MVDR system											
MCMF ADL-MVDR with cRF	**3.10**	**3.35**	**3.48**	**3.51**	**4.20**	**3.47**	**3.10**	**3.46**	**15.43**	**16.03**	**12.31**

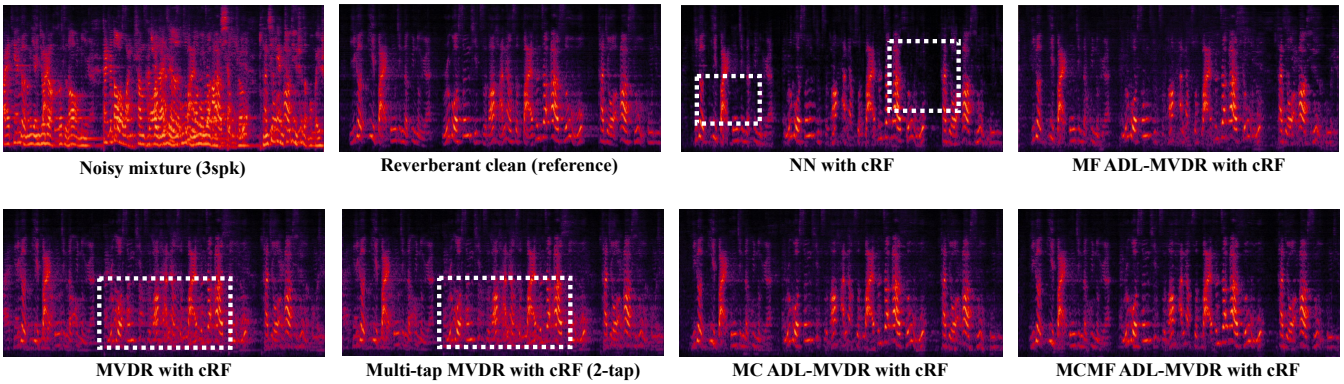

Fig. 5. (Color Online). Spectrograms of some evaluated systems in Table I, the nonlinear distortion and residual noise are highlighted by the dashed boxes.

NN systems (e.g., NN with cRF: 3.10). Our proposed MC ADL-MVDR system also demonstrates its superiority in ASR accuracy when compared to the multi-tap MVDR system with cRF (i.e., 12.73% vs. 13.52%). Considering that the current commercial ASR system is already very robust to low-level of noises, the nearly 6% relative improvement in WER is fairly good since our MC ADL-MVDR system can greatly remove the residual noise while introducing even fewer distortions to the target speech simultaneously. This can also be observed in the example spectrograms provided in Fig. 5, the conventional MVDR and multi-tap MVDR systems come with high levels of residual noise, whereas our MC ADL-MVDR system resolves this issue.

An example comparison of the beam patterns between conventional mask-based MVDR system and our proposed MC ADL-MVDR system is provided in Fig. 6. It represents the case of a 2-speaker mixture, with the target and interfering sources at directions of 63° and 131°, respectively. It is obvious that our proposed MC ADL-MVDR system can better capture the target source information with a sharper main lobe at the corresponding target direction. The frequency for these beam pattern plots are set to 968 Hz. We pick the representative time index for MC ADL-MVDR system in order to visualize its time-varying beamforming weights.

MF ADL-MVDR vs. NNs: The simulation results of the MF ADL-MVDR system are provided in the second block of Table I. By comparing the performance between MF ADL-MVDR system and the purely NN system with cRF, we observe that the MF ADL-MVDR system can lead to moderate improvements in all objective metrics (i.e., PESQ: 3.14 vs. 3.10, Si-SNR: 12.60 dB vs. 12.50 dB, and SDR: 13.17 dB vs. 13.01 dB) when only 1 channel is available to the ADL-MVDR module. We infer that the limited improvement here compared to MC ADL-MVDR system is due to the loss of spatial information. In the meantime, the proposed MF ADL-MVDR system achieves huge improvement on ASR accuracy, which is about 11.3% better in WER than NN with cRF (i.e., 19.57% vs. 22.07%). This implies that our proposed MF ADL-MVDR system can greatly reduce the nonlinear distortion introduced by conventional purely NN systems. Also reflected in Fig. 5, where the purely NN systems usually come with a large degree of nonlinear distortion such as the spectral 'black holes' (highlighted in the dashed boxes). Although not provided with spatial information, the proposed MF ADL-MVDR system drastically outperforms the purely NN systems in terms of ASR accuracy while achieving better objective scores.

MCMF ADL-MVDR vs. MC ADL-MVDR: The results of

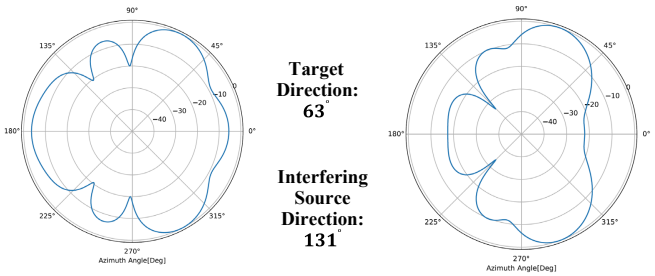

Fig. 6. (Color Online). Beam pattern examples for conventional mask-based MVDR system and our proposed MC ADL-MVDR system. Target source at 63°, interfering source at 131°.

the best performing MCMF ADL-MVDR system (9-channel 3-frame) are showed in the last block of Table I. Compared to its MC ADL-MVDR peer, we notice that the MCMF ADL-MVDR system can further improve the objective scores (i.e., PESQ: 3.46 vs. 3.42, Si-SNR: 15.43 dB vs. 14.80 dB and SDR: 16.03 dB vs. 15.45 dB), suggesting that MCMF ADL-MVDR system can even better remove the residual noise than MC ADL-MVDR systems. Slightly better performance is observed in terms of ASR accuracy (i.e., 12.31% vs. 12.73%) when compared to MC ADL-MVDR system. Results here suggest that on top of the spatial information, incorporating additional MF information is beneficial.

B. Ablation Study on MCMF ADL-MVDR Systems

As shown in Table II, we provide comparison results between several MCMF ADL-MVDR systems and MC ADL-MVDR systems (i.e., MF size = 1, current frame). Comparing the performance between MCMF and MC ADL-MVDR systems, we find that the inclusion of additional MF information can often lead to improved performance in both objective scores as well as the ASR accuracy. For example in 3-channel scenario, the PESQ, Si-SNR and SDR scores are 3.40 vs. 3.29, 14.97 dB vs. 13.85 dB, 15.50 dB vs. 14.43 dB between MCMF and MC ADL-MVDR systems. The WER for MCMF ADL-MVDR system is roughly 12% better than MC ADL-MVDR system (i.e., 12.74% vs. 14.46%) in 3-channel scenario.

Additionally, we observe that the gains from the additional MF information for MCMF ADL-MVDR systems become smaller as the number of available microphone channels increases. For instance, in 3-channel scenario, the MCMF ADL-MVDR system achieves relative improvement of 3.3% on PESQ scores (i.e. 3.40 vs. 3.29) and 12% on ASR accuracy. Whereas in 7-channel scenario, the relative improvement for MCMF ADL-MVDR system is only around 2.7% on PESQ scores (i.e., 3.43 vs. 3.34) and about 6% in terms of WER (i.e., 12.84% vs. 13.67%). This pattern suggests that the MF information becomes less important when more spatial information is available. When all 15 channels are provided, similar trend holds such that including additional MF information can further improve the objective scores (e.g., Si-SNR: 15.16 dB vs. 14.80 dB), while achieving similarly high ASR accuracy (i.e., 12.89% vs. 12.73%).

TABLE II
EVALUATION RESULTS FOR THE MCMF ADL-MVDR SYSTEMS AND MC ADL-MVDR SYSTEMS.

# of Channels	MF Size	PESQ	Si-SNR (dB)	SDR (dB)	WER (%)
3	3	**3.40**	**14.97**	**15.50**	**12.74**
3	1	3.29	13.85	14.43	14.46
7	3	**3.43**	**15.31**	**15.91**	**12.84**
7	1	3.34	14.14	14.80	13.67
9	3	**3.46**	**15.43**	**16.03**	**12.31**
9	1	3.39	14.67	15.15	13.17
15	2	**3.42**	**15.16**	**15.73**	12.89
15	1	**3.42**	14.80	15.45	**12.73**

We also want to point out that the feature space (i.e., size of the covariance matrix) is increasing exponentially with the number of microphone channels and the MF size. Therefore, when the number of available microphone channels is small, additional MF information (i.e., MCMF ADL-MVDR) may help to enhance the performance. But this could be redundant when more spatial information is available and a large feature size may hinder the learning process of NNs.

C. Ablation Study on cRF Sizes

The results for different sizes of the cRF are presented in Table III. The MF size is fixed at 5 (i.e., from 2 previous frames to 2 future frames) for all MF ADL-MVDR systems. We find that a 1×1 cRF (i.e., a cRM) results in the worst performance (e.g., WER: 22.49%) for purely NN systems when compared its peers with larger cRF sizes. The NN with a 5×5 cRF (i.e., the 3rd purely NN system) leads to the best performance in WER (i.e., 21.80%) and a 3×3 cRF can achieve the best objective scores (i.e., PESQ: 3.10, Si-SNR: 12.50 dB, SDR: 13.01 dB). In general, we find that NNs with cRF of sizes 3×3, 5×5 and 7×7 yield with similar performance in objective metrics and ASR accuracy, while the cRM alone is not sufficient to achieve the optimal performance. Larger cRF size could lead to slightly better performance for purely NNs but there is also a trade-off on the performance and run time efficiency of the system.

Similar patterns can be found in MC ADL-MVDR systems, where the cRM alone (i.e., the 1st MC ADL-MVDR system) is not leading to satisfactory performance (e.g., WER: 23.73%) and size of the cRF could be even more important. By comparing the systems with cRF of sizes 1×1 and 1×3 (i.e., the 1st and 2nd MC ADL-MVDR systems), we find that including nearby frequency information would help improve the system's performance (e.g., WER: 23.73% vs. 17.87%). Whereas substantial improvements (e.g., WER: 23.73% vs. 13.52%) can be achieved by introducing the nearby temporal information to the cRF (i.e., the 1st and 4th MC ADL-MVDR systems), which suggests that temporal information could be more important than nearby frequency information for our proposed ADL-MVDR systems. Meanwhile, we also find that including future frame information in cRF could help improve the system's performance. For example, by comparing the MC ADL-MVDR systems with cRF that contains information from 2 previous frames and the other one with temporal information from 1 previous frame to 1 future frame (i.e., the 3rd and 5th MC ADL-MVDR systems), slight improvements can be

TABLE III
EFFECTS OF THE CRF SIZES ON THE PERFORMANCE OF PURELY NNS AND MC/MF ADL-MVDR SYSTEMS. THE CRF SIZE IS REPRESENTED BY ITS TIME (T.) AND FREQUENCY (F.) RANGES, WHERE 0, NEGATIVE AND POSITIVE NUMBERS INDICATE THE CURRENT, PREVIOUS AND FUTURE TIME FRAME OR FREQUENCY BIN, RESPECTIVELY.

T. Range	F. Range	PESQ	Si-SNR (dB)	SDR (dB)	WER (%)
Purely NN Systems					
0	0	3.07	12.23	12.73	22.49
[-1,1]	[-1,1]	**3.10**	**12.50**	**13.01**	22.07
[-2,2]	[-2,2]	3.09	12.45	12.97	**21.80**
[-3,3]	[-3,3]	3.09	12.44	12.96	22.15
MC ADL-MVDR Systems					
0	0	2.83	11.51	11.98	23.73
0	[-1,1]	3.21	13.08	13.77	17.87
[-2,0]	[-1,1]	3.40	14.51	15.09	12.88
[-1,1]	0	3.37	14.31	14.93	13.52
[-1,1]	[-1,1]	**3.42**	**14.80**	**15.45**	**12.73**
MF ADL-MVDR Systems					
0	0	2.95	11.75	12.27	21.97
0	[-1,1]	3.12	12.26	12.81	21.08
[-1,0]	0	3.13	12.44	12.96	20.63
[-2,0]	0	3.13	12.44	12.95	21.11
[-2,0]	[-1,1]	3.11	12.35	12.90	21.41
[-1,1]	0	**3.16**	12.55	13.04	20.40
[-1,1]	[-1,1]	3.14	**12.60**	**13.17**	**19.57**
[-2,2]	[-2,2]	3.15	12.58	13.12	20.42

observed in both ASR accuracy (i.e., 12.73% vs. 12.88%) and objective scores (i.e., PESQ: 3.42 vs. 3.40, Si-SNR: 14.80 dB vs. 14.51 dB and SDR: 15.45 dB vs. 15.09 dB).

Several cRF setups for MF ADL-MVDR systems are also included and their results are generally consistent with those for MC ADL-MVDR and purely NN systems. Specifically, a 1×1 cRF (i.e., a cRM) does not perform well on MF ADL-MVDR system either (i.e., lowest ASR accuracy and objective scores). The inclusion of future frame information in cRF is crucial (e.g., comparing the 4th and 6th MF ADL-MVDR systems, WER: 21.11% vs. 20.40%) and that the nearby frequency information could also improve the ASR performance slightly while achieving similar objective scores (i.e., comparing the 6th and 7th MF ADL-MVDR systems, PESQ: 3.16 vs. 3.14, Si-SNR: 12.55 dB vs. 12.60 dB, SDR: 13.04 dB vs. 13.17 dB, and WER: 20.40% vs. 19.57%). Increasing the cRF size from 3×3 to 5×5 (i.e., the last two MF ADL-MVDR systems) does not improve the system's performance and even result in slightly poorer performance (except for PESQ scores), which indicates that adopting a very large size cRF may not be necessary or beneficial.

D. Ablation Study on MF Sizes

The cRF sizes for all MF ADL-MVDR systems presented in Table IV are fixed at 3×3 (i.e., ± 1 nearby frames and frequency bins) in order to investigate the influence of different MF sizes. As shown in Table IV, we include 6 different setups where the first 2 system setups represent the conditions when only information from previous frames is available, and the last 4 MF conditions further include information from future frames. By comparing the first 2 MF ADL-MVDR system in Table IV, we find that the inclusion of additional previous frame (i.e., t-2) could help improve the objective scores (e.g., Si-SNR: 12.39 dB vs. 12.06 dB) while achieving similar WER performance (i.e., 20.87% vs. 20.50%). Then, by comparing

TABLE IV
EFFECTS OF MF SIZES ON THE OBJECTIVE PERFORMANCE AND ASR ACCURACY OF MF ADL-MVDR SYSTEMS. t REPRESENTS THE CURRENT FRAME.

MF ADL-MVDR Systems					
MF Size	MF Range	PESQ	Si-SNR (dB)	SDR (dB)	WER (%)
2	[t-1,t]	3.02	12.06	12.58	20.50
3	[t-2,t]	3.12	12.39	12.95	20.87
3	[t-1,t+1]	3.09	12.13	12.71	20.27
5	[t-2,t+2]	3.14	12.60	13.17	19.57
7	[t-3,t+3]	**3.18**	**12.89**	**13.42**	**19.45**
9	[t-4,t+4]	3.17	12.76	13.27	19.52

the 1st and 3rd MF ADL-MVDR systems (i.e., [t-1,t] and [t-1,t+1]), we observe improvements in both the objective metrics (e.g., PESQ: 3.02 vs. 3.09) and ASR accuracy (i.e., 20.50% vs. 20.27%), indicating that it is beneficial to include future frames in the MF information. By increasing the MF range (e.g., the 3rd and 4th MF ADL-MVDR systems in Table IV), further improvements can be obtained (e.g., PESQ: 3.09 vs. 3.14 and WER: 20.27% vs. 19.57%). We also find that further expanding the MF size leads to even better performance (i.e., the 5th and 4th MF ADL-MVDR systems), where the system achieves better objective scores (e.g., PESQ: 3.18 vs. 3.14) as well as improved ASR accuracy (i.e., WER: 19.45% vs. 19.57%). However, when the MF size continually increases from 7 to 9 (i.e., the last 2 setups in Table IV), no further improvements are observed (e.g., PESQ: 3.17 vs. 3.18 and WER: 19.52% vs. 19.45%).

VI. CONCLUSIONS AND FUTURE WORK

In this work, we proposed an ADL-MVDR framework that can be applied for multi-channel, multi-frame, and multi-channel multi-frame target speech separation tasks. Our proposed ADL-MVDR framework has achieved the best performance across a set of objective evaluation metrics as well as ASR accuracy in both the multi-channel and multi-frame scenarios among its peers. The multi-channel multi-frame ADL-MVDR system can achieve even better performance by fully exploring spatio-temporal cross correlations. Leveraging on RNN-predicted filtering weights, the proposed ADL-MVDR system also resolves the numerical instability issue that often occurs in conventional mask-based MVDR systems. Additionally, the proposed ADL-MVDR systems can keep the residual noise at a minimum level (reflected by the high objective scores) while introducing hardly any nonlinear distortions (reflected by the high ASR accuracy).

Future work of this study may include the following directions. Firstly, we will evaluate this ADL-MVDR framework on purely single-channel speech separation tasks. Secondly, the complex ratio filter and multi-frame sizes are determined empirically in this study, we will further explore approaches to control these sizes adaptively. Thirdly, we will adapt this ADL-MVDR framework for joint separation and dereverberation tasks. And lastly, we could apply the idea of this ADL-MVDR beamformer to other types of beamformer to explore a more general solution.

REFERENCES

[1] J. Du, Q. Wang, T. Gao, Y. Xu, L.-R. Dai, and C.-H. Lee, "Robust speech recognition with speech enhanced deep neural networks," in *Interspeech*, 2014.

[2] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *LVA/ICA*. Springer, 2015, pp. 91–99.

[3] X. Zhang, Z.-Q. Wang, and D. Wang, "A speech enhancement algorithm by iterating single-and multi-microphone processing and its application to robust ASR," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 276–280.

[4] J. Yu, B. Wu, R. Gu, S.-X. Zhang, L. Chen, Y. Xu, M. Yu, D. Su, D. Yu, X. Liu, and H. Meng, "Audio-visual multi-channel recognition of overlapped speech," *Interspeech*, 2020.

[5] S. E. Eskimez, P. Soufleris, Z. Duan, and W. Heinzelman, "Front-end speech enhancement for commercial speaker verification systems," *Speech Communication*, vol. 99, pp. 101–113, 2018.

[6] T. Van den Bogaert, S. Doclo, J. Wouters, and M. Moonen, "Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids," *JASA*, vol. 125, no. 1, pp. 360–371, 2009.

[7] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE TASLP*, vol. 22, no. 12, pp. 1849–1858, 2014.

[8] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved MVDR beamforming using single-channel mask prediction networks." in *Interspeech*, 2016, pp. 1981–1985.

[9] Z. Zhang, C. Deng, Y. Shen, D. S. Williamson, Y. Sha, Y. Zhang, H. Song, and X. Li, "On loss functions and recurrency training for GAN-based speech enhancement systems," *arXiv preprint arXiv:2007.14974*, 2020.

[10] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *speech communication*, vol. 53, no. 4, pp. 465–494, 2011.

[11] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for joint enhancement of magnitude and phase," in *ICASSP*. IEEE, 2016, pp. 5220–5224.

[12] Y. Xu, M. Chen, P. LaFaire, X. Tan, and C.-P. Richter, "Distorting temporal fine structure by phase shifting and its effects on speech intelligibility and neural phase locking," *Scientific reports*, vol. 7, no. 1, pp. 1–9, 2017.

[13] Z. Zhang, D. S. Williamson, and Y. Shen, "Investigation of phase distortion on perceived speech quality for hearing-impaired listeners," *arXiv preprint arXiv:2007.14986*, 2020.

[14] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *ICASSP*. IEEE, 2015, pp. 708–712.

[15] J. Lee, J. Skoglund, T. Shabestary, and H.-G. Kang, "Phase-sensitive joint learning algorithms for deep learning-based speech enhancement," *IEEE Signal Processing Letters*, vol. 25, no. 8, pp. 1276–1280, 2018.

[16] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[17] Z.-Q. Wang, J. Le Roux, D. Wang, and J. R. Hershey, "End-to-end speech separation with unfolded iterative phase reconstruction," *arXiv preprint arXiv:1804.10204*, 2018.

[18] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *ICASSP*, 2018, pp. 696–700.

[19] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1570–1584, 2018.

[20] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[21] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE TASLP*, vol. 27, no. 8, pp. 1256–1266, 2019.

[22] Y. Xu, M. Yu, S.-X. Zhang, L. Chen, C. Weng, J. Liu, and D. Yu, "Neural spatio-temporal beamformer for target speech separation," *arXiv preprint arXiv:2005.03889*, 2020.

[23] 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 J-STSP*, 2020.

[24] Y. Huang, J. Benesty, and J. Chen, "Analysis and comparison of multichannel noise reduction methods in a common framework," *IEEE transactions on audio, speech, and language processing*, vol. 16, no. 5, pp. 957–968, 2008.

[25] M. Souden, J. Benesty, and S. Affes, "New insights into non-causal multichannel linear filtering for noise reduction," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 141–144.

[26] B. D. Van V. and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE assp magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[27] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[28] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer Science & Business Media, 2008, vol. 1.

[29] M. Souden, J. Benesty, and S. Affes, "A study of the LCMV and MVDR noise reduction filters," *IEEE Transactions on Signal Processing*, vol. 58, no. 9, pp. 4925–4935, 2010.

[30] C. Pan, J. Chen, and J. Benesty, "Performance study of the MVDR beamformer as a function of the source incidence angle," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 67–79, 2013.

[31] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *ICASSP*, 2016, pp. 5210–5214.

[32] 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, pp. 3246–3250.

[33] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on audio, speech, and language processing*, vol. 15, no. 5, pp. 1529–1539, 2007.

[34] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 444–451.

[35] 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.

[36] 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.

[37] D. Fischer and S. Doclo, "Robust constrained MFMVDR filtering for single-microphone speech enhancement," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 41–45.

[38] ——, "Sensitivity analysis of the multi-frame MVDR filter for single-microphone speech enhancement," in *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, pp. 603–607.

[39] 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, pp. 6745–6749.

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

[41] M. Souden, J. Chen, J. Benesty, and S. Affes, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2159–2169, 2011.

[42] 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," *arXiv preprint arXiv:2008.06994*, 2020.

[43] R. Gu, S.-X. Zhang, Y. Xu, L. Chen, Y. Zou, and D. Yu, "Multi-modal multi-channel target speech separation," *IEEE J-STSP*, 2020.

[44] W. Mack and E. A. Habets, "Deep filtering: Signal extraction and reconstruction using complex time-frequency filters," *IEEE Signal Processing Letters*, vol. 27, pp. 61–65, 2019.

[45] E. Oja, "Simplified neuron model as a principal component analyzer," *Journal of mathematical biology*, vol. 15, no. 3, pp. 267–273, 1982.

[46] J. Wang, "A recurrent neural network for real-time matrix inversion," *Applied Mathematics and Computation*, vol. 55, no. 1, pp. 89–100, 1993.

[47] C. Fyfe, "A neural network for PCA and beyond," *Neural Processing Letters*, vol. 6, no. 1-2, pp. 33–41, 1997.

[48] Y. Zhang and S. S. Ge, "Design and analysis of a general recurrent neural network model for time-varying matrix inversion," *IEEE Trans. on Neural Networks*, vol. 16, no. 6, pp. 1477–1490, 2005.

[49] E. Oja, H. Ogawa, and J. Wangviwattana, "Principal component analysis by homogeneous neural networks, part¡ cd02d35. gif¿: The weighted subspace criterion," *IEICE Transactions on Information and Systems*, vol. 75, no. 3, pp. 366–375, 1992.

[50] K. Shimada, Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Unsupervised beamforming based on multichannel nonnegative matrix factorization for noisy speech recognition," in *ICASSP*, 2018, pp. 5734–5738.

[51] Z.-Q. Wang and D. Wang, "All-neural multi-channel speech enhancement," in *Interspeech*, 2018, pp. 3234–3238.

[52] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *ICASSP*. IEEE, 2016, pp. 196–200.

[53] Z.-Q. Wang and D. Wang, "On spatial features for supervised speech separation and its application to beamforming and robust ASR," in *ICASSP*. IEEE, 2018, pp. 5709–5713.

[54] J. Benesty and Y. Huang, "A single-channel noise reduction MVDR filter," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 273–276.

[55] O. Schwartz, S. Gannot, and E. A. P. Habets, "Multi-microphone speech dereverberation and noise reduction using relative early transfer functions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 240–251, 2014.

[56] X. Xiao, C. Xu, and et al., "A study of learning based beamforming methods for speech recognition," in *CHiME 2016 workshop*, 2016.

[57] B. Li, T. N. Sainath, R. J. Weiss, K. W. Wilson, and M. Bacchiani, "Neural network adaptive beamforming for robust multichannel speech recognition," 2016.

[58] 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 *IEEE SLT*, 2018, pp. 558–565.

[59] G. L. and J. C., "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Antennas and Propagation*, vol. 30, pp. 27–34, 1982.

[60] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Transactions on signal processing*, vol. 50, no. 9, pp. 2230–2244, 2002.

[61] S.-X. Zhang, Y. Xu, M. Yu, L. Chen, and D. Yu, "Multi-modal multi-channel system and corpus for cocktail party problems," in *Preparation*, 2020.

[62] E. A. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, p. 1, 2006.

[63] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *ICASSP*. IEEE, 2019, pp. 626–630.

[64] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*, vol. 2, 2001, pp. 749–752.

[65] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE TASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

[66] 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," *arXiv preprint arXiv:1905.07497*, 2019.

[67] "Tencent ASR," https://ai.qq.com/product/aaiasr.shtml.