Neural Spectrospatial Filtering

Ke Tan, Zhong-Qiu Wang, and DeLiang Wang, *Fellow, IEEE*

***Abstract*—As the most widely-used spatial filtering approach for multi-channel speech separation, beamforming extracts the target speech signal arriving from a specific direction. An emerging alternative approach is multi-channel complex spectral mapping, which trains a deep neural network (DNN) to directly estimate the real and imaginary spectrograms of the target speech signal from those of the multi-channel noisy mixture. In this all-neural approach, the trained DNN itself becomes a nonlinear, time-varying spectrospatial filter. However, it remains unclear how this approach performs relative to commonly-used beamforming techniques on different array configurations and acoustic environments. This paper is devoted to examining this issue in a systematic way. Comprehensive evaluations show that multi-channel complex spectral mapping achieves separation performance comparable to or better than beamforming for different array geometries and speech separation tasks and reduces to monaural complex spectral mapping in single-channel conditions, demonstrating the general utility of this approach on multi-channel and single-channel speech separation. In addition, such an approach is computationally more efficient than widely-used mask-based beamforming. We conclude that this neural spectrospatial filter provides a strong alternative to traditional and mask-based beamforming.**

***Index Terms*—Beamforming, deep learning, multi-channel complex spectral mapping, spectrospatial filtering, speech separation.**

I. Introduction

SPATIAL filtering refers to microphone array processing that applies a filter to a multi-channel signal acquired by microphones spatially distributed in the physical space. In the context of multi-channel speech separation, spatial filtering is widely used to enhance a target speech source originating from a specific spatial location and suppress the signals from interfering sound sources from other locations. Such selective processing is based on distinct microphone positions relative to a radiating sound source, forming the geometry of a microphone array, i.e. the number and spatial arrangement of microphones.

Manuscript received September 3, 2021; revised December 23, 2021; accepted January 16, 2022. Date of publication January 25, 2022; date of current version February 2, 2022. This work was supported in part by the NSF under Grant ECCS-1808932 and in part by the Ohio Supercomputer Center. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiao-Lei Zhang. (*Corresponding author: Ke Tan.*)

Ke Tan is with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210-1277 USA (e-mail: tan.650@osu.edu).

Zhong-Qiu Wang is with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210-1277 USA, and also with the Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: wang.zhongqiu41@gmail.com).

DeLiang Wang is with the Department of Computer Science and Engineering and the Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, OH 43210-1277 USA (e-mail: dwang@cse.ohio-state.edu).

Digital Object Identifier 10.1109/TASLP.2022.3145319

Over the past decades, the most dominant spatial filtering approach is acoustic beamforming, which applies a linear filter to boost the signal from a specific direction while attenuating signals from other directions [26], [46]. A beamformer has different sensitivities to different arriving directions of signals, yielding a beam pattern with high sensitivity (i.e. main beam) towards a certain direction. This steering (beam) direction can be manipulated by, e.g., delaying a microphone signal by a certain time [46]. Various beamforming techniques have been developed in the signal processing community. These techniques can be broadly categorized into two classes. The first class is fixed beamformers, of which the filter coefficients are independent of the microphone signals. The simplest fixed beamformer uses a delay-and-sum technique consisting of two steps [11]. First, each microphone signal is time-shifted to compensate for the time difference of arrival (TDOA) between that microphone and a reference microphone. These time-shifted signals are subsequently summed to produce the beamformer output, where temporally aligned signals produce the highest output. The second class of beamforming techniques adaptively estimates the filter coefficients based on the signal characteristics. Such adaptive beamformers typically use the minimum variance principle first introduced by Capon [5]. A widely-used adaptive beamformer is the minimum variance distortionless response (MVDR) beamformer, which minimizes the average energy of the beamformer output while preserving the signal from the target direction (i.e. distortionless) [5], [16]. The use of a conventional beamformer requires the knowledge of the relative transfer function (RTF) between microphones, which must be estimated if not known *a priori* [16]. In reverberant and multi-source environments, however, accurate RTF estimation is fundamentally challenging.

Since the formulation of speech separation as supervised learning [51], data-driven methods have been extensively studied. In particular, deep learning has become the mainstream methodology of supervised speech separation, and remarkably advanced the separation performance over the past decade [52]. To address multi-channel speech separation, two independent studies [21], [22] first developed mask-based beamforming. In [21], Heymann *et al.* proposed to combine conventional beamforming with deep learning based mask estimation. They employed a DNN to estimate the ideal binary mask (IBM) from a monaural input for each channel. The estimated masks are then used to compute the steering vector and the spatial covariance matrix of noise, from which the beamformer coefficients are derived. Such an approach provides more accurate estimation of the steering vector and the noise covariance matrix than conventional algorithms. Other related studies include [12], [13], [64], [66], [68]. As in [21], these methods only utilize

magnitude-domain spectral features to estimate a magnitude-domain ideal time-frequency (T-F) mask. Subsequent studies additionally leverage spatial features, and provide more robust mask estimation [6], [54], [55], [65]. In multi-channel speech separation, mask-based beamforming has become the mainstream approach. In addition, some effort has been recently made on neural beamforming, which directly learns a set of beamforming filters using a DNN in either the time domain [28], [29], [39], [40] or the frequency domain [19], [24], [32], [35], [62], [63], [69].

Given the importance of phase in speech processing [33], [34], [36], complex-domain approaches have been studied for monaural separation [14], [43], [61]. In [43], we obtained significant improvements in objective intelligibility and perceptual quality of speech by performing complex spectral mapping, which directly estimates the real and imaginary spectrograms of target speech from those of the noisy mixture. For monaural separation, the advantage of complex spectral mapping over magnitude-domain approaches is two-fold. First, complex spectral mapping enhances phase in addition to magnitude, which benefits the speech quality. Second, real and imaginary spectrograms are used as the DNN input, more informative than magnitude-domain spectral features. To address multi-channel separation, Wang *et al.* [57] combined complex spectral mapping and adaptive beamforming. Specifically, monaural complex spectral mapping is performed on each channel individually, and the estimated complex spectrograms of the target speech and interference are used to directly compute the spatial covariance matrices. Experimental results show that this approach improves over mask-based beamforming and advances the state-of-the-art automatic speech recognition results on the CHiME-4 corpus [49].

Moreover, most studies on mask-based beamforming (e.g. [21], [57]) assume an unknown array geometry, and aim to derive a separation system applicable to arbitrary array configurations. While aiming for generality is admirable, the array geometry is fixed in many real-world applications (e.g. Amazon Echo), and one can potentially leverage this fixed geometry in supervised multi-channel separation.

With the fixed array geometry assumption, multi-channel complex spectral mapping (MC-CSM) has been recently developed [56]. Specifically, the real and imaginary spectrograms of the multi-channel mixture are concatenated and fed into a DNN to estimate those of the target speech signal. Rather than explicitly applying a beamformer, such an approach trains the DNN itself to become a filter. The MC-CSM approach has recently been shown to be effective in different speech separation tasks, including speech enhancement [44], speech dereverberation [56] and speaker separation [58]. The evaluation results in these studies show that MC-CSM can achieve much better results than traditional beamforming, and comparable or better separation performance compared to mask-based beamforming.

For fixed-geometry arrays, why would we expect this straightforward, all-neural approach be effective? With all available information encoded in the multi-channel complex spectrograms of the noisy mixture, this approach has the potential to extract all discriminative cues contained therein through deep learning, including both spectral and spatial cues. These cues could greatly benefit the extraction of a target signal arriving from a certain direction, containing monaural spectral characteristics, or both. Hence we call this processing spectrospatial filtering. It is worth emphasizing that the complex representation embodied in MC-CSM is inherently sensitive to phase relations between different microphones. We thus believe that, with MC-CSM, a DNN can learn to implicitly determine this target direction and enhance the signal coming from that direction by harnessing spatial and spectral information synergistically. The key is to train the DNN with a wide range of sound source positions and room acoustic properties while using a fixed array geometry, which provides stable inter-microphone phase relations to be captured by supervised learning.

Although MC-CSM has been explored in a few recent studies [44], [56], [58], its efficacy in different array configurations and acoustic environments has not yet been systematically investigated. The present study examines this approach with different array geometries in various multi-channel speech separation tasks. We comprehensively compare the approach with commonly-used beamforming techniques, including conventional and mask-based beamforming. Our experimental results show that the approach achieves separation performance comparable to or better than beamforming for various speech separation tasks and array geometries. This demonstrates that the MC-CSM approach with fixed-geometry arrays amounts to neural spectrospatial filtering that is both effective and general for speech separation, hence providing a major alternative to acoustic beamforming.

The rest of this paper is organized as follows. In Section II, we describe the signal model and formulate the multi-channel speech separation problem. In Section III, we briefly review several widely-used beamforming techniques. Section IV describes key components of the MC-CSM approach. Experimental setup is provided in Section V. We present and analyze experimental results in Section VI. Section VII concludes this paper.

II. Signal Model and Problem Formulation

In a noisy and reverberant environment, the signals received by a P-channel microphone array can be modeled as [15], [31], [48]

$$\begin{aligned}\mathbf{Y}(t,f) &= \mathbf{d}_q(f)S_q(t,f) + \mathbf{H}(t,f) + \mathbf{N}(t,f),\\ &= \mathbf{S}(t,f) + \mathbf{V}(t,f),\end{aligned} \tag{1}$$

where S_q is the short-time Fourier transform (STFT) of the target speech signal picked up by a reference microphone (the q-th microphone), and symbols t and f index the time frame and the frequency bin, respectively. Here $\mathbf{d}_q \in \mathbb{C}^{P\times 1}$ is the RTF vector for the target source with respect to the q-th microphone. Symbols $\mathbf{H}$ and $\mathbf{N}$ denote the STFTs of target speech reverberation and reverberant noise, respectively, and $\mathbf{S}(t,f) = \mathbf{d}_q(f)S_q(t,f)$ and $\mathbf{V}(t,f) = \mathbf{H}(t,f) + \mathbf{N}(t,f)$ the STFTs of the target and nontarget signals picked by the microphone array, respectively. Note that the reverberant noise or interference $\mathbf{N}$ may contain signals coming from multiple environmental noises, which may be directional or diffuse, and interfering speakers.

We assume that sound sources do not move within the duration of a single utterance.

In general, the goal of speech separation is to estimate S_q, which is anechoic, given the multi-channel mixture $\mathbf{Y}$ received by a microphone array. One may design a spectrospatial filter to estimate the target signal:

$$\hat{S}_q = \mathcal{F}(\mathbf{Y}), \tag{2}$$

where $\mathcal{F}$ is the mapping function represented by the neural spectrospatial filter.

III. Beamforming

For broadband signals, a beamformer typically identifies a frequency-dependent complex-valued weight vector $\mathbf{w}(t,f) = [w_1(t,f), \ldots, w_P(t,f)]^T$. The weight vector of a time-invariant beamformer can be expressed as $\mathbf{w}(f) = [w_1(f), \ldots, w_P(f)]^T$. Following (2), the output of the beamformer is given by

$$\hat{S}_q(t,f) = \mathcal{F}(\mathbf{Y}(t,f); \mathbf{w}(f)) = \mathbf{w}(f)^{\mathrm{H}}\mathbf{Y}(t,f), \tag{3}$$

where $(\cdot)^{\mathrm{H}}$ represents the conjugate transpose. Thus to determine the value of the weight vector $\mathbf{w}$ is the key to beamforming. Below we describe delay-and-sum and MVDR as representative fixed and adaptive beamformers, respectively.

A. Conventional Beamforming

The formulation of a beamformer requires the determination of the steering direction, represented as a steering vector. Typically, the RTF vector is chosen as the steering vector. This vector can be calculated from the direction-of-arrival (DOA) of the target source, if known, or estimated from the microphone signals [8], [9], [15].

1) Delay and Sum: A widely-used fixed beamformer is the delay-and-sum (DS) beamformer, of which the weight vector is derived directly from the DOA. The target speech signals captured by different microphones exhibit similar waveforms but different time delays (or phases). The relative delay between each microphone and the reference microphone can be determined from the DOA and the inter-microphone distance. This relative delay is compensated by time-shifting the corresponding microphone signal by a certain time. These time-shifted signals are then coherently summed. Equivalently, the weight vector can be derived by maximizing the white noise gain in the frequency domain [3].

2) MVDR: The popular MVDR aims to minimize the output power with the constraint that the signal arriving from the target direction is not distorted. This optimization problem is mathematically formulated as

$$\min_{\mathbf{w}(f)} \mathbf{w}(f)^{\mathrm{H}}\mathbf{\Phi}_{\mathbf{v}}(f)\mathbf{w}(f) \text{ subject to } \mathbf{w}(f)^{\mathrm{H}}\mathbf{d}_q(f) = 1, \tag{4}$$

where $\mathbf{\Phi}_{\mathbf{v}}$ is the spatial covariance matrix of the nontarget signals $\mathbf{V}$. The weight vector of the MVDR beamformer is determined by solving the optimization problem:

$$\mathbf{w}(f) = \frac{\mathbf{\Phi}_{\mathbf{v}}(f)^{-1}\mathbf{d}_q(f)}{\mathbf{d}_q(f)^{\mathrm{H}}\mathbf{\Phi}_{\mathbf{v}}(f)^{-1}\mathbf{d}_q(f)}. \tag{5}$$

B. Deep Learning Based Beamforming

1) Mask-Based Beamforming: Mask-based beamforming [21], [22] uses a DNN to estimate the ideal T-F masks, whereby the value of a T-F unit defines the relative level of the target signal within the unit. Thus the spatial covariance matrix of speech can be computed from speech-dominant T-F units, and that of noise from noise-dominant T-F units [13], [68].

In this study, we adopt the mask-based MVDR beamformer formulated in [54], where the spatial covariance matrices of speech and noise are calculated as follows:

$$\hat{\mathbf{\Phi}}_{\mathbf{s}}(f) = \frac{1}{T}\sum_t \eta(t,f)\mathbf{Y}(t,f)\mathbf{Y}(t,f)^{\mathrm{H}}, \tag{6}$$

$$\hat{\mathbf{\Phi}}_{\mathbf{v}}(f) = \frac{1}{T}\sum_t \xi(t,f)\mathbf{Y}(t,f)\mathbf{Y}(t,f)^{\mathrm{H}}, \tag{7}$$

where η and ξ represent the relative importance of each T-F unit for speech and noise covariance matrix computation, respectively. They can be computed as the median of the estimated masks in individual channels:

$$\eta(t,f) = \text{median}(\hat{M}_1(t,f), \ldots, \hat{M}_P(t,f)), \tag{8}$$

$$\xi(t,f) = \text{median}(1-\hat{M}_1(t,f), \ldots, 1-\hat{M}_P(t,f)), \tag{9}$$

where $\hat{M}_p$ is the estimated ratio mask for the p-th microphone. The RTF vector is estimated by utilizing the eigendecomposition of the speech covariance matrix [1], [10]:

$$\hat{\mathbf{r}}(f) = \mathcal{P}\{\hat{\mathbf{\Phi}}_{\mathbf{s}}(f)\},$$
$$\hat{\mathbf{d}}_q(f) = \frac{\hat{\mathbf{r}}(f)}{\hat{r}_q(f)}, \tag{10}$$

where $\mathcal{P}\{\cdot\}$ computes the principal eigenvector and $\hat{r}_q(f)$ denotes the q-th element of $\hat{\mathbf{r}}(f)$. Then the weight vector of an MVDR beamformer can be computed using (5). Note that the DNN trained for monaural mask estimation is usually also used as a post-filter that operates on the beamformer output, to further remove residual interference [53].

2) Complex Spectral Mapping Based Beamforming: Unlike mask-based beamforming that uses estimated T-F masks as a weighting mechanism, complex spectral mapping based beamforming developed in [57] directly calculates the speech and noise covariance matrices from the estimated complex spectrograms:

$$\hat{\mathbf{\Phi}}_{\mathbf{s}}(f) = \frac{1}{T}\sum_{t=1}^{T} \hat{\mathbf{S}}(t,f)\hat{\mathbf{S}}(t,f)^{\mathrm{H}}, \tag{11}$$

$$\hat{\mathbf{\Phi}}_{\mathbf{v}}(f) = \frac{1}{T}\sum_{t=1}^{T} \hat{\mathbf{V}}(t,f)\hat{\mathbf{V}}(t,f)^{\mathrm{H}}, \tag{12}$$

where $\hat{\mathbf{V}} = \mathbf{Y} - \hat{\mathbf{S}}$, and $\hat{\mathbf{S}}$ represents the estimated complex spectrogram of target speech produced by performing monaural complex spectral mapping individually for each channel. Symbol T denotes the total number of time frames in an utterance. The RTF vector is subsequently computed using (10).

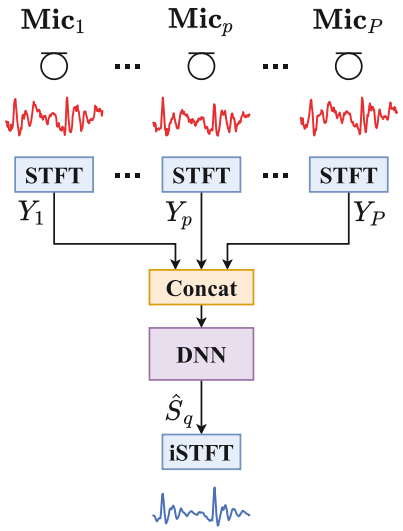


Fig. 1. Illustration of the MISO separation system based on MC-CSM, where "iSTFT" indicates the inverse STFT and "Concat" the concatenation operation.

Compared to using estimated T-F masks as an *ad hoc* weighting mechanism (see (6) and (7)), computing the spatial covariance matrices from the estimated complex spectra of speech and interference (see (11) and (12)) is a principled method, and it does not need to combine multiple monaural masks into a single one using operators (such as median) that are not easy to justify. If the complex spectra are perfectly estimated, (11) and (12) will provide the ground-truth spatial covariance matrices.

IV. Multi-Channel Complex Spectral Mapping

MC-CSM directly estimates the complex spectrogram of target speech for the reference microphone from that of the multi-channel noisy mixture via a DNN [56]. The DNN itself is used as a spectrospatial filter, yielding the output

$$\hat{S}_q = \mathcal{F}(\mathbf{Y}; \Theta), \tag{13}$$

where Θ denotes the set of all trainable parameters in the DNN. Specifically, the real and imaginary components of the mixture spectrograms at all microphones are concatenated and passed into a DNN as input. The output layer of the DNN produces an estimate of the real and imaginary components of the target spectrogram at the reference microphone. This approach essentially trains a DNN for nonlinear time-varying spectrospatial filtering, which amounts to a multiple-input single-output (MISO) system as illustrated in Fig. 1.

A. Complex-Domain T-F Representations

Akin to magnitude-domain spectra widely used in monaural separation algorithms, real and imaginary spectrograms both exhibit clear spectrotemporal structure, which is amenable to deep learning. We plot the real and imaginary spectrograms of a speech signal at two microphones of an array in Fig. 2. For a better illustration, we compress the real and imaginary components prior to plotting, using a symmetric logarithm function

$$z(x) = \begin{cases} -\log_{10}(-\alpha x + 1), & x < 0; \\ 0, & x = 0; \\ \log_{10}(\alpha x + 1), & x > 0, \end{cases} \tag{14}$$

where α is a pre-defined positive scaling factor and we set it to 100. Such a monotonically increasing function maintains the sign of the original value. As illustrated in Fig. 2(a), (d), (d), and (e), both real and imaginary spectra display spectrotemporal patterns.

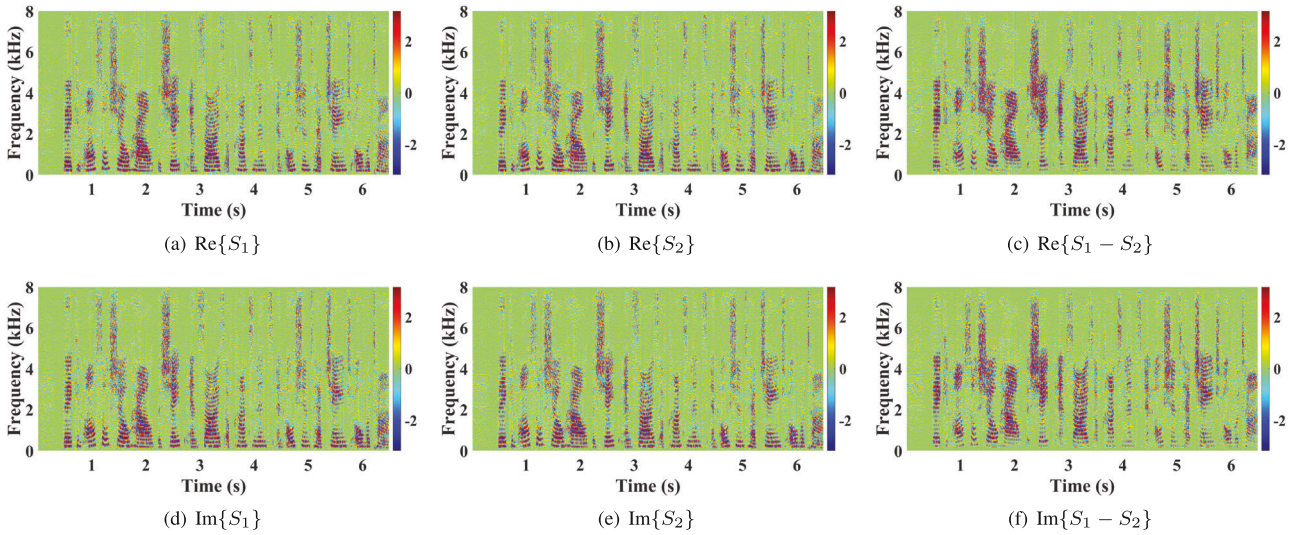
In addition, the complex spectrogram S_1 of the first-channel signal appear to be similar to S_2 of the second-channel signal, as shown in Fig. 2(a), (b), (d), and (e). To compare S_1 and S_2, we plot the real and imaginary components of $S_1 - S_2$ in Fig. 2(c) and (f), respectively, both of which also exhibit clear patterns. These patterns are strongly correlated with the inter-channel time difference (ITD) and the inter-channel intensity difference (IID) between the microphones, which are two main spatial cues in sound localization and multi-channel speech separation. Hence the difference between the complex spectrograms at two microphones would provide beneficial spatial information for speech separation.

Although the MC-CSM approach does not explicitly extract any spatial features, we believe that, with complex spectrograms from multiple channels, a DNN can learn to implicitly compare the spectra from different channels and extract effective inter-channel cues. These cues are associated with a specific TDOA between microphones. For fixed-geometry microphone arrays, this TDOA corresponds to certain DOAs, and thus the DNN would learn to suppress the interfering sounds arriving from other directions, by training on a wide range of source positions and room acoustic properties. Hence, in conjunction with spectral cues contained in the complex spectrograms of individual channels, the spatial cues implicitly encoded in inter-channel phase relations come to bear on the separation of the target speech signal.

B. DNN Architectures

In this study, we adopt two different DNN architectures for the complex spectral mapping based MISO system. The first is a densely-connected convolutional recurrent network (DC-CRN) developed in [44], and the second is a bidirectional long short-term memory (BLSTM) recurrent network.

The DC-CRN is an extension of the CRN originally designed for monaural speech enhancement in [42], which has an encoder-decoder architecture with recurrent layers between the encoder and the decoder. As depicted in Fig. 3, the encoder is a stack of convolutional densely-connected blocks (DC-blocks), and the decoder a stack of deconvolutional DC-blocks. As illustrated in Fig. 4, each DC-block is a stack of four convolutional layers and a gated convolutional or deconvolutional layer, with dense connections between layers. The output of each DC-block in the encoder is passed through a DC-block based skip pathway, and then concatenated with the features of the corresponding DC-block in the decoder. Between the encoder and the decoder, a two-layer BLSTM is used to model temporal dependencies. The



(a) Re$\{S_1\}$ (b) Re$\{S_2\}$ (c) Re$\{S_1 - S_2\}$

(d) Im$\{S_1\}$ (e) Im$\{S_2\}$ (f) Im$\{S_1 - S_2\}$

Fig. 2. Illustration of multi-channel real and imaginary spectrograms, where Re$\{\cdot\}$ and Im$\{\cdot\}$ compute the real and imaginary components, respectively. Here S_1 and S_2 denote the spectrograms of target speech at the 1st and the 2nd microphones of an array, respectively. All the spectrograms are compressed using a symmetric logarithm function prior to plotting. (a) Re$\{S_1\}$ (b) Re$\{S_2\}$ (c) Re$\{S_1 - S_2\}$ (d) Im$\{S_1\}$ (e) Im$\{S_2\}$ (f)Im$\{S_1 - S_2\}$

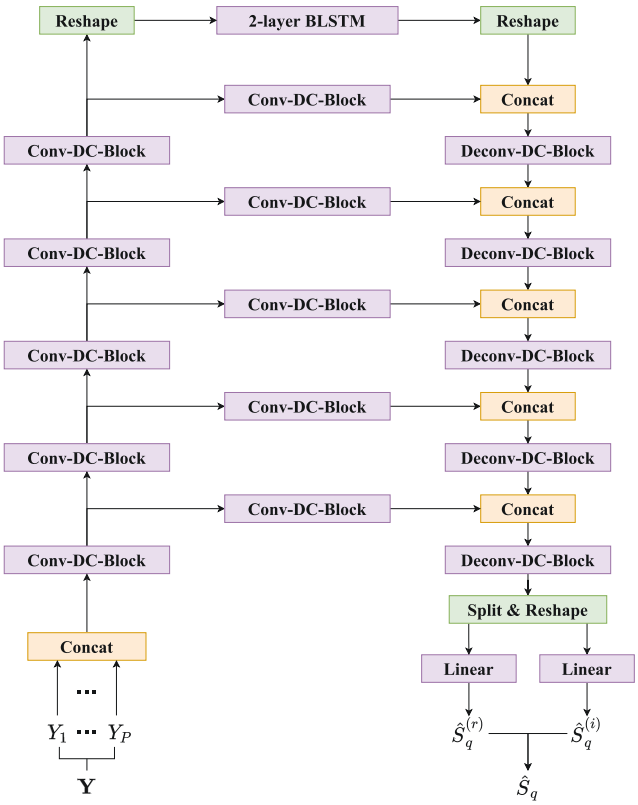
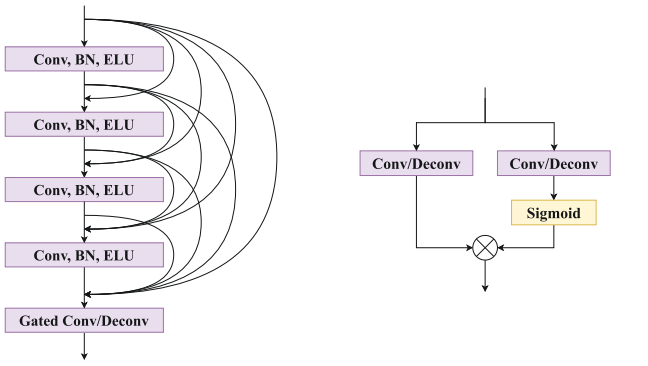


Fig. 3. Diagram of the DC-CRN for multi-channel complex spectral mapping, where $\hat{S}_q^{(r)}$ and $\hat{S}_q^{(i)}$ denote the real and imaginary components of $\hat{S}_q$, respectively.

real and imaginary components of the target spectrogram at the reference channel are estimated by two linear layers individually. We adopt the same network hyperparameters as the noncausal DC-CRN in [44].



(a) Densely-connected Block (b) Gated Convolution/Deconvolution

Fig. 4. Diagrams of the densely-connected block (a) and the gated convolution/deconvolution (b), where "BN" represents batch normalization and "ELU" the exponential linear unit. The symbol $\bigotimes$ represents the element-wise multiplication. (a) Densely-connected Block (b) Gated Convolution/Deconvolution.

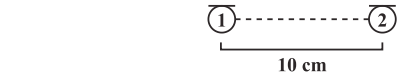
The recurrent network has four stacking BLSTM layers, each of which has 512 units for each time direction. A linear layer is employed to estimate the real and imaginary spectrograms of target speech at the reference channel. Thus, from the input layer to the output layer, this BLSTM network has $P\times2\times161$, 2×512, 2×512, 2×512, 2×512 and 2×161 units, respectively.

C. Training Objective

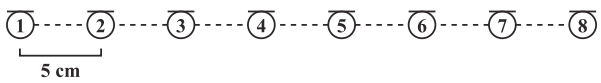
Following the complex spectral mapping approach originally developed for monaural speech enhancement [14], [43], the loss function is calculated by comparing the real and imaginary spectrograms between separated and target speech:

$$\mathcal{L}_{\text{RI}} = (\|\hat{S}_q^{(r)} - \text{Re}\{S_q\}\|_1 + \|\hat{S}_q^{(i)} - \text{Im}\{S_q\}\|_1), \qquad (15)$$

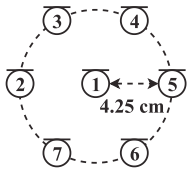
where $\|\cdot\|_1$ represents the ℓ_1 norm.



(a) 2-channel Linear Array



(b) 8-channel Linear Array



(c) 7-channel Circular Array

Fig. 5. Illustration of the three microphone arrays. For all the three arrays, microphone 1 is treated as the reference microphone. (a) 2-channel Linear Array (b)8-channel Linear Array (c) 7-channel Circular Array.

In [56], an additional magnitude loss term is introduced:

$$\mathcal{L}_{\text{Mag}} = \| |\hat{S}_q| - |S_q| \|_1, \tag{16}$$

where the magnitude spectrogram $|\hat{S}_q|$ of separated speech is calculated as $|\hat{S}_q| = \sqrt{(\hat{S}_q^{(r)})^2 + (\hat{S}_q^{(i)})^2}$. This term imposes a penalty on the magnitude estimation error, reflecting the relative importance of the magnitude over the phase in speech separation [50]. Hence we adopt the same loss function as in [56]:

$$\mathcal{L}_{\text{RI+Mag}} = \mathcal{L}_{\text{RI}} + \mathcal{L}_{\text{Mag}}. \tag{17}$$

V. Experimental Setup

A. Microphone Array Configurations

Our experiments use three different microphone arrays, as illustrated in Fig. 5. The first is a two-channel linear array, where the inter-microphone distance is set to 10 cm as in [44]. The second is an eight-channel linear array, where the distance between adjacent microphones is set to 5 cm as in [59]. The third is a circular array of seven microphones, of which the geometry is similar to the microphone array in Amazon Echo, comprising six microphones uniformly spaced on a circle plus one microphone at the center of the circle. Following [7], we set the radius of the circle to 4.25 cm. Note that all the three microphone arrays consist of omnidirectional microphones. For the two linear arrays, we treat the microphone at one endpoint as the reference microphone. For the circular array, the microphone at the center is treated as the reference microphone.

B. Data Preparation

We evaluate the MC-CSM approach on three multi-channel speech separation tasks, i.e. speech dereverberation, speech enhancement and speaker separation. All the tasks aim to estimate the direct-path (anechoic) signal at the reference microphone.

1) Speech Dereverberation: We use the training, development test and evaluation test sets of the WSJ0 dataset [17] as the speech corpus for training, validation and testing, respectively. Specifically, training, validation and test sets contain 12776 utterances from 101 speakers, 1206 utterances from 10 speakers and 651 utterances from 8 speakers, respectively.

Algorithm 1: Data Simulation Process for the Speech Dereverberation Task.

Input: WSJ0.
Output: Spatialized reverberant WSJ0.

```
REP[train] = 4; REP[validation] = 3; REP[test] = 4;
for dataset in {train, validation, test} set of WSJ0 do
  for each anechoic speech signal s in dataset do
    for count in {1, 2, ..., REP[dataset]} do
      Sample room length R_x and width R_y from [5, 10] m;
      Sample room height R_z from [3, 4] m;
      Sample array height a_z from [1, 2] m;
      Sample array displacement a_x and a_y from [−0.5, 0.5] m;
      Place array center at ⟨R_x/2 + a_x, R_y/2 + a_y, a_z⟩ m;
      Sample array orientation from [0, 2π];
      Sample target source location in the 0°-360° plane: ⟨b_x, b_y, b_z(= a_z)⟩, with the source-array distance between [1, 1.5] m;
      Sample T_60 value from [0, 1] s;
      Generate multi-channel room impulse responses using the image source method and convolve them with s to obtain the reverberant mixture;
    end for
  end for
end for
```

To generate multi-channel mixtures, we use the image source method [2] to simulate rectangular rooms. The detailed simulation procedure is provided in Algorithm 1. We assume that the microphone array is placed horizontally and has the same height as the target speech source. The reverberation time (T_{60}) is randomly sampled between 0 s and 1 s. For each of the three microphone arrays, we create training, validation and testing sets following the same procedure.

2) Speech Enhancement in the Presence of a Point-Source Noise: This speech enhancement task aims to address both denoising and dereverberation. For the scenarios where only a single point-source noise is present, we follow the same data simulation procedure in Algorithm 1, except that an additional noise source is simulated. The noise source is placed at the same height as the target source and the microphone array. We set the distance between the noise source and the array center to be the same as that between the target speech source and the array center. The DOA of the noise source is randomly sampled with the constraint that the angle between that DOA and the DOA of the target source is no smaller than 5°.

For noise source simulation, we use 18 noises from the Diverse Environments Multichannel Acoustic Noise Database

(DEMAND) [45]. Specifically, we use 15 noises for training and validation, and the remaining 3 noises for testing. For a mixture, a random cut of a randomly selected noise is used as the noise source. Note that each noise in the DEMAND dataset has 16 channels, and we only use the signal at the first channel. For the training and validation sets, the signal-to-noise ratio (SNR) is randomly sampled between -5 dB and 0 dB, where the SNR is with respect to the reverberant speech signal and the reverberant noise signal at the reference channel. For testing, we use three different SNRs, i.e. -5, 0 and 5 dB.

3) Speech Enhancement in the Presence of a Quasi-Diffuse Noise: We simulate a quasi-diffuse noise field in a way similar to [44]. Assuming that the DOA angle of the target speech source is θ, we use 72 noise source DOA angles, i.e. θ, θ+5°, θ+10°, θ+15°, ..., θ+345°, θ+350°, θ+355°. To simulate the noise sources, we first concatenate the utterances spoken by each of the 630 speakers in the TIMIT corpus [18], and then split them into 480 and 150 speakers for training and testing. Following [44], we randomly choose 72 speech clips from 72 randomly selected speakers, and place them on the 72 positions. Note that the distance between each noise source and the array center is set to be the same as that between the target speech source and the array center. We select SNRs in the same way as described in Section V-B2. Note that the sound sources are simulated in reverberant environments.

4) Speaker Separation: For the speaker separation task, we assume that the number of speakers in a mixture is 2 (see a 3-talker evaluation in Section VI-C). We consider 37 candidate azimuth positions for the speech sources, ranging from $-90°$ to 90° in 5° steps. For the linear microphone arrays, the azimuth of $-90°$ indicates the direction of looking towards microphone 1 (see Fig. 5 for the microphone numbering) from the array center, and the azimuth of 90° the opposite direction. For the circular array, the azimuth of $-90°$ is in the direction of looking towards microphone 2 from microphone 1, and the azimuth of 90° in the opposite direction. The source-array distance is set to 2 m.

We create the multi-channel mixtures by spatializing the WSJ0-2mix dataset [20], which contains 20000, 5000 and 3000 mixtures in the training, validation and test sets. Specifically, we convolve each pair of speech signals in the WSJ0-2mix dataset with a pair of multi-channel impulse responses, corresponding to two different source positions randomly selected from the 37 candidate positions. We investigate different approaches in both anechoic and reverberant conditions. In the reverberant condition, the reverberation time is randomly sampled between 0.2 s and 0.6 s.

C. STFT Settings and DNN Training Methodology

We assume that all signals are sampled at 16 kHz. We normalize the signals in the following way. Each noisy mixture at the reference microphone is rescaled by a factor such that the root mean square of that mixture waveform is 1. The same scaling factor is applied to the corresponding target speech waveform and the noisy mixtures at other microphones. In addition, we use a 20-ms Hamming window to segment the waveforms into a set of time frames, with a 50% overlap between adjacent frames. We applied a 320-point (16 kHz $\times$ 20 ms) discrete Fourier transform to each time frame, producing 161-dimensional one-sided spectra.

The models are trained on 4-second segments with the AMSGrad optimizer [37]. The minibatch size is set to 16. We use an initial learning rate of 0.001, which decays by 0.98 every two epochs. For testing, we select the model with the lowest validation loss among different epochs. Note that the models are trained separately for each experimental configuration.

D. Beamformer Baselines

1) DS Beamformer: We use an oracle DS beamformer, which is steered to the direction of the target speech source.

2) Time-Invariant MVDR Beamformer: An oracle time-invariant MVDR (TI-MVDR) beamformer can be derived by calculating the spatial covariance matrices using the ground-truth complex spectrograms:

$$\mathbf{\Phi}_{\mathbf{s}}(f) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{S}(t,f)\mathbf{S}(t,f)^{\mathrm{H}}, \tag{18}$$

$$\mathbf{\Phi}_{\mathbf{v}}(f) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{V}(t,f)\mathbf{V}(t,f)^{\mathrm{H}}, \tag{19}$$

where $\mathbf{V} = \mathbf{Y} - \mathbf{S}$. The RTF vector is estimated following (10).

In addition, a mask-based (MB) TI-MVDR beamformer is formulated as described in Section III-B1. A DNN (either DC-CRN or BLSTM) is used to monaurally estimate the ideal ratio mask (IRM):

$$M_p(t,f) = \frac{|S_p(t,f)|}{|S_p(t,f)| + |Y_p(t,f) - S_p(t,f)|}. \tag{20}$$

Similarly, we formulate a CSM based TI-MVDR as described in Section III-B2. For both MB TI-MVDR and CSM TI-MVDR, the DNN trained for mask or complex spectrum estimation can be used as a post-filter (PF).

3) Time-Varying MVDR Beamformer: Now we formulate a time-varying MVDR (TV-MVDR) beamformer. Like TI-MVDR, we estimate the RTF vector by performing eigenvalue decomposition on the speech covariance matrix computed using all time frames within an utterance. Following [27], [57], we compute the time-varying noise covariance matrix as

$$\hat{\mathbf{\Phi}}_{\mathbf{v}}(t,f) = (1-\beta)\frac{\hat{\mathbf{V}}(t,f)\hat{\mathbf{V}}(t,f)^{\mathrm{H}}}{\mathrm{trace}(\hat{\mathbf{V}}(t,f)\hat{\mathbf{V}}(t,f)^{\mathrm{H}})/P} + \beta\frac{\hat{\mathbf{\Phi}}_{\mathbf{v}}(f)}{\mathrm{trace}(\hat{\mathbf{\Phi}}_{\mathbf{v}}(f))/P}, \tag{21}$$

where β is empirically set to 0.5 and $\mathrm{trace}(\cdot)$ computes the trace of a matrix. A time-varying weight vector of the MVDR is then derived as

$$\hat{\mathbf{w}}(t,f) = \frac{\hat{\mathbf{\Phi}}_{\mathbf{v}}(t,f)^{-1}\hat{\mathbf{d}}_q(f)}{\hat{\mathbf{d}}_q(f)^{\mathrm{H}}\hat{\mathbf{\Phi}}_{\mathbf{v}}(t,f)^{-1}\hat{\mathbf{d}}_q(f)}, \tag{22}$$

and the beamformer output is computed as $\hat{S}_q(t,f) = \hat{\mathbf{w}}(t,f)^{\mathrm{H}}\mathbf{Y}(t,f)$.

TABLE I
COMPARISONS OF DIFFERENT APPROACHES ON SPEECH DEREVERBERATION

Mic. Array	linear-2ch		linear-8ch		circular-7ch	
Metric	PESQ	SI-SNR (dB)	PESQ	SI-SNR (dB)	PESQ	SI-SNR (dB)
Unprocessed	2.24	0.22	2.26	0.42	2.25	0.30
Oracle DS	2.35	1.21	2.53	2.74	2.40	1.30
Oracle TI-MVDR	2.41	4.42	2.79	10.73	2.91	13.94
MB TI-MVDR (BLSTM)	2.43	2.35	2.73	4.42	2.76	5.13
+ PF	3.19	4.37	3.28	5.59	3.35	6.27
MB TI-MVDR (DC-CRN)	2.43	2.36	2.74	4.59	2.77	5.35
+ PF	3.22	4.45	3.24	5.62	3.29	6.52
CSM TI-MVDR (BLSTM)	2.42	3.72	2.70	7.65	2.72	7.74
+ PF	3.41	8.85	3.38	10.36	3.34	9.95
CSM TI-MVDR (DC-CRN)	2.41	3.94	2.70	7.77	2.74	8.40
+ PF	**3.57**	10.47	3.54	**11.66**	3.54	11.50
Oracle TV-MVDR	2.68	7.52	3.38	15.06	3.44	15.44
CSM TV-MVDR (BLSTM)	2.60	5.15	3.11	9.09	3.12	8.74
+ PF	3.27	8.46	3.23	9.77	3.24	9.11
CSM TV-MVDR (DC-CRN)	2.61	5.55	3.11	9.29	3.14	9.47
+ PF	3.46	10.11	3.24	10.74	3.20	10.54
MC-CSM (BLSTM)	3.28	7.61	3.32	8.13	3.31	8.32
MC-CSM (DC-CRN)	3.55	**10.49**	**3.63**	11.65	**3.63**	**12.75**

With this formulation, we derive a CSM TV-MVDR beamformer. Moreover, an oracle TV-MVDR can be obtained by using the ground-truth noise spectrogram $\mathbf{V}$ in (21). The RTF vector is estimated in the same way as for the oracle TI-MVDR. Similar to the TI-MVDRs, the DNN trained for single-channel CSM can be applied to the beamformer output as a post-filter.

VI. EXPERIMENTAL RESULTS AND ANALYSES

As alluded to in Section V-B, our empirical analyses are conducted on three major separation tasks: speech dereverberation, speech enhancement, and speaker separation. In addition, we examine binaural speech enhancement.

A. Evaluation on Speech Dereverberation

We use perceptual evaluation of speech quality (PESQ) [38] and scale-invariant signal-to-noise ratio (SI-SNR) [30] to measure speech dereverberation performance. The test results are shown in Table I, where the numbers represent the averages over the test examples in each condition. The best score in each case is highlighted by boldface. We observe that the oracle DS beamformer provides limited PESQ and SI-SNR improvements over the unprocessed reverberant signals at the respective reference microphones, although the ground-truth DOA of the target source is used as the steering direction of the beamformer. In contrast, the oracle TI-MVDR produces higher PESQ and SI-SNR, and the oracle TV-MVDR further improves the dereverberation performance in both metrics. In the case of the 8-channel linear array, for example, the PESQ and SI-SNR improve from 2.53 and 2.74 dB for the oracle DS to 2.79 and 10.73 dB for the oracle TI-MVDR, and further to 3.38 and 15.06 dB for the oracle TV-MVDR.

CSM TI-MVDRs yield comparable PESQ scores to MB TI-MVDRs, and significantly higher SI-SNR than MB TI-MVDRs. Going from CSM TI-MVDRs to CSM TV-MVDRs further improves the dereverberation performance. For all these deep learning based beamformers, additional performance gains are obtained by applying a post-filter to the beamformer output, as shown in Table I. Specifically, the beamformer output is passed through the DNN trained for mask or complex spectrum estimation. For example, post-filtering produces a 0.79 PESQ improvement and a 2.09 dB SI-SNR improvement over the MB TI-MVDR with DC-CRN.

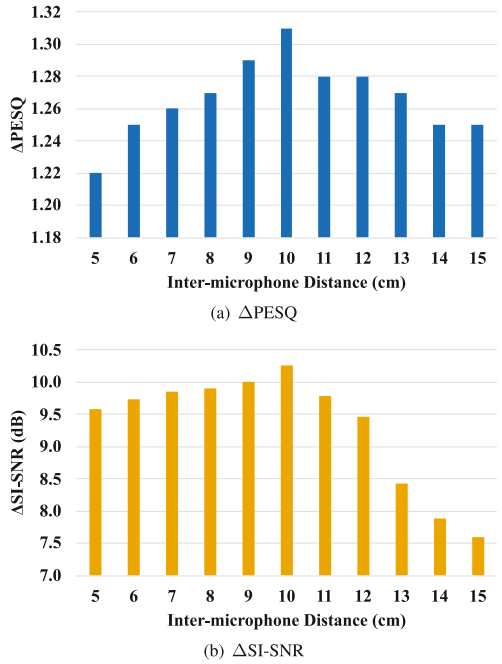
The two MC-CSM based systems with BLSTM and DC-CRN significantly elevate the PESQ and SI-SNR scores over the unprocessed signals for all the three microphone arrays. Both systems substantially outperform the oracle beamformers and the deep learning based beamformers, and they produce comparable PESQ and SI-SNR results to the strongest baselines, i.e. "CSM TI-MVDR (BLSTM) + PF" and "CSM TI-MVDR (DC-CRN) + PF". It should be noted that the beamforming based systems with a post-filter (e.g. "CSM TI-MVDR (BLSTM) + PF") apply a DNN P+1 times, P times for mask or complex spectrum estimation in the P channels and once for post-filtering. Thus these beamforming based systems have much higher computational costs than the corresponding MC-CSM systems. For example, we compare the numbers of multiply-accumulate (MAC) operations for "CSM TI-MVDR (DC-CRN) + PF" and "MC-CSM (DC-CRN)" on a 4-second reverberant signal in the 8-channel condition. Specifically, the number of MAC operations is 168.8 billion for "CSM TI-MVDR (DC-CRN) + PF" and 18.9 billion for "MC-CSM (DC-CRN)".

In addition, we investigate the MC-CSM approach on mismatched microphone array geometries using 2-channel linear arrays. Specifically, we utilize the DC-CRN trained for MC-CSM on the 2-channel linear array illustrated in Fig. 5(a), of which the inter-microphone distance is 10 cm. We test this model on 2-channel arrays with different inter-microphone distances ranging from 5 cm to 15 cm in 1-cm steps. As shown in Fig. 6, the PESQ and SI-SNR improvements relative to the unprocessed signals increase as the inter-microphone distance progressively goes from 5 cm to 10 cm, and decrease as that distance progressively goes from 10 cm to 15 cm. The greatest PESQ and SI-SNR improvements are obtained in the matched (fixed) array geometry case, not surprisingly. But the performance degrades only gradually as the distance deviates from the trained one.

When the input is monaural, multi-channel CSM naturally reduces to single-channel CSM (see Fig. 1). In this case, the DNN becomes a spectral filter. How does single-channel CSM perform relative to its multi-channel counterparts? Table II compares single- and multi-channel CSM on speech dereverberation. The numbers in the table represent the improvements

TABLE II
COMPARISON BETWEEN SINGLE- AND MULTI-CHANNEL COMPLEX SPECTRAL MAPPING ON SPEECH DEREVERBERATION. ΔPESQ AND ΔSI-SNR DENOTE PESQ AND SI-SNR IMPROVEMENTS OVER UNPROCESSED MIXTURES CAPTURED BY THE REFERENCE MICROPHONE

Mic. Setup	1ch		linear-2ch		linear-8ch		circular-7ch	
Metric	ΔPESQ	ΔSI-SNR (dB)	ΔPESQ	ΔSI-SNR (dB)	ΔPESQ	ΔSI-SNR (dB)	ΔPESQ	ΔSI-SNR (dB)
CSM (BLSTM)	0.95	6.59	1.04	7.39	1.06	7.71	1.06	8.02
CSM (DC-CRN)	1.16	7.97	1.31	10.27	1.37	11.23	1.38	12.45



(a) ΔPESQ

(b) ΔSI-SNR

Fig. 6. Investigation of the MC-CSM approach on mismatched microphone array geometries, where a DC-CRN is trained with an inter-microphone distance of 10 cm and tested at different inter-microphone distances. (a) ΔPESQ (b) ΔSI-SNR.

over unprocessed mixtures, i.e. ΔPESQ and ΔSI-SNR. In the single-channel case, CSM is performed on the reference channel of the 2-channel array. In the multi-channel cases, ΔPESQ and ΔSI-SNR are relative to the unprocessed mixtures captured by the reference channels of the corresponding microphone arrays, and these numbers are calculated from those in Table I. We observe that single-channel CSM performs well in both metrics. As expected, multi-channel CSM yields significantly better results than single-channel CSM, consistent with the previous observations for speech dereverberation [56] and speaker separation [58]. This demonstrates that complex spectral mapping clearly benefits from the spatial information encoded in the complex spectrograms of multiple microphone signals. That single-channel and multi-channel separation are handled in a unified framework is a major advantage of CSM over traditional and mask-based beamforming, which does not function for monaural inputs.

B. Evaluation on Speech Enhancement

We now compare different approaches on speech enhancement tasks. Tables III, IV and V present the short-time objective intelligibility (STOI) [41], PESQ and SI-SNR results at −5, 0 and 5 dB SNRs, respectively, in the presence of a point-source noise. We can see that the oracle DS beamformer improves the STOI, PESQ and SI-SNR over the unprocessed mixtures by a small margin. Oracle TI-MVDR and TV-MVDR yield consistently better enhancement performance in terms of all the three metrics. For MB and CSM MVDRs, similar performance trends as in Table I are observed. The CSM TV-MVDRs produce superior enhancement performance to the corresponding CSM TI-MVDRs, while additionally applying post-filters results in a different trend. Specifically, "CSM TV-MVDR (BLSTM/DC-CRN) + PF" underperforms "CSM TI-MVDR (BLSTM/DC-CRN) + PF" in terms of STOI, PESQ and SI-SNR. This is likely because TV-MVDR smears temporal dependencies across time frames in the original signal. Given that both BLSTM and DC-CRN leverage the dependencies across frames, the loss of these dependencies can be a drag for the effectiveness of the BLSTM and DC-CRN post-filters. Unlike TV-MVDRs, TI-MVDRs use the same spatial covariance matrix of noise for a whole utterance, and thus better preserve the cross-frame dependencies.

The DC-CRN is a stronger model than the BLSTM, and the DC-CRN MC-CSM system produces consistently higher STOI, PESQ and SI-SNR than the BLSTM MC-CSM system. Both these MC-CSM systems significantly outperform the oracle beamformers and the deep learning based beamformers on the 2-channel linear array. Note that all the beamformers perform better on the 8-channel linear array and the 7-channel circular array than on the 2-channel array. This is because arrays with more microphones produce a narrower beam towards a certain steering direction at a given frequency. Such beamformers provide greater attenuation of sounds arriving from other directions. Going from the 2-channel array to the 8-channel array or the 7-channel array leads to relatively smaller improvements for the MC-CSM systems than for the beamformers in the three metrics. We see that the BLSTM MC-CSM system underperforms the oracle TV-MVDR on the 8-channel and 7-channel arrays. Even on these two arrays, however, the BLSTM MC-CSM system yields similar enhancement performance to the strongest non-oracle beamformer in Tables III, IV and V, i.e. "CSM TV-MVDR (DC-CRN)".

The evaluation results for speech enhancement in the presence of a quasi-diffuse noise at −5, 0 and 5 dB SNRs are shown in Tables VI, VII and VIII, respectively. Similar performance trends as in Tables III, IV and V are observed. In addition, we find that beamforming provides clearly smaller STOI, PESQ and SI-SNR improvements over unprocessed mixtures in Tables VI, VII and VIII than in Tables III, IV and V. This confirms that the utility of beamforming is reduced in diffuse noise scenarios compared with directional noise scenarios. In addition, Tables IX and X compare single- and multi-channel CSM on speech enhancement, where ΔSTOI, ΔPESQ and ΔSI-SNR in

TABLE III
COMPARISONS OF DIFFERENT APPROACHES ON SPEECH ENHANCEMENT IN THE PRESENCE OF A POINT-SOURCE NOISE AT −5 DB SNR

Mic. Array	linear-2ch			linear-8ch			circular-7ch		
Metric	STOI (%)	PESQ	SI-SNR (dB)	STOI (%)	PESQ	SI-SNR (dB)	STOI (%)	PESQ	SI-SNR (dB)
Unprocessed	56.04	1.37	-9.14	56.40	1.38	-9.03	56.07	1.37	-9.10
Oracle DS	60.06	1.45	-8.41	67.40	1.64	-7.23	62.59	1.52	-8.36
Oracle TI-MVDR	67.42	1.61	-3.80	81.55	2.09	2.61	86.46	2.23	5.87
MB TI-MVDR (BLSTM)	63.13	1.53	-7.33	71.95	1.78	-6.63	74.20	1.87	-3.92
+ PF	78.00	2.08	-0.68	84.34	2.40	0.23	85.50	2.48	2.43
MB TI-MVDR (DC-CRN)	63.78	1.55	-7.16	73.21	1.82	-6.30	75.53	1.91	-3.71
+ PF	80.35	2.20	0.07	85.76	2.54	0.82	86.96	2.61	2.95
CSM TI-MVDR (BLSTM)	66.26	1.61	-4.24	77.47	1.99	1.00	78.89	2.00	1.34
+ PF	82.79	2.25	3.71	90.75	2.76	5.73	90.33	2.75	5.51
CSM TI-MVDR (DC-CRN)	66.81	1.61	-4.05	77.98	1.99	1.28	79.95	2.04	1.94
+ PF	**87.00**	**2.47**	**5.59**	**93.50**	**3.01**	**7.63**	**94.01**	**2.96**	**8.16**
Oracle TV-MVDR	73.81	1.79	-2.30	89.23	2.54	4.92	90.90	2.60	3.85
CSM TV-MVDR (BLSTM)	66.26	1.61	-4.24	81.36	2.28	2.39	78.89	2.00	1.34
+ PF	81.72	2.15	3.71	87.60	2.49	5.35	86.66	2.49	4.78
CSM TV-MVDR (DC-CRN)	68.72	1.68	-3.74	81.95	2.24	2.59	81.57	2.23	1.62
+ PF	84.83	2.30	5.01	90.49	2.67	7.08	90.07	2.67	6.95
MC-CSM (BLSTM)	80.37	2.08	2.73	82.55	2.27	3.17	81.42	2.09	2.17
MC-CSM (DC-CRN)	86.32	2.41	5.02	89.25	2.65	5.58	88.05	2.49	4.25

TABLE IV
COMPARISONS OF DIFFERENT APPROACHES ON SPEECH ENHANCEMENT IN THE PRESENCE OF A POINT-SOURCE NOISE AT 0 DB SNR

Mic. Array	linear-2ch			linear-8ch			circular-7ch		
Metric	STOI (%)	PESQ	SI-SNR (dB)	STOI (%)	PESQ	SI-SNR (dB)	STOI (%)	PESQ	SI-SNR (dB)
Unprocessed	64.74	1.58	-5.25	65.06	1.58	-5.11	64.89	1.58	-5.19
Oracle DS	68.70	1.68	-4.46	75.01	1.88	-3.18	70.81	1.75	-4.40
Oracle TI-MVDR	75.30	1.83	-0.45	86.79	2.28	5.54	90.92	2.42	8.66
MB TI-MVDR (BLSTM)	72.49	1.80	-3.23	81.09	2.13	-1.79	82.60	2.18	-0.20
+ PF	85.47	2.42	1.60	89.60	2.74	2.96	90.27	2.78	4.17
MB TI-MVDR (DC-CRN)	72.94	1.81	-3.07	81.56	2.15	-1.62	83.40	2.21	0.02
+ PF	87.04	2.55	2.24	90.34	2.86	3.34	91.15	2.90	4.61
CSM TI-MVDR (BLSTM)	74.42	1.83	-0.94	83.54	2.18	3.37	85.01	2.20	3.73
+ PF	88.86	2.60	5.20	93.51	3.00	6.88	92.95	2.98	6.48
CSM TI-MVDR (DC-CRN)	74.95	1.83	-0.69	83.99	2.18	3.80	85.90	2.24	4.46
+ PF	**92.35**	**2.85**	**7.46**	**95.66**	**3.23**	**9.21**	**95.92**	**3.22**	**9.50**
Oracle TV-MVDR	81.60	2.04	1.45	93.04	2.78	8.61	94.42	2.85	7.88
CSM TV-MVDR (BLSTM)	74.42	1.83	-0.94	87.39	2.51	4.77	85.01	2.20	3.73
+ PF	88.42	2.55	4.98	91.75	2.82	6.45	91.02	2.82	5.88
CSM TV-MVDR (DC-CRN)	77.71	1.95	0.01	87.87	2.49	5.30	88.19	2.50	4.80
+ PF	91.39	2.74	7.02	93.92	3.00	8.60	93.83	3.00	8.51
MC-CSM (BLSTM)	87.58	2.48	4.46	88.43	2.59	4.34	87.71	2.50	3.08
MC-CSM (DC-CRN)	91.74	2.80	7.00	93.18	2.96	7.25	92.91	2.86	6.96

TABLE V
COMPARISONS OF DIFFERENT APPROACHES ON SPEECH ENHANCEMENT IN THE PRESENCE OF A POINT-SOURCE NOISE AT 5 DB SNR

Mic. Array	linear-2ch			linear-8ch			circular-7ch		
Metric	STOI (%)	PESQ	SI-SNR (dB)	STOI (%)	PESQ	SI-SNR (dB)	STOI (%)	PESQ	SI-SNR (dB)
Unprocessed	71.81	1.79	-2.51	72.10	1.79	-2.33	71.97	1.79	-2.43
Oracle DS	75.47	1.90	-1.64	80.83	2.10	-0.21	77.11	1.96	-1.56
Oracle TI-MVDR	81.02	2.01	1.82	90.36	2.44	7.65	93.69	2.57	10.69
MB TI-MVDR (BLSTM)	78.73	2.01	-0.50	85.93	2.35	1.30	86.95	2.38	2.33
+ PF	88.87	2.62	2.70	91.73	2.91	4.09	92.22	2.93	4.91
MB TI-MVDR (DC-CRN)	79.13	2.02	-0.31	86.23	2.36	1.47	87.54	2.42	2.59
+ PF	90.06	2.75	3.27	92.29	3.02	4.44	92.85	3.05	5.26
CSM TI-MVDR (BLSTM)	80.16	2.01	1.15	87.59	2.32	4.89	88.66	2.34	5.05
+ PF	91.22	2.80	5.70	94.55	3.12	7.30	93.86	3.09	6.78
CSM TI-MVDR (DC-CRN)	80.70	2.02	1.48	88.07	2.34	5.46	89.58	2.39	6.01
+ PF	**94.42**	**3.06**	**8.31**	**96.56**	**3.34**	**9.92**	**96.70**	**3.33**	**10.10**
Oracle TV-MVDR	86.91	2.25	4.20	95.48	2.98	11.40	96.52	3.06	11.06
CSM TV-MVDR (BLSTM)	80.16	2.01	1.15	90.88	2.70	6.25	88.66	2.34	5.05
+ PF	90.97	2.77	5.48	93.42	3.00	6.87	92.59	2.99	6.23
CSM TV-MVDR (DC-CRN)	83.67	2.16	2.48	91.37	2.68	7.04	91.71	2.70	6.72
+ PF	93.91	2.99	7.92	95.38	3.19	9.31	95.33	3.18	9.20
MC-CSM (BLSTM)	90.62	2.73	5.01	90.86	2.76	4.64	90.55	2.74	4.69
MC-CSM (DC-CRN)	94.09	3.03	7.93	95.08	3.16	8.12	95.03	3.09	8.34

the multi-channel cases are calculated from the corresponding metric scores in Tables III, IV, V, VI, VII and VIII.

C. Evaluation on Speaker Separation

This section compares different approaches on the talker-independent multi-speaker separation task. We use four metrics to measure speaker separation performance, namely extended short-time objective intelligibility (ESTOI) [23], PESQ, SI-SNR and signal-to-distortion ratio (SDR) [47]. The evaluation results in the anechoic condition are shown in Table XI. Similar to the observation for the dereverberation and enhancement tasks, the oracle DS beamformer provides limited improvements in the four metrics. In contrast, the oracle TI-MVDR and TV-MVDR perform very well, different from the observation for the dereveberation and enhancement tasks. This is likely because the speaker separation task assumes an anechoic environment, in which adaptive beamforming would be more effective than in reverberant environments.

TABLE VI
COMPARISONS OF DIFFERENT APPROACHES ON SPEECH ENHANCEMENT IN THE PRESENCE OF A QUASI-DIFFUSE NOISE AT −5 DB SNR

Mic. Array	linear-2ch			linear-8ch			circular-7ch		
Metric	STOI (%)	PESQ	SI-SNR (dB)	STOI (%)	PESQ	SI-SNR (dB)	STOI (%)	PESQ	SI-SNR (dB)
Unprocessed	55.03	1.39	-9.08	55.33	1.40	-8.97	55.11	1.40	-9.03
Oracle DS	58.41	1.48	-7.78	64.93	1.58	-5.60	61.16	1.54	-7.84
Oracle TI-MVDR	62.19	1.47	-5.96	73.15	1.69	-1.47	76.00	1.70	-0.07
MB TI-MVDR (BLSTM)	59.69	1.49	-7.54	68.16	1.63	-4.78	68.60	1.66	-4.13
+ PF	74.77	1.91	-1.69	80.74	2.11	-0.27	82.38	2.21	1.43
MB TI-MVDR (DC-CRN)	60.13	1.49	-7.33	68.66	1.64	-4.49	68.97	1.64	-3.59
+ PF	77.27	2.00	-0.86	82.60	2.22	0.28	83.94	2.30	1.93
CSM TI-MVDR (BLSTM)	61.03	1.48	-6.44	69.69	1.53	-2.65	70.45	1.55	-2.50
+ PF	78.44	2.06	2.33	84.85	2.24	3.68	85.70	2.29	3.98
CSM TI-MVDR (DC-CRN)	61.50	1.48	-6.29	70.20	1.53	-2.47	71.13	1.57	-2.24
+ PF	82.44	**2.23**	3.59	**87.97**	2.43	4.64	**89.62**	**2.53**	5.52
Oracle TV-MVDR	69.24	1.60	-2.55	87.81	2.18	5.09	89.34	2.22	5.76
CSM TV-MVDR (BLSTM)	62.80	1.53	-4.56	77.69	1.99	1.32	76.28	1.95	1.04
+ PF	75.43	1.90	1.67	83.64	2.25	3.96	82.23	2.15	3.52
CSM TV-MVDR (DC-CRN)	63.82	1.53	-4.38	78.34	1.96	1.60	77.42	1.95	1.39
+ PF	80.38	2.14	3.12	86.60	2.33	5.54	85.77	2.22	5.21
MC-CSM (BLSTM)	78.54	2.05	2.37	81.91	2.20	2.93	81.19	2.27	4.11
MC-CSM (DC-CRN)	**83.01**	2.18	**3.92**	87.62	**2.46**	**5.58**	87.57	2.42	**6.47**

TABLE VII
COMPARISONS OF DIFFERENT APPROACHES ON SPEECH ENHANCEMENT IN THE PRESENCE OF A QUASI-DIFFUSE NOISE AT 0 DB SNR

Mic. Array	linear-2ch			linear-8ch			circular-7ch		
Metric	STOI (%)	PESQ	SI-SNR (dB)	STOI (%)	PESQ	SI-SNR (dB)	STOI (%)	PESQ	SI-SNR (dB)
Unprocessed	64.47	1.53	-5.16	64.86	1.54	-5.03	64.55	1.53	-5.10
Oracle DS	68.20	1.63	-3.93	74.42	1.77	-1.86	70.45	1.68	-3.94
Oracle TI-MVDR	72.31	1.65	-1.98	82.58	1.92	2.63	85.42	1.95	4.23
MB TI-MVDR (BLSTM)	70.12	1.66	-3.29	78.35	1.88	-0.55	79.21	1.91	0.22
+ PF	83.79	2.29	1.05	87.47	2.48	2.48	88.83	2.57	3.79
MB TI-MVDR (DC-CRN)	70.51	1.66	-3.12	78.69	1.89	-0.36	79.29	1.88	0.53
+ PF	85.74	2.40	1.60	88.73	2.60	2.90	89.77	2.69	4.25
CSM TI-MVDR (BLSTM)	71.39	1.65	-2.33	79.72	1.82	1.15	80.68	1.83	1.47
+ PF	87.02	2.51	4.55	91.68	2.77	6.19	91.64	2.77	6.09
CSM TI-MVDR (DC-CRN)	71.83	1.65	-2.19	80.09	1.81	1.33	81.33	1.85	1.79
+ PF	90.51	**2.72**	6.29	**94.08**	**2.98**	8.09	**94.36**	**3.01**	8.16
Oracle TV-MVDR	79.27	1.83	1.40	93.06	2.46	**8.95**	94.23	2.52	**9.57**
CSM TV-MVDR (BLSTM)	74.18	1.76	-0.42	86.18	2.27	4.38	85.63	2.26	4.21
+ PF	85.20	2.40	4.04	89.04	2.55	5.47	88.27	2.50	5.12
CSM TV-MVDR (DC-CRN)	75.18	1.77	-0.16	86.73	2.24	4.81	86.50	2.25	4.77
+ PF	89.37	2.64	5.79	91.75	2.71	7.61	90.66	2.53	6.89
MC-CSM (BLSTM)	86.68	2.48	4.14	88.64	2.58	4.35	88.96	2.64	6.39
MC-CSM (DC-CRN)	**90.73**	2.70	**6.43**	93.27	2.92	7.85	93.57	2.90	8.89

TABLE VIII
COMPARISONS OF DIFFERENT APPROACHES ON SPEECH ENHANCEMENT IN THE PRESENCE OF A QUASI-DIFFUSE NOISE AT 5 DB SNR

Mic. Array	linear-2ch			linear-8ch			circular-7ch		
Metric	STOI (%)	PESQ	SI-SNR (dB)	STOI (%)	PESQ	SI-SNR (dB)	STOI (%)	PESQ	SI-SNR (dB)
Unprocessed	71.92	1.71	-2.39	72.35	1.72	-2.23	72.00	1.71	-2.32
Oracle DS	75.54	1.82	-1.24	81.02	1.98	0.66	77.23	1.86	-1.21
Oracle TI-MVDR	79.70	1.85	0.99	88.64	2.15	5.91	91.24	2.21	7.84
MB TI-MVDR (BLSTM)	77.46	1.86	-0.52	84.52	2.11	1.87	85.22	2.13	2.65
+ PF	87.36	2.52	2.30	89.98	2.69	3.60	90.85	2.75	4.59
MB TI-MVDR (DC-CRN)	77.84	1.87	-0.35	84.80	2.12	2.05	85.52	2.10	3.10
+ PF	88.97	2.62	2.75	90.71	2.80	3.90	91.52	2.86	5.02
CSM TI-MVDR (BLSTM)	78.72	1.85	0.48	85.96	2.07	3.75	86.63	2.08	3.91
+ PF	89.40	2.66	5.04	93.02	2.90	6.59	92.75	2.89	6.43
CSM TI-MVDR (DC-CRN)	79.23	1.86	0.69	86.39	2.07	4.09	87.45	2.10	4.49
+ PF	92.98	2.93	7.24	95.51	**3.17**	9.22	95.53	**3.17**	9.01
Oracle TV-MVDR	85.96	2.07	4.31	**95.89**	2.73	**11.85**	**96.74**	2.79	**12.35**
CSM TV-MVDR (BLSTM)	81.49	2.00	2.06	90.46	2.50	6.13	89.98	2.49	5.78
+ PF	87.85	2.55	4.51	90.34	2.61	5.72	90.02	2.61	5.56
CSM TV-MVDR (DC-CRN)	82.47	2.01	2.48	90.99	2.47	6.78	90.88	2.49	6.62
+ PF	91.90	2.81	6.65	93.33	2.81	8.30	91.55	2.53	6.96
MC-CSM (BLSTM)	89.21	2.65	4.56	90.76	2.74	4.65	91.00	2.83	6.91
MC-CSM (DC-CRN)	**93.37**	**2.95**	**7.40**	95.19	3.14	8.84	95.52	3.13	9.93

For the deep learning based beamformers, we use the utterance-level permutation invariant training (uPIT) criterion for DNN training to achieve talker independency [25]. We always align the speaker permutations across channels prior to computing the candidate training losses corresponding to different permutations. Unlike the observations for speech dereverberation and enhancement, the TV-MVDRs underperform the corresponding TI-MVDRs. A possible interpretation is that the spatial covariance matrix of noise is less time-varying in anechoic environments than in reverberant environments, and (12) can provide a more accurate estimate of the noise covariance matrix than (21). Note that, for the deep

TABLE IX
COMPARISON BETWEEN SINGLE- AND MULTI-CHANNEL COMPLEX SPECTRAL MAPPING ON SPEECH ENHANCEMENT IN THE PRESENCE OF A POINT-SOURCE NOISE. ΔSTOI DENOTES STOI IMPROVEMENT OVER UNPROCESSED MIXTURES CAPTURED BY THE REFERENCE MICROPHONE

-5 dB SNR												
Mic. Setup	1ch			linear-2ch			linear-8ch			circular-7ch		
Metric	ΔSTOI (%)	ΔPESQ	ΔSI-SNR (dB)	ΔSTOI (%)	ΔPESQ	ΔSI-SNR (dB)	ΔSTOI (%)	ΔPESQ	ΔSI-SNR (dB)	ΔSTOI (%)	ΔPESQ	ΔSI-SNR (dB)
CSM (BLSTM)	17.33	0.38	9.17	24.33	0.71	11.87	26.15	0.89	12.20	25.35	0.72	11.27
CSM (DC-CRN)	23.29	0.63	11.47	30.28	1.04	14.16	32.85	1.27	14.61	31.98	1.12	13.35
0 dB SNR												
Mic. Setup	1ch			linear-2ch			linear-8ch			circular-7ch		
Metric	ΔSTOI (%)	ΔPESQ	ΔSI-SNR (dB)	ΔSTOI (%)	ΔPESQ	ΔSI-SNR (dB)	ΔSTOI (%)	ΔPESQ	ΔSI-SNR (dB)	ΔSTOI (%)	ΔPESQ	ΔSI-SNR (dB)
CSM (BLSTM)	18.46	0.61	7.54	22.84	0.90	9.71	23.37	1.01	9.45	22.82	0.92	8.27
CSM (DC-CRN)	23.24	0.89	10.17	27.00	1.22	12.25	28.12	1.38	12.36	28.02	1.28	12.15
5 dB SNR												
Mic. Setup	1ch			linear-2ch			linear-8ch			circular-7ch		
Metric	ΔSTOI (%)	ΔPESQ	ΔSI-SNR (dB)	ΔSTOI (%)	ΔPESQ	ΔSI-SNR (dB)	ΔSTOI (%)	ΔPESQ	ΔSI-SNR (dB)	ΔSTOI (%)	ΔPESQ	ΔSI-SNR (dB)
CSM (BLSTM)	15.84	0.68	5.63	18.81	0.94	7.52	18.76	0.97	6.97	18.58	0.95	7.12
CSM (DC-CRN)	19.90	0.97	8.62	22.28	1.24	10.44	22.98	1.37	10.45	23.06	1.30	10.77

TABLE X
COMPARISON BETWEEN SINGLE- AND MULTI-CHANNEL COMPLEX SPECTRAL MAPPING ON SPEECH ENHANCEMENT IN THE PRESENCE OF A QUASI-DIFFUSE NOISE

-5 dB SNR												
Mic. Setup	1ch			linear-2ch			linear-8ch			circular-7ch		
Metric	ΔSTOI (%)	ΔPESQ	ΔSI-SNR (dB)	ΔSTOI (%)	ΔPESQ	ΔSI-SNR (dB)	ΔSTOI (%)	ΔPESQ	ΔSI-SNR (dB)	ΔSTOI (%)	ΔPESQ	ΔSI-SNR (dB)
CSM (BLSTM)	17.58	0.38	8.99	23.51	0.66	11.45	26.58	0.80	11.90	26.08	0.87	13.14
CSM (DC-CRN)	22.42	0.56	10.86	27.98	0.79	13.00	32.29	1.06	14.55	32.46	1.02	15.50
0 dB SNR												
Mic. Setup	1ch			linear-2ch			linear-8ch			circular-7ch		
Metric	ΔSTOI (%)	ΔPESQ	ΔSI-SNR (dB)	ΔSTOI (%)	ΔPESQ	ΔSI-SNR (dB)	ΔSTOI (%)	ΔPESQ	ΔSI-SNR (dB)	ΔSTOI (%)	ΔPESQ	ΔSI-SNR (dB)
CSM (BLSTM)	18.68	0.73	7.48	22.21	0.95	9.30	23.78	1.04	9.38	24.41	1.11	11.49
CSM (DC-CRN)	22.91	0.97	9.57	26.26	1.17	11.59	28.41	1.38	12.88	29.02	1.37	13.99
5 dB SNR												
Mic. Setup	1ch			linear-2ch			linear-8ch			circular-7ch		
Metric	ΔSTOI (%)	ΔPESQ	ΔSI-SNR (dB)	ΔSTOI (%)	ΔPESQ	ΔSI-SNR (dB)	ΔSTOI (%)	ΔPESQ	ΔSI-SNR (dB)	ΔSTOI (%)	ΔPESQ	ΔSI-SNR (dB)
CSM (BLSTM)	14.62	0.75	5.34	17.29	0.94	6.95	18.41	1.02	6.88	19.00	1.12	9.23
CSM (DC-CRN)	18.79	1.04	7.83	21.45	1.24	9.79	22.84	1.42	11.07	23.52	1.42	12.25

TABLE XI
COMPARISONS OF DIFFERENT APPROACHES ON TWO-TALKER SPEAKER SEPARATION IN ANECHOIC CONDITIONS

Mic. Array	linear-2ch				linear-8ch				circular-7ch				Training Criterion
Metric	ESTOI (%)	PESQ	SI-SNR (dB)	SDR (dB)	ESTOI (%)	PESQ	SI-SNR (dB)	SDR (dB)	ESTOI (%)	PESQ	SI-SNR (dB)	SDR (dB)	
Unprocessed	56.84	1.89	0.00	0.08	56.85	1.89	0.00	0.08	56.85	1.89	0.00	0.08	-
Oracle DS	61.18	2.05	1.50	1.57	68.97	2.38	4.46	4.61	62.06	2.05	1.13	1.19	-
Oracle TI-MVDR	94.88	3.70	22.41	22.91	95.89	3.88	24.47	25.65	**99.62**	**4.38**	**29.53**	**31.10**	-
MB TI-MVDR (BLSTM)	84.02	2.96	10.43	11.93	82.61	2.87	8.52	10.59	86.20	3.03	11.28	12.82	uPIT
MB TI-MVDR (DC-CRN)	84.44	3.01	11.05	12.54	85.66	3.05	10.33	12.49	88.80	3.21	13.02	14.70	uPIT
CSM TI-MVDR (BLSTM)	80.50	2.79	7.89	9.95	81.46	3.32	7.46	12.05	82.34	3.30	7.28	11.68	uPIT
CSM TI-MVDR (DC-CRN)	84.89	3.05	13.41	14.02	86.81	3.21	13.81	15.54	84.39	3.07	12.82	13.84	uPIT
Oracle TV-MVDR	87.96	3.24	16.30	16.41	95.93	**3.95**	**25.34**	**26.38**	92.39	3.62	21.93	22.18	-
CSM TV-MVDR (BLSTM)	76.86	2.71	6.98	8.76	83.27	3.29	7.28	11.58	78.27	3.07	6.29	9.90	uPIT
CSM TV-MVDR (DC-CRN)	79.12	2.79	9.88	10.14	85.33	3.20	13.04	14.39	79.76	2.89	10.49	11.09	uPIT
MC-CSM (DC-CRN)	**97.32**	**3.91**	**23.27**	**23.58**	**96.92**	3.87	23.54	23.85	98.28	4.02	26.36	26.70	uPIT
MC-CSM (DC-CRN)	97.02	3.88	23.04	23.41	96.19	3.86	23.31	23.70	98.45	4.05	26.00	26.33	LBT
MC-CSM (BLSTM)	86.72	3.15	8.12	11.57	87.17	3.19	6.92	10.89	87.60	3.17	7.63	11.43	LBT

learning based beamformers, we do not use the single-channel DNN trained for mask or complex spectrum estimation as a post-filter. This is because the beamformer output is a separated speech signal with residual interference, rather than a 2-speaker mixture. Thus the DNN trained on monaural 2-speaker mixtures would be an inappropriate post-filter.

For the MC-CSM systems, we train the DNN using a new criterion to achieve talker independency, which assigns the speakers to the two output layers based on speaker locations. As illustrated in Fig. 7, we always assign the speaker located at a smaller azimuth (i.e. speaker 1) to the first output layer and the other speaker (i.e. speaker 2) to the second output layer. We refer to this criterion as location-based training (LBT). This criterion would work because the label permutation is selected consistently based on the relative positions of the two speakers. Given the spatial cues encoded in the multi-channel input spectrograms, such a way of assigning labels could discriminatively guide DNN training for MC-CSM. For comparison, we additionally train a DC-CRN for MC-CSM using uPIT. As shown in Table XI, the DC-CRN MC-CSM system with location-based training produces

TABLE XII
COMPARISONS OF DIFFERENT APPROACHES ON TWO-TALKER SPEAKER SEPARATION IN REVERBERANT CONDITIONS

Mic. Array	linear-2ch				linear-8ch				circular-7ch				Training Criterion
Metric	ESTOI (%)	PESQ	SI-SNR (dB)	SDR (dB)	ESTOI (%)	PESQ	SI-SNR (dB)	SDR (dB)	ESTOI (%)	PESQ	SI-SNR (dB)	SDR (dB)	
Unprocessed	37.45	1.58	-6.73	-1.78	37.45	1.59	-6.74	-1.77	37.65	1.59	-6.68	-1.74	-
Oracle DS	40.99	1.69	-5.51	-0.88	48.02	1.89	-3.59	0.85	43.45	1.74	-5.53	-0.95	-
Oracle TI-MVDR	47.63	1.81	-2.60	1.40	63.88	2.23	3.71	6.27	68.80	2.31	5.96	8.06	-
MB TI-MVDR (BLSTM)	37.77	1.61	-7.19	-1.77	37.94	1.59	-8.71	-1.78	38.31	1.62	-6.81	-1.71	uPIT
MB TI-MVDR (DC-CRN)	37.80	1.60	-7.18	-1.76	37.87	1.58	-8.76	-1.81	38.35	1.62	-6.82	-1.70	uPIT
CSM TI-MVDR (BLSTM)	40.53	1.74	-5.42	0.24	40.77	1.89	-2.87	2.71	41.91	1.89	-3.49	2.28	uPIT
CSM TI-MVDR (DC-CRN)	45.20	1.79	-4.06	1.00	55.57	2.05	-0.89	4.14	52.78	1.99	-1.85	3.47	uPIT
Oracle TV-MVDR	56.80	2.02	0.46	3.76	**79.23**	**2.75**	**8.31**	**10.86**	**79.71**	2.75	**7.95**	**10.72**	-
CSM TV-MVDR (BLSTM)	40.70	1.76	-4.68	0.94	54.91	2.25	-1.44	4.28	50.00	2.16	-2.14	3.55	uPIT
CSM TV-MVDR (DC-CRN)	46.74	1.84	-3.48	1.61	59.59	2.26	-0.22	5.02	54.00	2.10	-1.82	3.51	uPIT
MC-CSM (DC-CRN)	71.67	2.53	1.64	5.69	75.76	2.70	3.39	7.59	78.23	2.76	5.76	9.32	uPIT
MC-CSM (DC-CRN)	**72.32**	**2.56**	**2.03**	**5.97**	75.10	2.67	3.36	7.58	78.71	**2.78**	5.87	9.44	LBT
MC-CSM (BLSTM)	46.99	1.87	-3.14	1.40	58.22	2.16	-1.19	3.06	60.65	2.24	0.25	3.96	LBT

TABLE XIII
EVALUATION OF MC-CSM SYSTEMS ON THREE-TALKER SPEAKER SEPARATION

Anechoic Environment													
Mic. Array	linear-2ch				linear-8ch				circular-7ch				Training Criterion
Metric	ESTOI (%)	PESQ	SI-SNR (dB)	SDR (dB)	ESTOI (%)	PESQ	SI-SNR (dB)	SDR (dB)	ESTOI (%)	PESQ	SI-SNR (dB)	SDR (dB)	
Unprocessed	39.21	1.51	-3.16	-3.04	39.22	1.51	-3.18	-3.05	39.22	1.52	-3.16	-3.04	-
MC-CSM (DC-CRN)	82.05	2.83	11.95	12.32	85.75	3.08	13.39	13.77	91.87	3.41	18.49	18.82	uPIT
MC-CSM (DC-CRN)	**83.81**	**2.92**	**12.55**	**12.92**	**86.45**	**3.13**	**13.70**	**14.05**	**92.49**	**3.45**	**19.19**	**19.56**	LBT
MC-CSM (BLSTM)	65.77	2.44	4.09	7.36	75.95	2.76	5.17	9.09	76.19	2.71	5.36	9.05	LBT
Reverberant Environment													
Mic. Array	linear-2ch				linear-8ch				circular-7ch				Training Criterion
Metric	ESTOI (%)	PESQ	SI-SNR (dB)	SDR (dB)	ESTOI (%)	PESQ	SI-SNR (dB)	SDR (dB)	ESTOI (%)	PESQ	SI-SNR (dB)	SDR (dB)	
Unprocessed	26.62	1.35	-8.79	-4.41	26.62	1.35	-8.81	-4.43	26.74	1.35	-8.75	-4.39	-
MC-CSM (DC-CRN)	**55.72**	**1.96**	**-0.72**	**3.57**	**61.71**	**2.16**	**1.15**	5.00	**67.71**	**2.39**	**3.84**	**7.10**	uPIT
MC-CSM (DC-CRN)	55.41	1.95	-0.84	3.49	61.30	**2.16**	1.10	**5.07**	66.73	2.31	3.81	7.03	LBT
MC-CSM (BLSTM)	34.75	1.59	-4.95	-0.27	41.60	1.75	-3.56	1.07	44.40	1.79	-2.20	1.55	LBT

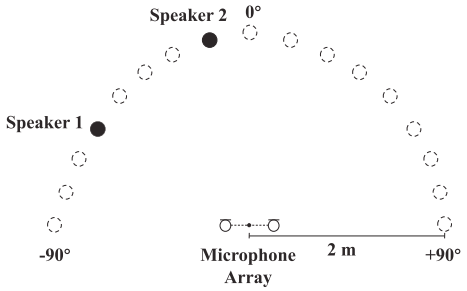


Fig. 7. Illustration of the new training criterion for the MC-CSM models. The solid circles indicate the speaker source locations, and the dashed circles the other candidate source locations in our data simulation.

similar results to the DC-CRN MC-CSM system with uPIT in all the four metrics. Moreover, these DC-CRN MC-CSM systems yield comparable results to the oracle TI-MVDR, and the BLSTM MC-CSM system produces comparable results to the strongest non-oracle beamformer baseline, i.e. "CSM TI-MVDR (DC-CRN)".

In the reverberant condition (see Table XII), all the scores decrease substantially compared to the anechoic condition. We see that the non-oracle beamformers provide limited improvements over the unprocessed mixtures. In addition, the BLSTM MC-CSM system produces comparable performance to the best-performing non-oracle beamformer, i.e. "CSM TV-MVDR (DC-CRN)". The DC-CRN MC-CSM system performs significantly better than the BLSTM MC-CSM system, suggesting that the design of the DNN architecture can greatly impact the performance of an MC-CSM system.

We additionally evaluate the MC-CSM systems on 3-talker speaker separation. The training, validation and test sets are created by spatializing the WSJ0-3mix dataset. As shown in Table XIII, the MC-CSM systems are capable of separating three concurrent speakers in both anechoic and reverberant environments, which further demonstrates the general utility of MC-CSM on speech separation.

For MC-CSM, the two training criteria, PIT and LBT, are based on different principles. The widely-used PIT criterion makes assignments by comparing the losses corresponding to all possible label permutations, while LBT assigns speakers consistently based on the source locations, enforcing the DNN to leverage spatial information to address the label ambiguity. As shown in Tables XI, XII and XIII, the two different criteria yield almost the same separation performance.

In addition, Table XIV compares single- and multi-channel CSM on speaker separation. We use DC-CRN based single- and multi-channel CSM systems trained with uPIT for this comparison. In the multi-channel cases, ΔESTOI, ΔPESQ, ΔSI-SNR and ΔSDR are calculated from the metric scores in Tables XI, XII and XIII. Note that LBT is inapplicable to single-channel CSM training due to the lack of spatial information. As expected, multi-channel CSM substantially outperforms single-channel CSM in all conditions.

TABLE XIV
COMPARISON BETWEEN SINGLE- AND MULTI-CHANNEL COMPLEX SPECTRAL MAPPING ON SPEAKER SEPARATION. ΔESTOI AND ΔSDR DENOTE ESTOI AND SDR IMPROVEMENTS OVER UNPROCESSED MIXTURES CAPTURED BY THE REFERENCE MICROPHONE

2-talker								
	Anechoic Environment				Reverberant Environment			
Metric	ΔESTOI (%)	ΔPESQ	ΔSI-SNR (dB)	ΔSDR (dB)	ΔESTOI (%)	ΔPESQ	ΔSI-SNR (dB)	ΔSDR (dB)
1ch	25.91	0.87	10.23	10.55	21.48	0.42	5.30	4.18
linear-2ch	40.48	2.02	23.27	23.50	34.22	0.95	8.37	7.47
linear-8ch	40.07	1.98	23.54	23.77	38.31	1.11	10.13	9.36
circular-7ch	41.43	2.13	26.36	26.62	40.58	1.17	12.44	11.06
3-talker								
	Anechoic Environment				Reverberant Environment			
Metric	ΔESTOI (%)	ΔPESQ	ΔSI-SNR (dB)	ΔSDR (dB)	ΔESTOI (%)	ΔPESQ	ΔSI-SNR (dB)	ΔSDR (dB)
1ch	20.65	0.33	6.55	7.07	7.95	0.03	2.57	2.85
linear-2ch	42.84	1.32	15.11	15.36	29.10	0.61	8.07	7.98
linear-8ch	46.53	1.57	16.57	16.82	35.09	0.81	9.96	9.43
circular-7ch	52.65	1.89	21.65	21.86	40.97	1.04	12.59	11.49

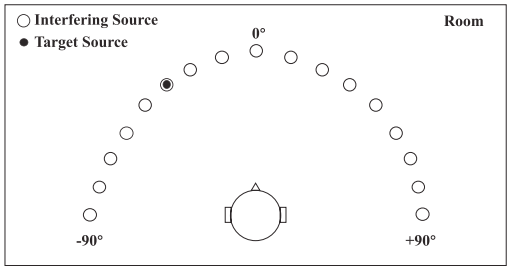


Fig. 8. Illustration of the data simulation process for binaural speech enhancement.

D. Investigation on Binaural Speech Enhancement

Finally, we evaluate different approaches on binaural speech enhancement, which is another classical multi-channel enhancement problem [51]. We use WSJ0 utterances as speech sources. In order to simulate binaural room impulse responses (BRIRs), we utilize the RAZR room acoustics simulator[1] [60], in which the spherical-head model of Brown and Duda [4] is used to simulate the head-related transfer function (HRTF). Specifically, we randomly sample the room size between $4.5\times4.5\times2.2$ m^3 and $10\times10\times4.5$ m^3. The reverberation time is randomly sampled between 0.2 s and 1.2 s, and the radius of the sphere (head) between 8 cm and 11 cm. As illustrated in Fig. 8, we use 37 sound source positions located on a semicircle in the front of the head, ranging from -90° to 90° in 5° steps. Similar to the procedure described in Section V-B3, TIMIT speakers are used as the interfering sources placed at the 37 positions [67]. The target source is randomly placed at one of the 37 positions. The source-array distance is set to 1 m. We randomly sample the height of the head between 1 m and 2 m, and assume that all the sound sources are located at the same height.

Apart from the test set created from simulated BRIRs, we create another test set using a set of real BRIRs recorded at the University of Surrey.[2] These BRIRs were captured using a Cortex Instruments Mk2 head and torso simulator (HATS). Specifically, the impulse responses were obtained by replaying sinesweeps through a loudspeaker and then deconvolving the responses. The loudspeaker was placed on a semicircle around the HATS, which has a radius of 1.5 m. The azimuth positions of the loudspeaker range from -90° to 90° in 5° intervals. Four rooms with different sizes and reflective characteristics were used for recording, corresponding to different reverberation times, i.e. 0.32 s, 0.47 s, 0.68 s and 0.89 s.

We treat the left ear as the reference channel. Tables XV and XVI show the evaluation results on the simulated BRIRs and the real BRIRs, respectively. In both tables, similar performance trends are observed for the beamformers. Note that we do not use the oracle DS beamforming in the real BRIRs case, because the geometric information of the head is not publicly available. Moreover, we see that the MC-CSM systems yield superior enhancement performance to the beamformers. The BLSTM and DC-CRN MC-CSM systems produce comparable STOI, PESQ and SI-SNR to "CSM TV-MVDR (BLSTM) + PF" and "CSM TV-MVDR (DC-CRN) + PF," respectively. Moreover, single-channel CSM, denoted as "SC-CSM" in Tables XV and XVI, performs well on both simulated and real BRIRs, although it significantly underperforms multi-channel CSM unsurprisingly.

In addition, by comparing the results between Tables XV and XVI, we find that the MC-CSM models trained on simulated BRIRs generalize reasonably well to real BRIRs, and exhibit comparable decreases in the results (from Tables XV and XVI) to the "CSM TI/TV-MVDR + PF" systems. Given the lack of a large, publicly available real BRIR set, this suggests the effectiveness of the BRIR simulation method in generating arbitrary numbers of BRIRs for DNN training, although the spherical-head model only provides a crude approximation of the HRTF.

VII. CONCLUDING REMARKS

In this study, we have comprehensively examined the multi-channel complex spectral mapping approach as a neural spectrospatial filter. Although the approach does not explicitly utilize any spatial features, spectral and spatial cues are available in the complex spectrograms of microphone signals, and exploited through complex spectral mapping. In this approach, a trained DNN itself is a spectrospatial filter. Such an approach is conceptually simpler, computationally more efficient and easier to adapt to real-time processing, than deep learning based beamforming. For instance, an MC-CSM separation system requires

[1][Online]. Available: http://medi.uni-oldenburg.de/razr/
[2][Online]. Available: https://github.com/IoSR-Surrey/RealRoomBRIRs

TABLE XV
EVALUATION OF DIFFERENT APPROACHES FOR BINAURAL SPEECH ENHANCEMENT ON SIMULATED BRIRS

SNR	-5 dB			0 dB			5 dB		
Metric	STOI (%)	PESQ	SI-SNR (dB)	STOI (%)	PESQ	SI-SNR (dB)	STOI (%)	PESQ	SI-SNR (dB)
Unprocessed	54.07	1.44	-9.51	63.82	1.55	-5.65	71.64	1.74	-3.10
Oracle DS	58.86	1.46	-7.79	68.95	1.64	-4.27	76.58	1.85	-2.12
Oracle TI-MVDR	62.48	1.52	-6.09	72.64	1.70	-2.23	80.05	1.91	0.52
MB TI-MVDR (BLSTM)	60.48	1.50	-7.61	70.95	1.70	-3.67	78.28	1.92	-1.43
+ PF	75.22	1.90	-2.35	84.53	2.31	0.13	88.03	2.54	1.07
MB TI-MVDR (DC-CRN)	60.63	1.51	-7.50	71.08	1.70	-3.59	78.43	1.92	-1.35
+ PF	75.95	1.95	-2.00	85.09	2.38	0.41	88.44	2.62	1.32
CSM TI-MVDR (BLSTM)	60.91	1.57	-6.61	71.84	1.73	-2.58	79.36	1.93	0.05
+ PF	75.89	1.96	1.01	85.48	2.46	3.57	88.59	2.69	4.51
CSM TI-MVDR (DC-CRN)	61.44	1.56	-6.48	72.09	1.72	-2.53	79.61	1.93	0.12
+ PF	79.77	2.14	2.02	89.27	2.67	4.74	92.32	2.92	5.88
Oracle TV-MVDR	70.30	1.65	-2.61	80.22	1.88	1.23	86.79	2.12	3.96
CSM TV-MVDR (BLSTM)	63.67	1.62	-4.61	75.69	1.84	-0.44	83.11	2.08	1.97
+ PF	77.53	2.00	1.44	85.98	2.48	3.68	88.56	2.67	4.32
CSM TV-MVDR (DC-CRN)	64.48	1.61	-4.48	76.15	1.83	-0.36	83.54	2.07	2.15
+ PF	81.17	2.15	1.91	89.70	2.67	4.54	92.38	2.91	5.62
MC-CSM (BLSTM)	76.43	1.94	1.20	85.37	2.42	3.14	88.28	2.63	3.55
MC-CSM (DC-CRN)	**83.55**	**2.21**	**2.84**	**91.04**	**2.74**	**5.17**	**93.58**	**3.00**	**6.13**
SC-CSM (BLSTM)	69.69	1.67	-1.08	81.30	2.21	1.52	85.40	2.46	2.20
SC-CSM (DC-CRN)	75.20	1.91	0.80	86.24	2.47	3.58	90.29	2.76	4.74

TABLE XVI
EVALUATION OF DIFFERENT APPROACHES FOR BINAURAL SPEECH ENHANCEMENT ON REAL BRIRS

SNR	-5 dB			0 dB			5 dB		
Metric	STOI (%)	PESQ	SI-SNR (dB)	STOI (%)	PESQ	SI-SNR (dB)	STOI (%)	PESQ	SI-SNR (dB)
Unprocessed	54.41	1.39	-9.82	64.51	1.55	-6.12	72.76	1.77	-3.63
Oracle TI-MVDR	63.47	1.53	-6.85	73.15	1.74	-3.29	80.16	1.96	-0.84
MB TI-MVDR (BLSTM)	61.31	1.51	-7.74	72.14	1.75	-4.33	79.57	1.99	-2.38
+ PF	72.33	1.83	-3.32	83.11	2.33	-1.20	86.87	2.60	-0.51
MB TI-MVDR (DC-CRN)	61.47	1.51	-7.68	72.26	1.75	-4.31	79.70	1.99	-2.37
+ PF	74.07	1.89	-3.16	84.05	2.38	-1.12	87.78	2.66	-0.41
CSM TI-MVDR (BLSTM)	60.37	1.57	-8.35	71.21	1.77	-4.75	78.53	1.99	-2.55
+ PF	68.77	1.60	-5.15	82.02	2.22	-2.96	86.08	2.51	-2.72
CSM TI-MVDR (DC-CRN)	61.91	1.56	-7.85	72.63	1.77	-4.33	79.86	2.00	-2.28
+ PF	77.68	1.96	-3.61	87.91	2.58	-1.66	90.86	2.87	-1.06
Oracle TV-MVDR	69.71	1.66	-3.33	79.32	1.92	**0.07**	85.63	2.19	**2.35**
CSM TV-MVDR (BLSTM)	62.36	1.61	-6.99	74.09	1.88	-3.43	81.22	2.14	-1.65
+ PF	70.90	1.66	-4.84	82.60	2.25	-3.29	86.03	2.50	-3.26
CSM TV-MVDR (DC-CRN)	65.26	1.63	-6.20	76.42	1.90	-2.88	83.07	2.16	-1.26
+ PF	79.19	1.99	-3.87	88.19	2.59	-1.94	90.75	2.86	-1.35
MC-CSM (BLSTM)	68.75	1.62	-6.16	80.55	2.18	-4.29	84.46	2.43	-4.09
MC-CSM (DC-CRN)	**78.66**	**1.99**	**-2.48**	**88.49**	**2.62**	-0.65	**91.56**	**2.94**	-0.16
SC-CSM (BLSTM)	62.16	1.41	-6.76	76.49	1.91	-4.49	82.01	2.22	-4.12
SC-CSM (DC-CRN)	67.94	1.54	-3.83	82.47	2.23	-1.28	87.96	2.62	-0.58

much fewer MAC operations than the corresponding deep learning based beamforming system, as detailed in Section VI-A. Monaural complex spectral mapping is a special case of the MC-CSM approach; in the monaural case the DNN becomes a nonlinear spectral filter. Therefore, complex spectral mapping presents a unified framework for single- and multi-channel speech separation.

We have investigated the MC-CSM approach with different array geometries for multi-channel speech dereverberation, speech enhancement and speaker separation. Comprehensive comparisons have been conducted between this approach and other widely-used beamforming techniques, including both conventional and deep learning based beamforming. Evaluation results show that the MC-CSM approach yields separation results comparable to or better than beamforming for different array geometries and speech separation tasks. This suggests that the MC-CSM spectrospatial filtering approach is generally effective for speech separation with fixed-geometry microphone arrays. In addition, the MC-CSM approach works well on binaural speech enhancement, which further demonstrates the capacity of this approach in multi-channel speech separation. As many real-world applications are equipped with a fixed microphone array like Amazon Echo, the MC-CSM approach is potentially a very practical choice, providing a competitive alternative to the dominant beamforming approach in multi-channel speech separation.

In addition to examining the MC-CSM approach, this paper proposes a new training criterion, i.e. location-based training, to achieve talker independency for multi-channel speaker separation. Different from widely-used PIT, this criterion makes label assignments based on the relative positions of speakers, and yields comparable separation performance without considering various speaker label permutations. We have also investigated a new method of generating BRIRs for DNN training in binaural speech separation. This method can generate arbitrary numbers of simulated BRIRs, just like the image source method, which would be very useful for advancing supervised binaural speech separation. Experimental results show that the MC-CSM models trained on simulated BRIRs generalize reasonably well to real BRIRs.

We conclude this paper by highlighting three additional insights.

- With a convolutional recurrent architecture, DC-CRN benefits from both the feature extraction capability of convolutional layers and the temporal modeling capability of recurrent layers. Such an architecture is advantageous over BLSTM especially for complex spectral mapping, given that spectrotemporal patterns in real and imaginary spectrograms are highly structured. This is confirmed by the evaluation results in Section VI, demonstrating that the DC-CRN system significantly outperforms the BLSTM system in all conditions.
- Mask-based beamforming combines multiple monaural T-F masks into a single one using a certain pooling operator (like median). The resulting mask is used as an *ad hoc* weighting mechanism to compute the spatial covariance matrices of speech and noise. In contrast, CSM based beamforming is a principled method, and provides the ground-truth spatial covariance matrices if the complex spectra are perfectly estimated.
- The inclusion of a magnitude term (see (16)) in the loss function is important for complex spectral mapping, due to the relative importance of magnitude over phase in speech separation. Although monaural complex spectral mapping works well without this term [43], its inclusion seems more important in the multi-channel case.

ACKNOWLEDGMENT

The authors would like to thank Steven van de Par and Stephan Ewert from the University of Oldenburg for providing their RAZR program.

REFERENCES

[1] S. Affes and Y. Grenier, "A source subspace tracking array of microphones for double talk situations," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. Conf.*, 1996, pp. 909–912.
[2] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoustical Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.
[3] J. Benesty, I. Cohen, and J. Chen, *Fundamentals of Signal Enhancement and Array Signal Processing*. Hoboken, NJ, USA: Wiley, 2018.
[4] C. P. Brown and R. O. Duda, "A structural model for binaural sound synthesis," *IEEE Speech Audio Process.*, vol. 6, no. 5, pp. 476–488, Sep. 1998.
[5] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.
[6] S. Chakrabarty and E. A. Habets, "Time-frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 4, pp. 787–799, Aug. 2019.
[7] Z. Chen *et al.*, "Continuous speech separation: Dataset and analysis," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 7284–7288.
[8] I. Cohen, "Identification of speech source coupling between sensors in reverberant noisy environments," *IEEE Signal Process. Lett.*, vol. 11, no. 7, pp. 613–616, Jul. 2004.
[9] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Speech Audio Process.*, vol. 12, no. 5, pp. 451–459, Sep. 2004.
[10] S. Doclo and M. Moonen, "Combined frequency-domain dereverberation and noise reduction technique for multi-microphone speech enhancement," in *Proc. Int. Workshop Acoust. Echo Noise Control*, 2001, pp. 31–34.
[11] D. E. Dudgeon, "Fundamentals of digital array processing," *Proc. IEEE*, vol. 65, no. 6, pp. 898–904, Jun. 1977.
[12] H. Erdogan *et al.*, "Multi-channel speech recognition: LSTMs all the way through," in *Proc. CHiME-4 Workshop*, 2016, pp. 1–4.
[13] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved MVDR beamforming using single-channel mask prediction networks," in *Proc. Interspeech*, 2016, pp. 1981–1985.
[14] S.-W. Fu, T.-Y Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *Proc. IEEE 27th Int. Workshop Mach. Learn. Signal Process.*, 2017, pp. 1–6.
[15] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.
[16] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multi-microphone speech enhancement and source separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.
[17] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete LDC93S6A," *Web Download. Philadelphia: Linguistic Data Consortium*, vol. 83, 1993.
[18] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1.," *Space-Terr. Integr. Netw.*, vol. 93, p. 27 403, 1993.
[19] R. Gu, S.-X. Zhang, Y. Zou, and D. Yu, "Complex neural spatial filter: Enhancing multi-channel target speech separation in complex domain," *IEEE Signal Process. Lett.*, vol. 28, pp. 1370–1374, Apr. 2021.
[20] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2016, pp. 31–35.
[21] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2016, pp. 196–200.
[22] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2016, pp. 5210–5214.
[23] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.
[24] M. J. Jo, G. W. Lee, J. M. Moon, C. Cho, and H. K. Kim, "Estimation of MVDR beamforming weights based on deep neural network," in *Proc. Audio Eng. Soc. 145th Conv. Audio Eng. Soc.*, 2018.
[25] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.
[26] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *IEEE Signal Process. Mag.*, vol. 13, no. 4, pp. 67–94, Jul. 1996.
[27] Y. Kubo, T. Nakatani, M. Delcroix, K. Kinoshita, and S. Araki, "Mask-based MVDR beamformer for noisy multisource environments: Introduction of time-varying spatial covariance model," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2019, pp. 6855–6859.
[28] B. Li, T. N. Sainath, R. J. Weiss, K. W. Wilson, and M. Bacchiani, "Neural network adaptive beamforming for robust multichannel speech recognition," *Interspeech*, pp. 1976–1980, 2016.
[29] Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S.-C. Liu, "FaSNet: Low-latency adaptive beamforming for multi-microphone audio processing," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 260–267.
[30] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.
[31] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.
[32] Z. Meng, S. Watanabe, J. R. Hershey, and H. Erdogan, "Deep long short-term memory adaptive beamforming networks for multichannel robust speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2017, pp. 271–275.

[33] P. Mowlaee and R. Saeidi, "On phase importance in parameter estimation in single-channel speech enhancement," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2013, pp. 7462–7466.
[34] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Phase importance in speech processing applications," in *Proc. Interspeech*, 2014, pp. 1623–1627.
[35] 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 J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1274–1288, Dec. 2017.
[36] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Commun.*, vol. 53, no. 4, pp. 465–494, 2011.
[37] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of Adam and beyond," in *Proc. Int. Conf. Learn. Representations*, 2018.
[38] 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 *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2001, pp. 749–752.
[39] T. N. Sainath *et al.*, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 5, pp. 965–979, May 2017.
[40] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, M. Bacchiani, and A. Senior, "Speaker location and microphone spacing invariant acoustic modeling from raw multichannel waveforms," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2015, pp. 30–36.
[41] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
[42] K. Tan and D. L. Wang, "A. convolutional recurrent neural network for real-time speech enhancement," in *Proc. Interspeech*, 2018, pp. 3229–3233.
[43] K. Tan and D. L. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 380–390, Nov. 2019.
[44] K. Tan, X. Zhang, and D. L. Wang, "Deep learning based real-time speech enhancement for dual-microphone mobile phones," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1853–1863, May 2021.
[45] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *J. Acoustical Soc. Amer.*, vol. 133, no. 5, pp. 3591–3591, 2013.
[46] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Acoust. Speech Signal Process. Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.
[47] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.
[48] E. Vincent, T. Virtanen, and S. Gannot. *Audio Source Separation and Speech Enhancement*. Hoboken, NJ, USA: Wiley, 2018.
[49] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Comput. Speech Lang.*, vol. 46, pp. 535–557, 2017.
[50] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 30, no. 4, pp. 679–681, Aug. 1982.
[51] D. L. Wang and G. J. Brown, editors. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ, USA: Wiley, 2006.
[52] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.
[53] Z.-Q. Wang and D. L. Wang, "All-neural multi-channel speech enhancement," in *Proc. Interspeech*, 2018, pp. 3234–3238.
[54] Z.-Q. Wang and D. L. Wang, "Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 2, pp. 457–468, Feb. 2018.
[55] Z.-Q. Wang and D. L. Wang, "On spatial features for supervised speech separation and its application to beamforming and robust ASR," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2018, pp. 5709–5713.
[56] Z.-Q. Wang and D. L. Wang, "Multi-microphone complex spectral mapping for speech dereverberation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 486–490.
[57] Z.-Q. Wang, P. Wang, and D. L. Wang, "Complex spectral mapping for single- and multi-channel speech enhancement and robust ASR," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1778–1787, May 2020.
[58] Z.-Q. Wang, P. Wang, and D. L. Wang, "Multi-microphone complex spectral mapping for utterance-wise and continuous speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2001–2014, May 2021.
[59] J. Y. Wen, N. D. Gaubitch, E. A. Habets, T. Myatt, and P. A. Naylor, "Evaluation of speech dereverberation algorithms using the MARDY database," in *Proc. Int. Workshop Acoust. Echo Noise Control*, 2006, pp. 1–4.
[60] T. Wendt, S. Van De Par, and S. D. Ewert, "A computationally-efficient and perceptually-plausible algorithm for binaural room impulse response simulation," *J. Audio Eng. Soc.*, vol. 62, no. 11, pp. 748–766, 2014.
[61] D. S. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 483–492, Mar. 2016.
[62] X. Xiao, S. Watanabe, E. S. Chng, and H. Li, "Beamforming networks using spatial covariance features for far-field speech recognition," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2016, pp. 1–6.
[63] X. Xiao *et al.*, "Deep beamforming networks for multi-channel speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 5745–5749.
[64] 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 *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2017, pp. 3246–3250.
[65] T. Yoshioka, H. Erdogan, Z. Chen, and F. Alleva, "Multi-microphone neural speech separation for far-field multi-talker speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2018, pp. 5739–5743.
[66] T. Yoshioka *et al.*, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2015, pp. 436–443.
[67] X. Zhang and D. L. Wang, "Deep learning based binaural speech separation in reverberant environments," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 5, pp. 1075–1084, May 2017.
[68] X. Zhang, Z.-Q. Wang, and D. L. Wang, "A. speech enhancement algorithm by iterating single- and multi-microphone processing and its application to robust ASR," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2017, pp. 276–280.
[69] 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," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 6089–6093.