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 to improve objective assessment scores, but they often introduce nonlinear distortions that are harmful to modern automatic speech recognition (ASR) systems. Minimum variance distortionless response (MVDR) filters are often adopted to remove nonlinear distortions, however, conventional neural mask-based MVDR systems still result in relatively high levels of residual noise. Moreover, the matrix inverse involved in the MVDR solution is sometimes numerically unstable during joint training 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 proposed MCMF ADL-MVDR system addresses linear and nonlinear distortions. Spatio-temporal cross correlations are also fully utilized in the proposed approach. The proposed systems are evaluated using a Mandarin audio-visual corpus and are compared with several state-of-the-art approaches. Experimental results demonstrate the superiority of our proposed systems under different scenarios and across several objective evaluation metrics, including ASR performance.

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

I. INTRODUCTION

TARGET speech separation algorithms extract target speech signal from noisy background when interfering sources and background noise exist [1]. These algorithms serve as important front-ends for many speech communication systems such as automatic speech recognition (ASR) [2], [3], [4], [5], speaker verification [6], and digital hearing-aid devices [7]. 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 [8], [9], [10]. 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 [11], [12], [13], [14]. Phase-aware T-F masks have later been

proposed, including the phase-sensitive mask [15], [16], [10], and complex ratio mask [12]. Yet an accurate estimate of the phase component is still difficult for an 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 [17], [18], [19], [20], [21], [22], to avoid directly estimating the magnitude and phase components. Some of these approaches (e.g., Wave-U-Net [21] and TasNet [19]) 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 timedomain fully-convolutional Conv-TasNet [22] 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 [19]. Alternatively, other approaches implicitly combine the feature extraction and the separation steps as reported in [17], [20].

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 [23], [22], [24]. To alleviate the nonlinear distortion issue and achieve better performance on source separation, speech separation systems in practice often adopt a multi-channel processing scheme to further leverage spatial information in addition to original spectral information. Many approaches have been developed, including the multi-channel Wiener filter (MWF) [25], [26], the linearly constrained minimum variance (LCMV) filter [27] 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 [28], [29], [30], [31]. MVDR filters have been widely used in speech separation systems to reduce the amount of nonlinear distortions, which is helpful to ASR systems [32], [33], [23]. 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 [34], [35] aim 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 [36], [37], [38] have been adopted in single-channel speech separation systems to remove the noise 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 [36], [39].

Recent MVDR approaches are often combined with an NN-based T-F mask estimator [33], [40], [41], [23] 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 neural mask-based MVDR systems result in high levels of residual noise (e.g., linear distortions), since segment- or utterance-level beamforming weights cannot fully eliminate residual noise on each frame [42], [9], [33], [23]. Our recent work further incorporates multi-frame (MF) information during beamforming weights derivation [23], where it exploits extra inter-frame correlation in addition to the spatial correlation between the microphones in conventional multi-channel MVDR approaches. Results showed better ASR accuracy and higher PESQ scores when compared to conventional neural mask-based MVDR approaches. Unfortunately, the amount of residual noise in the separated signal is still high.

Many online beamforming weights estimation approaches have been proposed recently for real-time or time-varying purposes [43], [44], [45], [46], [47], [48]. In [43], Souden et al. 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. Systems such as [44], [47], [48] also used smoothing factors to estimate the time-varying covariance matrices, although these approaches allow better performance in time-varying conditions, the residual noise could still be present and postfiltering is sometimes needed. [46] proposed a frame-level beamforming method, however, it achieved slightly worse ASR accuracy compared to conventional segment-level beamforming systems [46].

In the current study, we propose a novel all deep learning MVDR (ADL-MVDR) framework that can be adapted for speech separation under different microphone configurations, including multi-channel (single-frame), multi-frame (i.e., when only one channel is available to the beamforming module), and multi-channel multi-frame (MCMF) scenarios. This study extends our preliminary work on the ADL-MVDR beamformer [49], 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 [24], [50]) that consists of dilated 1-D 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) [51] 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 inverse [52], [53], [54], [55] and principal component analysis (PCA) [56], [54]. The proposed ADL-MVDR module deploys two recurrent neural networks (RNNs) to replace the matrix operations (i.e., matrix inverse and principal eigenvector extraction) performed 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 non-stationary noise at each frame. The system also uses visual information (described in our prior work [24], [50]) to extract the direction of arrival (DOA) of the target speaker. Results from our prior study [49] 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 neural 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 the beamforming module, to further evaluate generalization. Secondly, we further adapt the ADL-MVDR framework to an 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 neural 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 NEURAL MASK-BASED MVDR FILTER

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

A. Conventional neural 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 M-dimensional 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 indices. 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 $\mathbf{x}(n) = \mathbf{g}(n) * \mathbf{s}(n)$, where $\mathbf{x}(n)$ and $\mathbf{s}(n)$ are the time-domain MC reverberant and anechoic speech

signals, n is the time index, $\mathbf{g}(n)$ represents the room impulse response and '$*$' denotes linear convolution.

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

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

where $\mathbf{h}_{\text{MC-MVDR}}(f) \in \mathbb{C}^M$ 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

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

where $\boldsymbol{v}(f) \in \mathbb{C}^M$ stands for the target speech steering vector, which can be estimated as the principal eigenvector of speech covariance matrix (i.e., $\boldsymbol{v}(f) = \mathcal{P}\{\mathbf{\Phi}_{\mathbf{XX}}\}$) [32], [57], [58]. $\mathbf{\Phi}_{\mathbf{NN}}$ represents the noise covariance matrix. The MVDR solution based on the steering vector is [59], [60]

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

In practice, many studies adopt the other MVDR solution that is based on the reference channel [42], [23], [61]

$$\mathbf{h}_{\text{MC-MVDR}}(f) = \frac{\mathbf{\Phi}_{\mathbf{NN}}^{-1}(f)\mathbf{\Phi}_{\mathbf{XX}}(f)}{\text{Trace}(\mathbf{\Phi}_{\mathbf{NN}}^{-1}(f)\mathbf{\Phi}_{\mathbf{XX}}(f))}\mathbf{u}, \qquad (5)$$

where $\mathbf{u}$ is the one-hot vector that selects the reference microphone channel. Note that the matrix inverse involved in the MVDR solution is sometimes not numerically stable during joint training with NNs. For instance, the estimated noise covariance matrix could be singular, which will cause numerical instability when computing the matrix inverse [62], [63], [64], where diagonal loading is often used to alleviate this issue [65], [40], [23], [64].

In a typical neural mask-based MVDR system, the covariance matrices are estimated chunk-wisely with T-F masks [57], [9], [41], [66], [23]. 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

$$\hat{\mathbf{\Phi}}_{\mathbf{XX}}(f) = \frac{\sum_{t=1}^{T} \text{RM}_{\mathrm{X}}^2(t,f)\mathbf{Y}(t,f)\mathbf{Y}^{\mathrm{H}}(t,f)}{\sum_{t=1}^{T} \text{RM}_{\mathrm{X}}^2(t,f)}, \qquad (6)$$

where RM_{X} stands for the RM for estimating the speech component and T is the total number of frames. The power mask is used for normalization. The noise covariance matrix $\hat{\mathbf{\Phi}}_{\mathbf{NN}}$ can be computed in a similar manner. Nevertheless, we want to point out that the chunk-wise beamforming cannot fully eliminate residual noise on each frame and therefore the relatively high levels of residual noise becomes a hand-in-hand problem for conventional neural 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 interframe correlations instead of performing spatial beamforming. Many MF-MVDR systems have been proposed recently [36], [69], [41]. Analogous to the MC-MVDR speech separation scenario, the process of obtaining the MF-MVDR enhanced speech can be described as

$$\hat{\mathrm{X}}^{(0)}(t,f) = \mathbf{h}_{\text{MF-MVDR}}^{\mathrm{H}}(t,f)\overline{\mathbf{Y}}^{(\mathbf{0})}(t,f), \qquad (7)$$

where $\mathbf{h}_{\text{MF-MVDR}}(t,f) \in \mathbb{C}^L$ 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 [36], [69], [23], respectively,

$$\begin{aligned} \mathbf{h}_{\text{MF-MVDR}}(t,f) &= [\mathrm{h}_0(t,f), \mathrm{h}_1(t,f), ..., \mathrm{h}_{L-1}(t,f)]^T, \\ \overline{\mathbf{Y}}^{(\mathbf{0})}(t,f) &= [\mathrm{Y}^{(0)}(t,f), \mathrm{Y}^{(0)}(t-1,f), ..., \qquad (8) \\ &\quad \mathrm{Y}^{(0)}(t-L+1,f)]^T, \end{aligned}$$

where $h_l(t,f)$ represents the l-th filter coefficient and $\mathrm{Y}^{(0)}(t,f)$ is the single-channel noisy speech STFT. This is similar to MC beamforming methods by viewing different frames as microphone inputs of different channels. Note that Eq. (8) can be extended to use information from future frames, which benefits speech processing. The MF speech and noise can be constructed by concatenating the delayed (or shifted for future frames) version of estimated speech or noise components following similar steps in Eq. (8). 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

$$\mathbf{h}_{\text{MF-MVDR}} = \arg\min_{\overline{\mathbf{h}}} \overline{\mathbf{h}}^{\mathrm{H}}\mathbf{\Phi}_{\mathbf{VV}}^{\text{MF}}\overline{\mathbf{h}} \quad \text{s.t.} \quad \overline{\mathbf{h}}^{\mathrm{H}}\boldsymbol{\gamma}_{\mathbf{x}} = \mathbf{1}, \quad (9)$$

where $\mathbf{\Phi}_{\mathbf{VV}}^{\text{MF}}(t,f) \in \mathbb{C}^{L \times L}$ denotes the covariance matrix of the MF undesired signal component [36], [37], [41] which consists of the noise and the uncorrelated speech components. $\boldsymbol{\gamma}_{\mathbf{x}}(t,f) \in \mathbb{C}^L$ is the speech interframe correlation (IFC) vector that describes the correlation between the previous and current frames. According to [69], [41], the speech IFC vector $\boldsymbol{\gamma}_{\mathbf{x}}$ can be formulated as

$$\boldsymbol{\gamma}_{\mathbf{x}}(t,f) = \frac{\mathbf{\Phi}_{\mathbf{XX}}^{\text{MF}}(t,f)\mathbf{e}}{E[|\mathrm{X}^{(0)}(t,f)|^2]}, \qquad (10)$$

where $\mathbf{\Phi}_{\mathbf{XX}}^{\text{MF}}$ stands for the covariance matrix of the MF speech and $\mathrm{X}^{(0)}$ 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. (9), the MF-MVDR filter vector can be obtained as [36], [69], [41]

$$\mathbf{h}_{\text{MF-MVDR}}(t,f) = \frac{\mathbf{\Phi}_{\mathbf{VV}}^{\text{MF}^{-1}}(t,f)\boldsymbol{\gamma}_{\mathbf{x}}(t,f)}{\boldsymbol{\gamma}_{\mathbf{x}}^{\mathrm{H}}(t,f)\mathbf{\Phi}_{\mathbf{VV}}^{\text{MF}^{-1}}(t,f)\boldsymbol{\gamma}_{\mathbf{x}}(t,f)}. \qquad (11)$$

Note that in [41], $\mathbf{\Phi}_{\mathbf{VV}}^{\text{MF}}$ was replaced by the MF noise covariance matrix $\mathbf{\Phi}_{\mathbf{NN}}^{\text{MF}}$ under the assumption that the uncorrelated speech component is negligible, which imposes an upper bound on the system's performance.
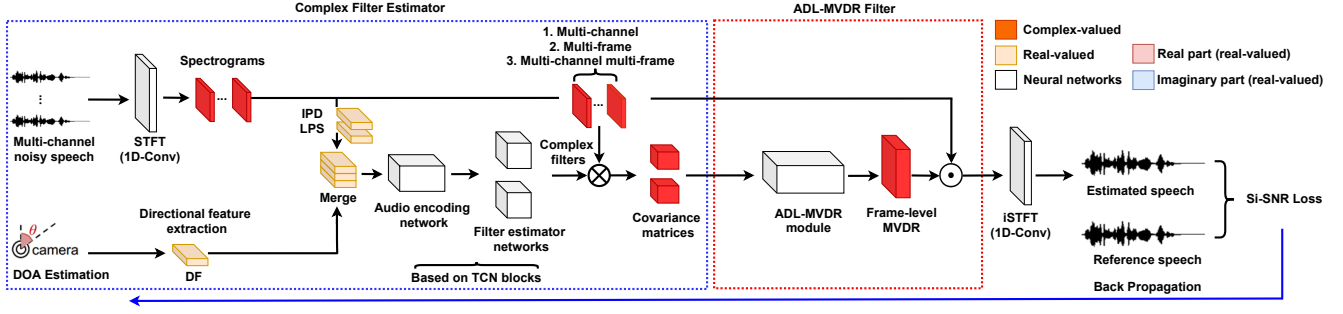
Fig. 1. (Color Online). Network architecture of our proposed ADL-MVDR framework. It consists of a complex filter estimator (i.e., highlighted in blue dashed box) based on temporal convolution network (TCN) blocks for components estimation (depending on the case), and an ADL-MVDR filter (i.e., highlighted 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. (15) and (19), (22) or (24), respectively, depending on the situation.
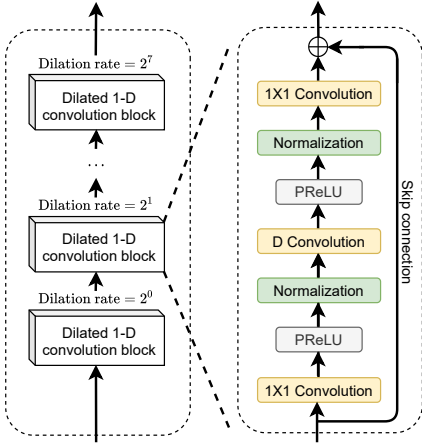

Fig. 2. (Color Online). Illustration of one TCN block (left) that consists of 8 dilated 1-D convolution blocks (right). Each dilated 1-D convolution block consists of 1×1 (i.e., pointwise) convolution layers and a depth-wise separable convolution layer (D convolution) [67]. PReLU [68] activation and normalization are applied between the convolution layers. Skip connection is also used.
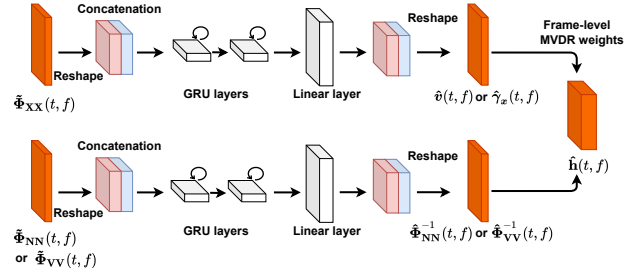

Fig. 3. (Color Online). Detailed network architecture of the ADL-MVDR module. The real and imaginary parts of the covariance matrices are concatenated before fed into the GRU-Nets. The estimated MVDR coefficients are reshaped back to their original forms before computing the frame-level MVDR weights.

III. PROPOSED ADL-MVDR BEAMFORMER

As mentioned in previous sections, the covariance matrices for most of the conventional neural mask-based MC-MVDR systems are computed at the chunk-level that discards the temporal information. This results in relatively high levels of residual noise in the separated speech. Additionally, the matrix inverse involved in conventional neural mask-based MVDR systems is sometimes not numerically stable when jointly trained with NNs [64]. 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 [43], [70], [41]. 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 idea of ADL-MVDR that can be applied to many different configurations, including MC beamforming, MF filtering and MCMF beamforming (note that a separate system needs to be trained for each microphone setup). The key idea of our proposed ADL-MVDR framework is about using two separate gated recurrent unit (GRU) based networks (denoted as GRU-Nets) to replace the matrix

inverse and principal eigenvector extraction processes involved in MVDR solution. 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 bypasses the numerical instability issue caused by the matrix inverse, whereas conventional methods often use diagonal loading and gradient clipping to address this issue [23], [64]. Note that previous approaches that adopt NNs to directly learn the beamforming filtering weights [71], [72] 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.

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 [24], [50] (i.e., a Conv-TasNet [22] variant) for covariance matrices estimation, followed by another ADL-MVDR module (depicted in Fig. 3) for frame-level MVDR weights derivation.

As described in our previous works [50], [24], inside the complex filter estimator, the interaural phase difference (IPD) and log-power spectra (LPS) features are extracted from the
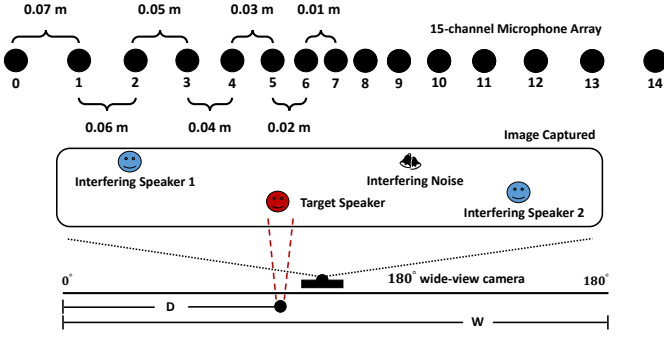
Fig. 4. (Color Online). Schematic diagrams for the 15-channel linear microphone array and the DOA estimation process with a wide view camera.

15-channel noisy speech, where the IPD between the p-th pair of microphone channels (e.g., m_1 and m_2) is

$$\text{IPD}^{(p)}(t, f) = \angle Y^{(m_1)}(t, f) - \angle Y^{(m_2)}(t, f), \quad (12)$$

where $\angle$ extracts the phase angle. We use the IPD computed from five pairs of channels, i.e., (0,14), (1,13), (2,11), (4,11), and (6,8), corresponding to five different distances between microphones [24], [50]. A directional feature (DF) [73] is also used in our experiment, which is defined as the cosine distance between the steering vector and IPD:

$$\text{DF}(t, f) = \sum_{p=1}^{P} \langle \mathbf{e}^{\text{TPD}^{(p)}(\theta_t, f)}, \mathbf{e}^{\text{IPD}^{(p)}(t, f)} \rangle, \quad (13)$$
$$\text{TPD}^{(p)}(\theta_t, f) = 2\pi f \Delta_p \cos\theta_t / (f_s c),$$

where $\langle \rangle$ denotes cosine distance and vector $\mathbf{e}^{(\cdot)} = \begin{bmatrix} \cos(\cdot) \\ \sin(\cdot) \end{bmatrix}$. $\text{TPD}^{(p)}(\theta_t, f)$ is the target-dependent phase difference that describes the phase delay of a plane wave (with frequency f), at the p-th pair of microphones (total number of P pairs), with target direction of θ_t, Δ_p is the distance between the p-th pair of microphones, f_s is the sampling frequency and c is the sound velocity. The target DOA (i.e., θ_t) can be estimated using a wide-view camera by locating the target speaker's face. Once the DF is obtained, it is further merged with the IPD and LPS features by concatenating along the feature dimension to form the input vector (1799 feature size) to the separation network [24], [50]. 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 $\text{DOA} = \frac{\text{D}}{\text{W}} \times 180°$.

In the front-end complex filter estimator, the audio encoding network first reduces the feature dimension of the input vector using a 1-D convolution layer with 256 1×1 kernels (i.e., pointwise convolution) [24], followed by a stack of two successive TCN blocks that consists of dilated 1-D convolution layers with dilation rates exponentially increased from 2^0 to 2^7. The details of TCN are illustrated in Fig. 2. Within each dilated 1-D convolution block, the number of output channels

is set to 512 for the first 1×1 convolution layer, the depth-wise separable convolution layer has a kernel size of 3 with 512 output channels, followed by another 1×1 convolution layer with 256 output channels. After audio encoding network, two separate filter estimator networks are used to estimate the speech and noise (or undesired signal) components. Each filter estimator network consists of two successive TCN blocks (with dilation rates from 2^0 to 2^7) followed by another 1×1 convolution layer, where the number of output channels for the last 1×1 convolution layer depends on the estimation target (i.e., cRM or cRF).

B. Complex ratio filtering

Our recent work [23] suggests that the complex ratio mask (denoted as cRM) can lead to better system performance, the procedure of using cRM to derive the covariance matrix is described below as

$$\hat{\mathbf{X}}_{\mathbf{cRM}}(t, f) = (\text{cRM}_r + j\text{cRM}_i) \cdot (\mathbf{Y}_r + j\mathbf{Y}_i)$$
$$= \text{cRM}_X(t, f) \cdot \mathbf{Y}(t, f),$$
$$\hat{\mathbf{\Phi}}_{\mathbf{XX}}(f) = \frac{\sum_{t=1}^{T} \hat{\mathbf{X}}_{\mathbf{cRM}}(t, f) \hat{\mathbf{X}}_{\mathbf{cRM}}^{\text{H}}(t, f)}{\sum_{t=1}^{T} \text{cRM}_X^{\text{H}}(t, f) \text{cRM}_X(t, f)}, \quad (14)$$

where $\hat{\mathbf{X}}_{\mathbf{cRM}}(t, f)$ represents the estimated MC speech component via the complex speech mask cRM_X. r and i denote the real and imaginary parts, respectively. '·' is the complex multiplier and j is the complex number. The power of the complex mask is used for normalization.

In this study, different from prior neural 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 [51] method for estimation. As depicted in Fig. 5, 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. 5 can be formulated as

$$\hat{\mathbf{X}}_{\mathbf{cRF}}(t, f) = \sum_{\tau_1=-J_1}^{J_2} \sum_{\tau_2=-K_1}^{K_2} \text{cRF}(t, f, \tau_1, \tau_2)$$
$$\cdot \mathbf{Y}(t + \tau_1, f + \tau_2), \quad (15)$$

where $\hat{\mathbf{X}}_{\mathbf{cRF}}$ is the estimated MC speech component using the cRF method, the cRF has a size of $(J_2+J_1+1)\times(K_2+K_1+1)$ for each T-F pixel in the estimated speech. J_1, J_2 and K_1, K_2 represent the number of previous/lower and future/higher frames and frequency bins used for filtering, respectively. The noise or the undesired signal components can be obtained in a similar manner. Different from [51], 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.

C. Multi-channel ADL-MVDR

An accurate estimate of the steering vector is very important for an MC-MVDR system as it contains information about
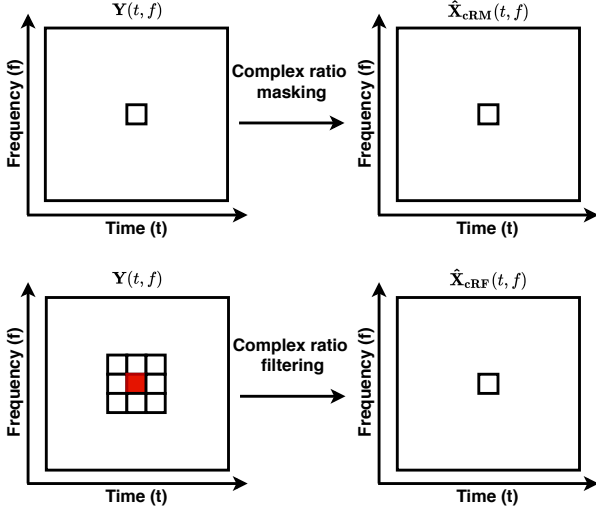
Fig. 5. (Color Online). Schematic diagrams for complex ratio masking and complex ratio filtering. The cRM is a one-to-one mapping, whereas the cRF is a many-to-one mapping. In this example, the cRF has a size of 3×3, where each T-F bin is estimated using nine T-F bins in its neighboring filter range. The center mask is marked with a red fill.

which direction the signal should be preserved [74], [75]. Yet previous approaches involving extracting the principal eigenvector on the speech covariance matrix sometimes could introduce large gradients during back propagation and gradient clipping is often adopted [35], [40], [23]. A similar issue exists for the inversion process of the estimated noise covariance matrix when it becomes singular during joint training [62], [63], [64], where diagonal loading [65] is often used to stabilize the process [76], [40], [23], [64]. In order to perform time-varying beamforming and stabilize the joint training process, we deploy two GRU-Nets to replace the principal eigenvector extraction and matrix inverse involved in MVDR solution. 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. 3.

$$\begin{aligned} \hat{\boldsymbol{v}}(t,f) &= \mathbf{GRU\text{-}Net}_{\boldsymbol{v}}(\tilde{\mathbf{\Phi}}_{\mathbf{XX}}(t',f)), \\ \hat{\mathbf{\Phi}}_{\mathbf{NN}}^{-1}(t,f) &= \mathbf{GRU\text{-}Net}_{\mathbf{NN}}(\tilde{\mathbf{\Phi}}_{\mathbf{NN}}(t',f)), \end{aligned} \quad (16)$$

where $0 \leq t' \leq t$. Leveraging on the temporal properties of RNNs, the frame-wise covariance matrices are fed into the GRU-Nets for MVDR coefficients estimation. The hidden internal states of the GRU-Nets can help capture the previous temporal information in the estimated MVDR coefficients (i.e., steering vector and inverse of the noise covariance matrix). 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 (note they are frame-wise matrices without taking expectation over time) can be obtained from

the cRF estimated speech and noise components as

$$\begin{aligned} \tilde{\mathbf{\Phi}}_{\mathbf{XX}}(t,f) &= \frac{\hat{\mathbf{X}}_{\mathbf{cRF}}(t,f)\hat{\mathbf{X}}_{\mathbf{cRF}}^{\mathrm{H}}(t,f)}{\sum_{t=1}^{T} \mathrm{cRM}_{\mathrm{X}}^{\mathrm{H}}(t,f)\mathrm{cRM}_{\mathrm{X}}(t,f)}, \\ \tilde{\mathbf{\Phi}}_{\mathbf{NN}}(t,f) &= \frac{\hat{\mathbf{N}}_{\mathbf{cRF}}(t,f)\hat{\mathbf{N}}_{\mathbf{cRF}}^{\mathrm{H}}(t,f)}{\sum_{t=1}^{T} \mathrm{cRM}_{\mathrm{N}}^{\mathrm{H}}(t,f)\mathrm{cRM}_{\mathrm{N}}(t,f)}, \end{aligned} \quad (17)$$

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. 5) that is used for normalization, respectively. The same notation also applies to $\mathrm{cRM}_{\mathrm{X}}$. Different from Eq. (14), we do not sum over the temporal dimension to preserve the frame-level information as input to the GRU-Nets.

Based on these RNN-derived frame-wise MVDR coefficients, i.e., $\hat{\mathbf{\Phi}}_{\mathbf{NN}}^{-1}(t,f) \in \mathbb{C}^{M \times M}$ and $\hat{\boldsymbol{v}}(t,f) \in \mathbb{C}^{M}$, the MC MVDR beamforming weights can be derived at frame-level as

$$\hat{\mathbf{h}}_{\text{MC ADL-MVDR}}(t,f) = \frac{\hat{\mathbf{\Phi}}_{\mathbf{NN}}^{-1}(t,f)\hat{\boldsymbol{v}}(t,f)}{\hat{\boldsymbol{v}}^{\mathrm{H}}(t,f)\hat{\mathbf{\Phi}}_{\mathbf{NN}}^{-1}(t,f)\hat{\boldsymbol{v}}(t,f)}, \quad (18)$$

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

$$\hat{\mathrm{X}}_{\text{MC ADL-MVDR}}^{(0)}(t,f) = \hat{\mathbf{h}}_{\text{MC ADL-MVDR}}^{\mathrm{H}}(t,f)\mathbf{Y}(t,f). \quad (19)$$

D. Multi-frame ADL-MVDR

We also introduce an MF setup to our ADL-MVDR framework to simulate an extreme condition when only one channel of the signal is available to the beamforming module. 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.

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

$$\begin{aligned} \hat{\overline{\mathbf{X}}}_{\mathbf{cRF}}^{(0)}(t,f) &= [\hat{\mathrm{X}}_{\mathrm{cRF}}^{(0)}(t-L_1+1,f),...,\hat{\mathrm{X}}_{\mathrm{cRF}}^{(0)}(t,f), \\ &\quad ...,\hat{\mathrm{X}}_{\mathrm{cRF}}^{(0)}(t+L_2,f)]^T, \\ \tilde{\mathbf{\Phi}}_{\mathbf{XX}}^{\mathrm{MF}}(t,f) &= \frac{\hat{\overline{\mathbf{X}}}_{\mathbf{cRF}}^{(0)}(t,f)\hat{\overline{\mathbf{X}}}_{\mathbf{cRF}}^{(0)\,\mathrm{H}}(t,f)}{\sum_{t=1}^{T} \mathrm{cRM}_{\mathrm{X}}^{\mathrm{H}}(t,f)\mathrm{cRM}_{\mathrm{X}}(t,f)}, \end{aligned} \quad (20)$$

where $\hat{\overline{\mathbf{X}}}_{\mathbf{cRF}}^{(0)}$ is the estimated L-frame ($L = L_1 + L_2$, where L_1 and L_2 indicate the number of previous and future frames)

single-channel speech component in T-F domain using the cRF method. $\hat{X}_{\text{cRF}}^{(0)}$ is the estimated single-channel speech component via cRF. The time-varying covariance matrix of the undesired signal $\tilde{\Phi}_{\mathbf{VV}}^{\text{MF}}$ 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 $\tilde{\Phi}^{\text{MF}}(t, f) \in \mathbb{C}^{L \times L}$ estimated via the cRF method. This is formulated as

$$
\begin{aligned}
\hat{\gamma}_{\mathbf{x}}(t, f) &= \mathbf{GRU\text{-}Net}_{\gamma}(\tilde{\Phi}_{\mathbf{XX}}^{\text{MF}}(t', f)), \\
\hat{\Phi}_{\mathbf{VV}}^{\text{MF}-1}(t, f) &= \mathbf{GRU\text{-}Net}_{\mathbf{VV}}(\tilde{\Phi}_{\mathbf{VV}}^{\text{MF}}(t', f)).
\end{aligned}
\tag{21}
$$

Once these variables are obtained, the MF ADL-MVDR filter weights $\hat{\mathbf{h}}_{\text{MF ADL-MVDR}}(t, f) \in \mathbb{C}^{L}$ are applied to the MF noisy speech $\overline{\mathbf{Y}}^{(0)}(t, f) \in \mathbb{C}^{L}$ and the estimated speech are obtained as

$$
\begin{aligned}
\hat{\mathbf{h}}_{\text{MF ADL-MVDR}}(t, f) &= \frac{\hat{\Phi}_{\mathbf{VV}}^{\text{MF}-1}(t, f)\hat{\gamma}_{\mathbf{x}}(t, f)}{\hat{\gamma}_{x}^{\text{H}}(t, f)\hat{\Phi}_{\mathbf{VV}}^{\text{MF}-1}(t, f)\hat{\gamma}_{\mathbf{x}}(t, f)}, \\
\hat{X}_{\text{MF ADL-MVDR}}^{(0)}(t, f) &= \hat{\mathbf{h}}_{\text{MF ADL-MVDR}}^{\text{H}}(t, f)\overline{\mathbf{Y}}^{(0)}(t, f).
\end{aligned}
\tag{22}
$$

E. 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

$$
\begin{aligned}
\hat{\overline{\mathbf{X}}}_{\mathbf{cRF}}(t, f) &= [\hat{\mathbf{X}}_{\mathbf{cRF}}(t - L_1 + 1, f), ..., \hat{\mathbf{X}}_{\mathbf{cRF}}(t, f), \\
&\quad ..., \hat{\mathbf{X}}_{\mathbf{cRF}}(t + L_2, f)]^{T}, \\
\tilde{\Phi}_{\mathbf{XX}}^{\text{MCMF}}(t, f) &= \frac{\hat{\overline{\mathbf{X}}}_{\mathbf{cRF}}(t, f)\hat{\overline{\mathbf{X}}}_{\mathbf{cRF}}^{\text{H}}(t, f)}{\sum_{t=1}^{T} \text{cRM}_{\text{X}}^{\text{H}}(t, f)\text{cRM}_{\text{X}}(t, f)}.
\end{aligned}
\tag{23}
$$

The MCMF estimated noise $\hat{\overline{\mathbf{N}}}_{\mathbf{cRF}}$ component and its time-varying covariance matrix $\tilde{\Phi}_{\mathbf{NN}}^{\text{MCMF}}$ can be estimated using the same method. Once the time-varying speech and noise covariance matrices $\tilde{\Phi}^{\text{MCMF}}(t, f) \in \mathbb{C}^{ML \times ML}$ are obtained, we can follow similar steps as the MC ADL-MVDR described in Eq. (16) to estimate the MCMF steering vector $\hat{v}^{\text{MCMF}}(t, f) \in \mathbb{C}^{ML}$ and inverse of the MCMF noise covariance matrix. After that, the MCMF ADL-MVDR beamforming weights $\hat{\mathbf{h}}_{\text{MCMF ADL-MVDR}}(t, f) \in \mathbb{C}^{ML}$ are derived in a similar manner as described in Eq. (18). Finally, the MCMF ADL-MVDR enhanced speech $\hat{X}_{\text{MCMF ADL-MVDR}}^{(0)}$ is obtained

$$
\hat{X}_{\text{MCMF ADL-MVDR}}^{(0)}(t, f) = \hat{\mathbf{h}}_{\text{MCMF ADL-MVDR}}^{\text{H}}(t, f)\overline{\mathbf{Y}}(t, f),
\tag{24}
$$

where $\overline{\mathbf{Y}}(t, f) \in \mathbb{C}^{ML}$ is the MCMF noisy speech.

IV. EXPERIMENTAL SETUP

A. Speech Materials

We adopt a Mandarin audio-visual speech corpus (will be released soon) collected from Youtube, which has been reported in our prior works [23], [24], [50]. An SNR estimator together with a face detector is used to filter out the low-quality ones [24], [50]. A total number of 205,500 clean video segments from around 1500 speakers are gathered. In contrast to our prior works [24], [50], 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 455 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 and the noise source contains random clips from 255 noises recorded indoors.

To generate the multi-channel data, the microphone array and all the sound sources are randomly placed in a simulated room, the distance between any sound source and the microphone array is between 0.5 m to 6 m [24]. The number of overlapped speakers ranges from 1 to 3. We also consider a wide range of reverberant conditions, where 2000 different rooms with 6000 room impulse responses (RIRs) are simulated via the image method [77]. The room size is randomly chosen from 4 m × 4 m × 3 m to 10 m × 10 m × 6 m, with RT60s ranging from 0.05 s to 0.7 s. As a result, the simulated multi-channel speech data has an SNR range of 18 to 30 dB and the signal-to-interference ratio (SIR) is between -6 to 6 dB.

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 entire system is trained to minimize the time-domain scale-invariant source-to-noise ratio (Si-SNR) loss [78], which is the negative of Si-SNR, i.e.,

$$
\begin{aligned}
\mathcal{L}_{\text{Si-SNR}} &= -20\log_{10}\frac{\|\alpha \cdot \mathbf{x}\|}{\|\hat{\mathbf{x}} - \alpha \cdot \mathbf{x}\|}, \\
\alpha &= \frac{\hat{\mathbf{x}}^{\text{T}}\mathbf{x}}{\mathbf{x}^{\text{T}}\mathbf{x}},
\end{aligned}
\tag{25}
$$

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 [22] for complex filter estimation, which contains bunch of dilated 1-D convolution networks together with a pair of fixed STFT/iSTFT encoder/decoder implemented with 1-D convolution layers [24], [50]. Details on the audio encoding network is introduced in Section III-A. In this study, we focus on the ADL-MVDR framework which can be adapted 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 **GRU-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 **GRU-Net$_v$** 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 eight T-F bins as depicted in Fig. 5) where it can utilize temporal information from one previous frame to one future frame and nearby frequency information from one frequency bin below to one frequency bin above. Note that the number of the output channels of the final 1×1 convolution layer in the filter estimator networks could be changed accordingly for different configurations of cRF sizes. For example, we set the number of output channels to 257×9 for a 3×3 cRF.

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 an MF size of five, i.e., from two previous frames to two 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 **GRU-Net$_\gamma$** and 50 units for **GRU-Net$_{VV}$**.

To investigate the influence of incorporating additional MF information on top of the MC spatial information, a 9-channel (i.e., mics: 0, 2, 3, 5, 7, 9, 11, 12, 14) 3-frame (i.e., from one previous frame to one future frame) MCMF ADL-MVDR system is included. Here the **GRU-Net$_{NN}$** consists of two GRU layer with 500 units each, followed by another 1458-unit fully connected layer. The **GRU-Net$_v$** 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 for MCMF ADL-MVDR systems, including when only three (i.e., mics: 0, 7, 14), seven (i.e., mics: 0, 3, 5, 7, 9, 11, 14), nine (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 three (i.e., from one previous frame to one future frame) and two (i.e., one 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 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. The statistics on model size, speed and memory usage are further provided. These results are reported in the following section. Note that the current configurations of ADL-MVDR systems are not specifically designed for online processing, although the audio encoding network and the ADL-MVDR beamformer could be configured to a causal system for online processing.

D. Evaluation Metrics

A set of objective evaluation metrics are used to evaluate the systems' performance from different perspectives. These metrics include PESQ [79], source-to-distortion ratio (SDR) [80] for speech quality assessment and STOI for intelligibility estimation [81]. The Si-SNR score is also included as it has been utilized for many recent speech separation systems [22], [82], [23]. Moreover, we use a Tencent commercial speech recognition API [83] (based on deep feed-forward sequential memory networks [84]) 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., the audio encoding network described in Section III-A, denoted as NN with cRM/cRF) are also included as baselines. We further include conventional neural mask-based MC-MVDR systems [40] and multi-tap MVDR systems [23] for comparison approaches (denoted as MVDR with cRM/cRF and multi-tap MVDR with cRM/cRF), they incorporate the same front-end network structure for mask/filter estimation as the ADL-MVDR systems, followed by conventional segment-level MVDR beamforming [40], [23]. Furthermore, we implement another MVDR systems with recursive updating rules for mini-block covariance matrices as described in [32], [45] (denoted as MVDR with cRF - recursive), where the mini-block size is set to 30 frames with hop size of 10 frames. For fair comparison, the same front-end complex filter estimator was used for speech and noise covariance matrices estimation (with 3×3 cRF), the center mask is used to compute the transition factors between successive mini-blocks. Note that all of our implemented conventional neural mask-based MVDR systems feature an end-to-end training scheme similar to [85], [61], [40], [23], we further adopt diagonal loading [65] and gradient clipping [35], [23] to alleviate the numerical instability issue during joint training.

In Table I, 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, SDR and STOI) for brevity.

In the ablation studies, we report the performance with a set of speech evaluation metrics (e.g., PESQ, Si-SNR) as well as WER. 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

---

[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, SDR AND STOI 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)	STOI (%)	WER (%)
	0-15°	15-45°	45-90°	90-180°	1spk	2spk	3spk	Avg.	Avg.	Avg.	Avg.	Avg.
Reverberant clean (reference)	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50	∞	∞	100	8.26
Noisy Mixture	1.88	1.88	1.98	2.03	3.55	2.02	1.77	2.16	3.39	3.50	70.8	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	88.6	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	89.2	22.07
MF ADL-MVDR	**2.80**	**2.99**	**3.16**	**3.11**	**4.01**	**3.10**	**2.80**	**3.14**	**12.60**	**13.17**	**89.5**	**19.57**
Conventional neural mask-based MVDR systems and our proposed MC ADL-MVDR system												
MVDR with cRF - recursive [32]	2.56	2.75	2.93	2.90	3.75	2.87	2.56	2.90	10.12	11.10	89.0	16.74
MVDR with cRM [23]	2.55	2.76	2.96	2.84	3.73	2.88	2.56	2.90	10.62	12.04	88.5	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	88.9	15.91
Multi-tap MVDR with cRM [23]	2.70	2.96	3.18	3.09	3.80	3.07	2.74	3.08	12.56	14.11	91.5	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	91.4	13.52
MC ADL-MVDR	**3.04**	**3.30**	**3.48**	**3.48**	**4.17**	**3.41**	**3.07**	**3.42**	**14.80**	**15.45**	**93.3**	**12.73**
Our proposed MCMF ADL-MVDR system												
MCMF ADL-MVDR	**3.10**	**3.35**	**3.48**	**3.51**	**4.20**	**3.47**	**3.10**	**3.46**	**15.43**	**16.03**	**93.7**	**12.31**
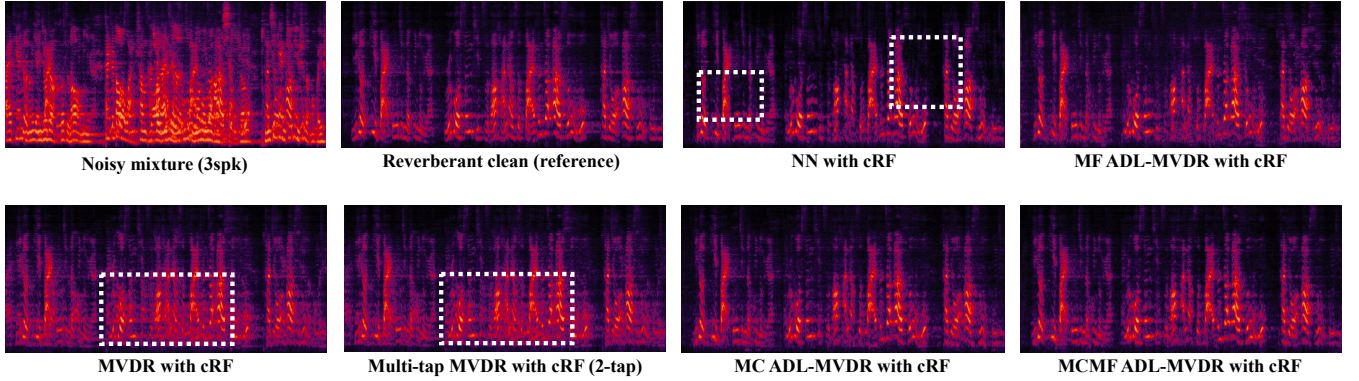
Fig. 6. (Color Online). Spectrograms of some evaluated systems in Table I, the nonlinear distortion and residual noise are highlighted by the dashed boxes.

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. Lastly, Table V provides statistics on the model size, running speed and memory usage for a set of evaluated systems.

A. Overview Results on ADL-MVDR Systems

MC ADL-MVDR vs. Conventional neural 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 neural mask-based 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). Similar patterns could be

found for intelligibility scores, where our proposed MC ADL-MVDR system achieves around 5% better performance than recursive and chunk-level MVDR systems (e.g., STOI: 93.3% vs. 89.0% and 88.9%), respectively.

Compared to MVDR with cRF (i.e., chunk-level covariance estimation), the MVDR with recursive updating rules results in slightly worse performance (e.g., PESQ; 2.90 vs. 2.92, Si-SNR: 10.12 dB vs. 11.31 dB), which might be caused by the less accurate estimations of the covariance matrices within a mini-block rather than using a full chunk of audio. Yet it still achieves significant ASR gains compared to the purely NNs (i.e., 16.74% vs. 22.49% and 22.07%). Note that the residual noise in recursive MVDR and conventional neural mask-based MVDR systems is still at relatively high level according to their PESQ scores (i.e., average scores of 2.90 and 2.92, respectively) when compared to the purely NN systems (e.g., NN with cRF: 3.10). The multi-tap MVDR with cRF can better remove the residual noise than MVDR with cRF, however, the residual noise is still high compared to MC ADL-MVDR system. 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. 6, the conventional neural mask-based 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 neural mask-based MVDR system and our proposed MC ADL-MVDR system is provided in Fig. 7. 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 is 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, SDR: 13.17 dB vs. 13.01 dB, and STOI: 89.5% vs. 89.2%) when only one 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. The purely NN systems (i.e., without MVDR) usually come with a relatively higher degree of nonlinear distortion on separated speech as they only focus on noise reduction [86], [87]. As reflected in Fig. 6, the highlighted spectral 'black holes' (i.e., zero or close to zero values on spectrogram) is an example of the purely NN systems where too much speech information is removed when suppressing the noise, where modern ASR systems are known to be sensitive to this type of distortion [88], [86]. Although not provided with spatial information for the beamforming module, 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 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, SDR: 16.03 dB vs. 15.45 dB, and STOI: 93.7% vs. 93.3%), suggesting that MCMF ADL-MVDR system can even better remove the residual noise and generate more intelligible speech signals than MC ADL-MVDR systems. Slightly better performance is
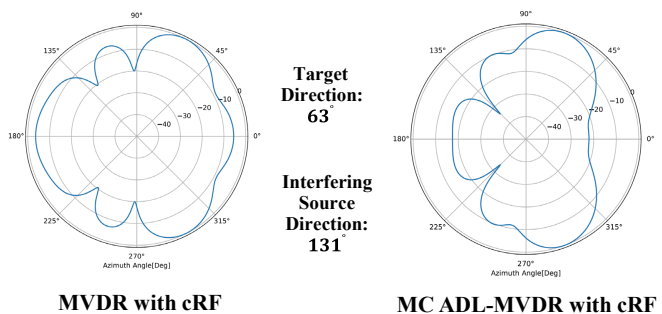

Fig. 7. (Color Online). Beam pattern examples for conventional neural mask-based MVDR system and our proposed MC ADL-MVDR system. RT60 = 0.31 s. Target source at 63°, interfering source at 131°.

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**

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

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/LOWER AND
FUTURE/HIGHER 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

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

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 two previous frames and the other one with temporal information from one previous frame to one future frame (i.e., the 3rd and 5th MC ADL-MVDR systems), slight improvements can be 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. In general, incorporating future T-F pixels to the cRF can improve the systems performance but in a limited scale, which could be caused by the non-causality of the complex filter estimator that already uses some future information.

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%).

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 five (i.e., from two previous frames to two 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

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 six different setups where the first two system setups represent the conditions

TABLE V
MODEL SIZE, MEMORY USAGE AND RUNNING SPEED INFORMATION OF SOME EVALUATED SYSTEMS.

Systems	Model Size	Training Mem. (MB)	Inference Mem. (MB)	Training Speed (ms)	Inference Speed (ms)	PESQ	Si-SNR (dB)	WER (%)
Purely NNs and our proposed MF ADL-MVDR System								
NN with cRM	**16.40 M**	**176.4**	**67.9**	**53.7**	**13.5**	3.07	12.23	22.49
NN with cRM (large)	31.62 M	335.0	126.3	54.5	14.1	3.07	12.26	22.06
NN with cRF	17.46 M	193.3	70.3	66.4	15.7	3.10	12.50	22.07
MF ADL-MVDR	18.99 M	288.5	78.7	108.4	20.1	**3.14**	**12.60**	**19.57**
Conventional neural mask-based MVDR and our proposed MC/MCMF ADL-MVDR systems								
MVDR with cRF - recursive [32]	**18.32 M**	281.2	74.1	174.8	31.3	2.90	10.12	16.74
MVDR with cRF	**18.32 M**	282.8	74.5	**83.1**	**16.4**	2.92	11.31	15.91
MVDR with cRF (large)	35.84 M	547.1	141.5	83.7	17.0	2.94	11.38	15.75
Multi-tap MVDR with cRF	**18.32 M**	284.4	74.6	98.4	20.8	3.08	12.66	13.52
MC ADL-MVDR (compact)	**18.32 M**	**280.5**	**74.0**	131.9	27.0	3.37	14.36	13.01
MC ADL-MVDR	23.47 M	359.9	93.6	222.2	38.9	3.42	14.80	12.73
MCMF ADL-MVDR	27.01 M	416.2	106.4	302.6	55.3	**3.46**	**15.43**	**12.31**

when only information from previous frames is available, and the last four MF conditions further include information from future frames. By comparing the first two 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 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 seven to nine (i.e., the last two setups in Table IV), no further improvements are observed (e.g., PESQ: 3.17 vs. 3.18 and WER: 19.52% vs. 19.45%). Better performance can be achieved by increasing the MF size, however, additional information from future frames did not make significant contributions to the performance, which is likely caused by the non-causality of the current front-end audio encoding network.

E. Ablation Study on Model Statistics

Comparison results of the statistics in the systems' training and inference stages are reported in Table V. A single Nvidia Tesla V100 GPU is used to measure the statistics, where we set the batch size to 1 for evaluation in both training and inference stage. The model size is first reported by the total number of parameters in a model, memory usage information[2] (training and inference stages) and running speed (average running time for processing 1 s of audio input) are also included. The systems' performance on PESQ, Si-SNR and WER are provided for comparison.

[2] pytorch_memlab, https://github.com/Stonesjtu/pytorch_memlab

The purely NN with cRM has the smallest model size (i.e., 16.40 M) since it does not come with a cRF estimator or RNN-based beamforming module. The proposed MF ADL-MVDR with cRF has slightly larger model size compared to NN with cRF (i.e., 18.99 M vs. 17.46 M), but with similar inference memory (i.e., 78.7 MB vs. 70.3 MB), inference speed (20.1 ms vs. 15.7 ms) while better performance in speech quality (e.g., PESQ: 3.14 vs. 3.10) and ASR accuracy (i.e., 19.57% vs. 22.07%).

The proposed MC ADL-MVDR systems utilize additional spatial information (i.e., microphone channels) in the GRU-Nets that leads to increased model size compared to conventional neural mask-based MVDR systems (e.g., 23.47 M vs. 18.32 M) that is based on mathematical derivation. It obtains much better performance compared to multi-tap MVDR system on the separated speech quality (i.e., PESQ: 3.42 vs. 3.08, Si-SNR: 14.80 dB vs. 12.66 dB) and WER (i.e., 12.73% vs. 13.52%), at the cost of additional inference memory (i.e., 93.6 MB vs. 74.6 MB) and inference time (i.e., 38.9 ms vs. 20.8 ms). When more computation resources are available, the MCMF ADL-MVDR system leads to the best performance (i.e., PESQ; 3.46, Si-SNR: 15.43 dB and WER 12.31%) with a model size of 27.01 M and slightly increased inference memory and speed compared to MC ADL-MVDR (i.e., 106.4 MB and 55.3 ms, respectively).

To investigate whether the performance gain is introduced by increased number of network parameters, we further include three models, (1) a lager purely NN with cRM, denoted with NN with cRM (large), (2) a larger neural mask-based MVDR system with cRF, denoted as MVDR with cRF (large), and (3) a compact version of the MC ADL-MVDR system, denoted as MC ADL-MVDR (compact). For the two enlarged models, we increase the embedding dimension of the audio encoding block from 256 to 512, with everything else remained unchanged. Meanwhile, for MC ADL-MVDR (compact), we reduce the hidden feature space of the GRU-Nets (i.e., 250 and 50 for **GRU-Net**$_v$, 250 for **GRU-Net**$_{NN}$) and the cRF region (i.e., 3×1 cRF, only using nearby temporal information).

As showed in Table V, simply increasing the network size does not lead to major improvements on the systems' performance. For purely NNs with cRM, we have PESQ: 3.07

vs. 3.07, Si-SNR: 12.26 dB vs. 12.23 dB and WER: 22.06% vs. 22.49% between the large and original models. On the other hand, for MVDR systems with cRF, increasing the model size leads to 0.02 and 0.07 dB increments in PESQ and Si-SNR, respectively. Even with a reduced model size in MC ADL-MVDR (compact), it still achieves relative 15%, 27% and 18% gains on PESQ, Si-SNR, and WER compared to MVDR with cRF. For fair comparison on model statistics, the proposed ADL-MVDR systems should be compared with the recursive MVDR system, as they estimate time-varying beamforming weights and can be adapted for online processing with streaming input. Note that conventional chunk-level MVDR systems cannot be directly used for this purpose as they require a chunk of audio input to derive the beamforming weights. When compared to MVDR system with recursive updating rules, the MC ADL-MVDR (compact) achieves faster inference speed (i.e., 27.0 ms vs. 31.3 ms), lower inference memory (i.e., 74.0 MB vs. 74.1 MB) while gaining substantial gains on speech quality (e.g., PESQ: 3.37 vs. 2.90) and ASR accuracy (13.01% vs. 16.74%).

Results here indicate that improvements are limited for purely NNs and conventional neural mask-based MVDR systems with increased model size, and our proposed MC ADL-MVDR system still achieves substantial improvements than its peers with the same model size. We infer that the additional inference time and memory cost in our proposed ADL-MVDR systems compared to purely NNs and conventional neural mask-based MVDR systems are coming from the RNN-based networks in the ADL-MVDR module. Moreover, when compared to conventional neural mask-based MVDR system with recursive updating rules, the proposed ADL-MVDR systems consume similar computing resources and time while delivering much better performance on speech quality as well as the ASR accuracy.

VI. Conclusions and Future Work

In this work, we proposed an ADL-MVDR framework that can be configured and applied for multi-channel, multi-frame, and multi-channel multi-frame target speech separation tasks. Our proposed ADL-MVDR systems have achieved the best performance in terms of objective speech quality and intelligibility, 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 bypasses the numerical instability issue that occurs in conventional neural mask-based MVDR systems during joint training with neural networks. Additionally, the proposed ADL-MVDR systems keep the residual noise at a minimum level (reflected by highest objective scores) while introducing hardly any nonlinear distortions (reflected by lowest WER). However, one limitation of the current ADL-MVDR framework is that it requires additional computation resources compared to conventional neural mask-based MVDR approaches.

Future work of this study may include the following directions. First, we will evaluate this ADL-MVDR framework on purely single-channel speech separation tasks. Second, the complex ratio filter and multi-frame sizes are determined empirically in this study, we will further explore approaches to control these sizes adaptively. Third, we will adapt this ADL-MVDR framework for joint separation and dereverberation tasks. Fourth, we will explore a more generalized solution of neural beamformer that is independent of microphone geometries. Lastly, we will further explore on different network structures to provide a more efficient model, for instance, we could try replacing the RNN-based networks with feed-forward mechanisms for improved efficiency.

References

[1] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.
[2] 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.
[3] 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.
[4] 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.
[5] 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.
[6] 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.
[7] 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.
[8] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE TASLP*, vol. 22, no. 12, pp. 1849–1858, 2014.
[9] 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.
[10] 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," *Interspeech*, 2020.
[11] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *speech communication*, vol. 53, no. 4, pp. 465–494, 2011.
[12] 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.
[13] 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.
[14] Z. Zhang, D. S. Williamson, and Y. Shen, "Investigation of phase distortion on perceived speech quality for hearing-impaired listeners," *Interspeech*, 2020.
[15] 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.
[16] 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.
[17] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.
[18] Z.-Q. Wang, J. Le Roux, D. Wang, and J. R. Hershey, "End-to-end speech separation with unfolded iterative phase reconstruction," *Interspeech*, 2018.
[19] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *ICASSP*, 2018, pp. 696–700.

[20] 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.

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

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

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

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] ——, "Sensitivity analysis of the multi-frame MVDR filter for single-microphone speech enhancement," in *EUSIPCO*. IEEE, 2017, pp. 603–607.

[40] 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.

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

[42] E. A. Habets and J. Benesty, "A two-stage beamforming approach for noise reduction and dereverberation," *IEEE TASLP*, vol. 21, no. 5, pp. 945–958, 2013.

[43] 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.

[44] M. Taseska and E. A. Habets, "Nonstationary noise PSD matrix estimation for multichannel blind speech extraction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2223–2236, 2017.

[45] T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, and T. Nakatani, "Online mvdr beamformer based on complex gaussian mixture model with spatial prior for noise robust ASR," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 780–793, 2017.

[46] T. Higuchi, K. Kinoshita, N. Ito, S. Karita, and T. Nakatani, "Frame-by-frame closed-form update for mask-based adaptive MVDR beamforming," in *ICASSP*, 2018, pp. 531–535.

[47] S. Chakrabarty and E. A. Habets, "Time–frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 787–799, 2019.

[48] J. M. Martín-Doñas, J. Jensen, Z.-. Tan, A. M. Gomez, and A. M. Peinado, "Online multichannel speech enhancement based on recursive EM and DNN-based speech presence estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 3080–3094, 2020.

[49] 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," *ICASSP*, pp. 6089–6093, 2021.

[50] 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.

[51] 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.

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

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

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

[55] 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.

[56] 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.

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

[58] Y. Liu, A. Ganguly, K. Kamath, and T. Kristjansson, "Neural network based time-frequency masking and steering vector estimation for two-channel MVDR beamforming," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6717–6721.

[59] 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.

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

[61] 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.

[62] S. Zhao and D. L. Jones, "A fast-converging adaptive frequency-domain mvdr beamformer for speech enhancement," in *Interspeech*, 2012.

[63] C. Y. Lim, C.-H. Chen, and W.-Y. Wu, "Numerical instability of calculating inverse of spatial covariance matrices," *Statistics & Probability Letters*, vol. 129, pp. 182–188, 2017.

[64] W. Zhang, C. Boeddeker, S. Watanabe, T. Nakatani, M. Delcroix, K. Kinoshita, T. Ochiai, N. Kamo, R. Haeb-Umbach, and Y. Qian, "End-to-end dereverberation, beamforming, and speech recognition with improved numerical stability and advanced frontend," in *ICASSP*. IEEE, 2021, pp. 6898–6902.

[65] X. Mestre and M. A. Lagunas, "On diagonal loading for minimum variance beamformers," in *ISSPIT*. IEEE, 2003, pp. 459–462.

[66] 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.

[67] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1800–1807.

[68] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[69] 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.

[70] 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.

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

[72] B. Li, T. N. Sainath, R. J. Weiss, K. W. Wilson, and M. Bacchiani, "Neural network adaptive beamforming for robust multichannel speech recognition," in *Interspeech*, 2016, pp. 1976–1980.

[73] 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.

[74] 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.

[75] 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.

[76] Y. L. Chen and J.-H. Lee, "Finite data performance analysis of mvdr antenna array beamformers with diagonal loading," *Progress In Electromagnetics Research*, vol. 134, pp. 475–507, 2013.

[77] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

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

[79] 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.

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

[81] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[82] 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.

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

[84] S. Zhang, M. Lei, Z. Yan, and L. Dai, "Deep-FSMN for large vocabulary continuous speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5869–5873.

[85] J. Heymann, L. Drude, C. Boeddeker, P. Hanebrink, and R. Haeb-Umbach, "BEAMNET: End-to-end training of a beamformer-supported multi-channel ASR system," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5325–5329.

[86] Z.-Q. Wang, P. Wang, and D. Wang, "Complex spectral mapping for single-and multi-channel speech enhancement and robust ASR," *IEEE TASLP*, vol. 28, pp. 1778–1787, 2020.

[87] T. Ochiai, S. Watanabe, T. Hori, J. R. Hershey, and X. Xiao, "Unified architecture for multichannel end-to-end speech recognition with neural beamforming," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1274–1288, 2017.

[88] N. Tawara, T. Kobayashi, M. Fujieda, K. Katagiri, T. Yazu, and T. Ogawa, "Adversarial autoencoder for reducing nonlinear distortion," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2018, pp. 1669–1673.