Combining Spectral and Spatial Features for Deep Learning Based Blind Speaker Separation

Zhong-Qiu Wang, *Student Member, IEEE*, and DeLiang Wang, *Fellow, IEEE*

***Abstract*—This study tightly integrates complementary spectral and spatial features for deep learning based multi-channel speaker separation in reverberant environments. The key idea is to localize individual speakers so that an enhancement network can be trained on spatial as well as spectral features to extract the speaker from an estimated direction and with specific spectral structures. The spatial and spectral features are designed in a way such that the trained models are blind to the number of microphones and microphone geometry. To determine the direction of the speaker of interest, we identify time-frequency (T-F) units dominated by that speaker and only use them for direction estimation. The T-F unit level speaker dominance is determined by a two-channel chimera++ network, which combines deep clustering and permutation invariant training at the objective function level, and integrates spectral and inter-channel phase patterns at the input feature level. In addition, T-F masking based beamforming is tightly integrated in the system by leveraging the magnitudes and phases produced by beamforming. Strong separation performance has been observed on reverberant talker-independent speaker separation, which separates reverberant speaker mixtures based on a random number of microphones arranged in arbitrary linear-array geometry.**

***Index Terms*—Spatial features, beamforming, deep clustering, permutation invariant training, chimera++ networks, blind source separation.**

I. Introduction

RECENT years have witnessed major advances of monaural talker-independent speaker separation since the introduction of deep clustering [1]–[4], deep attractor networks [5] and permutation invariant training (PIT) [6], [7]. These algorithms address the label permutation problem in the challenging monaural speaker-independent setup [8], [9] and demonstrate substantial improvements over conventional algorithms, such as spectral clustering [10], computational auditory scene analysis based approaches [11] and target- or speaker-dependent systems [12], [8].

Manuscript received June 17, 2018; revised September 19, 2018 and November 9, 2018; accepted November 13, 2018. Date of publication November 19, 2018; date of current version December 6, 2018. This work was supported in part by an AFRL contract FA8750-15-1-0279, in part by the National Science Foundation under Grant IIS-1409431, and in part by the Ohio Supercomputer Center. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tuomas Virtanen. *(Corresponding author: Zhong-Qiu Wang.)*

Z.-Q. Wang is with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210-1277 USA (e-mail: wangzhon@cse.ohio-state.edu).

D. Wang is with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210-1277 USA, and also with 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.2018.2881912

When multiple microphones are available, spatial information can be leveraged to alleviate the label permutation problem, as speaker sources are directional and typically spatially separated in real-world scenarios. One conventional stream of research is focused on spatial clustering [13]–[15], where individual T-F units are clustered into sources using complex Gaussian mixture models (GMMs) or their variants based on spatial cues such as interchannel time, phase or level differences (ITDs, IPDs or ILDs) and spatial spread, under the speech sparsity assumption. However, such spatial cues degrade significantly in reverberant environments and lead to inadequate separation when the sources are co-located, close to one another or when spatial aliasing occurs. In addition, conventional spatial clustering typically does not exploit spectral information. In contrast, recent developments in deep learning based monaural speaker separation suggest that, even with spectral information alone, remarkable separation can be obtained [9], although most of such studies are only evaluated in anechoic conditions.

One promising research direction is hence to harness the merits of these two streams of research so that spectral and spatial processing can be tightly combined to improve separation and at the same time, make the trained models as blind as possible to microphone array configuration. In [16], [17], monaural deep clustering is employed for T-F masking based beamforming. Their methods follow the success of T-F masking based beamforming in the CHiME challenges [18]. Although beamforming is found to be very helpful in tasks such as robust automatic speech recognition (ASR), where distortionless response is a major concern, for tasks such as speaker separation and speech enhancement, it typically cannot achieve sufficient separation in reverberant environments, when sources are close to each other, or when the number of microphones is limited. For such tasks, performing further spectral masking would be very helpful. The studies in [19], [20] apply single-channel deep attractor networks on the outputs of a set of fixed beamformers. A major motivation in [20] is that fixed beamformers together with a separate beam prediction network can be efficient to compute in an online low-latency system. However, their approach requires the information of microphone geometry to carefully design the fixed beamformers, which are manually designed for a single fixed device based on its microphone geometry and hence are typically not as powerful as data-dependent beamformers that can exploit signal statistics for significant noise reduction, especially in offline scenarios. In addition, the fixed beamformers point towards a set of discretized directions. This could lead to resolution problems and would become cumbersome to apply

when elevation is a consideration. Different from the approaches that apply deep clustering and its variants on monaural spectral information, our recent study [21] includes interchannel phase patterns for the training of deep clustering networks to better resolve the permutation problem. The trained model can be directly applied to arrays with any number of microphones in different arrangements, and can be potentially applied to separating any number of sources. However, this approach only produces a magnitude-domain binary mask and does not exploit beamforming, which is capable of phase enhancement and is known to perform very well especially in modestly reverberant conditions or when many microphones are available.

In this context, our study tightly integrates spectral and spatial processing for blind source separation (BSS), where spatial information is encoded as additional input features to leverage the representational power of deep learning for better separation. The overall proposed approach is a *Separate-Localize-Enhance* strategy. More specifically, a two-channel chimera++ network that takes interchannel phase patterns into account is first trained to resolve the label permutation problem and perform initial separation. Next, the resulting estimated masks are used in a localization-like procedure to estimate speaker directions and signal statistics. After that, directional (or spatial) features, computed by compensating IPDs or by using data-dependent beamforming, are designed to combine all the microphones for the training of an enhancement network to further separate each source. Here, beamforming is incorporated in two ways: one uses the magnitude produced by beamforming as additional input features of the enhancement networks to improve the magnitude estimation of each source and the other further considers the phase provided by beamforming as the enhanced phase. We emphasize that the proposed approach aligns with human ability to focus auditory attention on one particular source with its associated spectral structures and arriving from a particular direction, and suppress the other sources [22].

Our study makes five major contributions. First, interchannel phase and level patterns are incorporated for the training of two-channel chimera++ networks. This approach, although straightforward, is found to be very effective for exploiting two-channel spatial information. Second, two effective spatial features are designed for the training of an enhancement network to utilize the spatial information contained in all the microphones. Third, data-dependent beamforming based on T-F masking is effectively integrated in our system by means of its magnitudes and phases. Fourth, a run-time iterative approach is proposed to refine the estimated masks for T-F masking based beamforming. Fifth, the trained models are blind to the number of microphones and microphone geometry. On reverberant versions of the speaker-independent wsj0-2mix and wsj0-3mix corpus [1], spatialized by measured and simulated room impulse responses (RIRs), the proposed approach exhibits large improvements over various algorithms including MESSL [23], oracle and estimated multi-channel Wiener filter, GCC-NMF [24], ILRMA [25] and multi-channel deep clustering [21].

In the rest of this paper, we first introduce the physical model in Section II, followed by a review of the monaural chimera++ networks [3] in Section III. Next, we extend them to two-microphone cases in Section IV.A. Based on the estimated masks obtained from pairwise microphone processing, Section IV.B encodes the spatial information contained in all the microphones as directional features to train an enhancement network for further separation, with or without utilizing the estimated phase produced by beamforming. An optional run-time iterative mask refining algorithm is presented in Section IV.C. Fig. 1 illustrates the proposed system. We present our experimental setup and evaluation results in Section V and VI, respectively, and conclude this paper in Section VII.

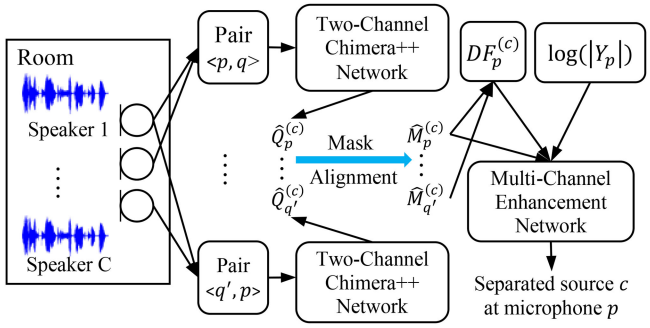

Fig. 1. Illustration of proposed system for BSS. A two-channel chimera++ network is applied to each microphone pair of interest for initial mask estimation. A multi-channel enhancement network is then applied for each source at a reference microphone for further separation.

II. Physical Model

Given a reverberant P-channel C-speaker time-domain mixture $\boldsymbol{y}[n] = \sum_{c=1}^{C} \boldsymbol{s}^{(c)}[n]$, the physical model in the short-time Fourier transform (STFT) domain is formulated as:

$$\boldsymbol{Y}(t,f) = \sum_{c=1}^{C} \boldsymbol{S}^{(c)}(t,f), \tag{1}$$

where $\boldsymbol{S}^{(c)}(t,f)$ and $\boldsymbol{Y}(t,f)$ respectively represent the P-dimensional STFT vectors of the reverberant image of source c and the reverberant mixture captured by the microphone array at time t and frequency f. Our study proposes multiple algorithms to separate the mixture Y_p captured at a reference microphone p to individual reverberant sources $\hat{S}_p^{(c)}$, by integrating single- and multi-channel processing under a deep learning framework. To improve the usability, it is highly desirable to make the trained models of our algorithms directly applicable to microphone arrays with various numbers of microphones arranged in diverse layouts. This property is especially useful for cloud-based services, where the client setup can vary significantly in terms of microphone array configuration or when array configuration is not available. Note that the proposed algorithms focus on separation and do not address de-reverberation, although they can be straightforwardly modified for that purpose.

III. Monaural Chimera++ Networks

Our recent study [3] proposed for monaural speaker separation a novel multi-task learning approach, which combines the permutation resolving capability of deep clustering [1], [2] and

the mask inference ability of PIT [6], [7], yielding significant improvements over the individual models. The objective function of deep clustering pulls in the T-F units dominated by the same speaker and pushes away those dominated by different speaker, creating hidden representations that can be utilized by PIT to predict continuous mask values more easily and more accurately. The objective function is also considered as a regularization term to improve the permutation resolving ability of utterance-level PIT. In this section, we first introduce deep clustering and permutation invariant training, and then review the chimera++ networks.

The key idea of deep clustering [1] is to learn a unit-length embedding vector for each T-F unit using a deep neural network such that for the T-F units dominated by the same speaker, their embeddings are close to one another, while farther otherwise. This way, simple clustering algorithms such as k-means can be applied to the embeddings at run time to determine the speaker assignment at each T-F unit. More specifically, let $\boldsymbol{v}_i$ denote the D-dimensional embedding vector of the ith T-F unit and $\boldsymbol{u}_i$ represent a C-dimensional one-hot vector denoting which of the C sources dominates the ith T-F unit. Vertically stacking them yields the embedding matrix $V \epsilon \mathbb{R}^{TF \times D}$ and the label matrix $U \epsilon \mathbb{R}^{TF \times C}$. The embeddings are learned to approximate the affinity matrix UU^T:

$$\mathcal{L}_{DC} = \left\| VV^T - UU^T \right\|_F^2 \tag{2}$$

Recent studies [3] suggested that a variant deep clustering loss function that whitens the embeddings based on a k-means objective leads to better separation performance.

$$\mathcal{L}_{DC,W} = \left\| V(V^TV)^{-\frac{1}{2}} - U\left(U^TU\right)^{-1}U^TV(V^TV)^{-\frac{1}{2}} \right\|_F^2 \tag{3}$$

$$= D - trace\left((V^TV)^{-1}V^TU\left(U^TU\right)^{-1}U^TV\right) \tag{4}$$

It is important in deep clustering to discount the importance of silence T-F units, as their labels are ambiguous and they do not carry directional phase information for multi-channel separation [21]. Following [3], the weight of each T-F is computed as the magnitude of each T-F unit over the sum of the magnitudes of all the T-F units. This weighting mechanism can be simply implemented by broadcasting the weight vector to V and U before computing the loss.

A recurrent neural network with bi-directional long short-term memory (BLSTM) units is usually utilized to model the contextual information from past and future frames. The network architecture of deep clustering is shown in the left branch of Fig. 2.

A permutation-free objective function was proposed in [1], and later reported to work well when combined with deep clustering in [2]. In [6], [7], a permutation invariant training technique was proposed, first showing that such objective function can produce comparable results by itself. The key idea is to train a neural network to minimize the minimum utterance-level loss of all the permutations. The phase-sensitive mask (PSM) [26] is typically used as the training target. Following [7], the loss function for phase-sensitive spectrum approximation (PSA) is defined as:

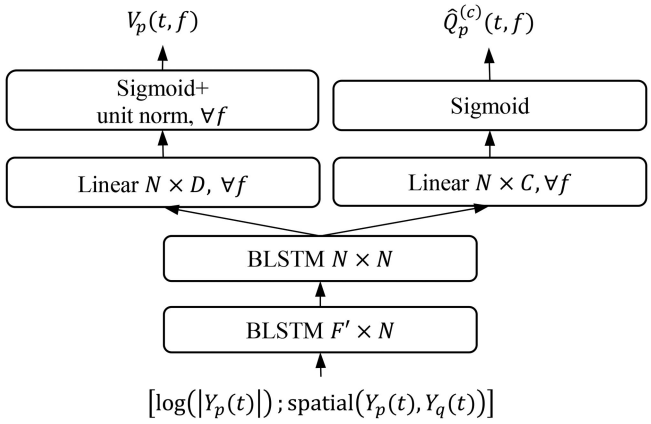

Fig. 2. Illustration of two-channel chimera++ networks on microphone pair $\langle p, q \rangle$. $\text{spatial}(Y_p(t), Y_q(t))$ can be a combination of $\cos(\angle Y_p - \angle Y_q)$, $\sin(\angle Y_p - \angle Y_q)$ and $\log(|Y_p|/|Y_q|)$ for microphones p and q. F' represents input feature dimension and N is number of units in each BLSTM layer.

$$\mathcal{L}_{PIT} = \min_{\varphi_p \epsilon \Psi} \sum\nolimits_c \left\| \hat{Q}_p^{\varphi_p(c)} |Y_p| - T_0^{|Y_p|}\left(\left|S_p^{(c)}\right| \cos\left(\angle S_p^{(c)} - \angle Y_p\right)\right) \right\|_1, \tag{5}$$

where p indexes a microphone channel, Ψ is a set of permutations over C sources, $S_p^{(c)}$ and Y_p are the STFT representations of source c and the mixture captured at microphone p, $T_0^{|Y_p|}(\cdot) = \max(0, \min(|Y_p|, \cdot))$ truncates the PSM to the range $[0, 1]$, $\hat{Q}$ denotes the estimated masks, $|\cdot|$ computes magnitude, and $\angle(\cdot)$ extracts phase. We denote the best permutation as $\hat{\varphi}_p(\cdot)$. Following our recent studies [27], [3], the L_1 loss is used as the loss function, as it leads to consistently better separation than the L_2 loss. Following [3], sigmoidal units are utilized in the output layer to obtain $\hat{Q}_p^{(c)}$ for separation. See the right branch of Fig. 2 for the network structure.

In [3], a multi-task learning approach is proposed to combine the merits of both algorithms. The objective function is a combination of the two loss functions:

$$\mathcal{L}_{chi++} = \alpha \mathcal{L}_{DC,W} + (1 - \alpha)\, \mathcal{L}_{PIT} \tag{6}$$

At run time, only the PIT output is needed to make predictions: $\hat{S}_p^{(c)} = \hat{Q}_p^{(c)} Y_p$. Here, the mixture phase is used for time-domain signal re-resynthesis.

IV. Proposed Algorithms

A. Two-Channel Extension of Chimera++ Networks

Following our previous studies on multi-channel speech enhancement [28], [29] and speaker separation [21], the key idea of the proposed approach for two-channel separation is to utilize not only spectral but also spatial features for model training. This way, complementary spectral and spatial information can be simultaneously utilized to benefit from the representational power of deep learning to better resolve the permutation problem and achieve better mask estimation. See Fig. 2 for an illustration of the network architecture.

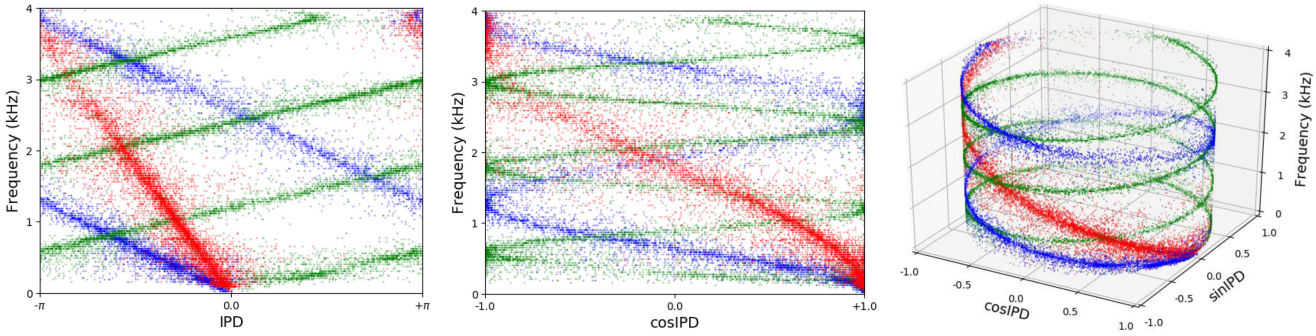

Fig. 3. Distribution of interchannel phase patterns of an example reverberant three-speaker mixture with $T_{60} = 0.54\ s$ and microphone spacing 21.6 cm. Each T-F unit is colored according to its dominant source. (a) IPD vs. Frequency; (b) cosIPD vs. Frequency; (c) cosIPD and sinIPD vs. Frequency.

Given a pair of microphones p and q with a random spacing, it is well-known that, because of speech sparsity, the STFT ratio $Y_p / Y_q = |Y_p| / |Y_q| e^{j(\angle Y_p - \angle Y_q)}$, which is indicative of the relative transfer function [30], naturally forms clusters within each frequency for spatially separated speaker sources with different time delays to the array [14], [13]. This property establishes the foundations of conventional narrowband spatial clustering [31]–[34], which typically first employs spatial information such as directional statistics and mixture STFT vectors for within-frequency bin-wise clustering based on complex GMM and its variants, and then aligns the clusters across frequencies. However, such approaches perform clustering largely based on spatial information, and typically do not leverage spectral cues, although there are recent attempts at using spectral embeddings produced by deep clustering for spatial clustering [16]. In addition, the clustering is usually only conducted independently within each frequency because of the IPD ambiguity, and thus does not exploit inter-frequency structures. By IPD ambiguity we mean that IPD varies with frequency and the underlying time delay cannot be uniquely determined only from the IPD at a frequency when spatial aliasing and phase wrapping occur.

Our study investigates the incorporation of the spatial information contained in Y_p/Y_q for the training of a two-channel chimera++ network. We consider the following interchannel phase and level patterns:

$$\text{IPD} = \angle e^{j(\angle Y_p - \angle Y_q)} = \text{mod}\left(\angle Y_p - \angle Y_q + \pi, 2\pi\right) - \pi \quad (7)$$

$$\cos \text{IPD} = \cos\left(\angle Y_p - \angle Y_q\right) \quad (8)$$

$$\sin \text{IPD} = \sin\left(\angle Y_p - \angle Y_q\right) \quad (9)$$

$$\text{ILD} = \log(|Y_p| / |Y_q|) \quad (10)$$

In our experiments, the combination of cosIPD and sinIPD leads to consistently better performance than the individual ones and the IPD. Our insight is that according to the Euler's formula, the distribution of cosIPD and sinIPD for directional sources naturally follows a helix-like structure with respect to frequency. See Fig. 3(c) for an illustration of the cosIPD and sinIPD distribution of a reverberant three-speaker mixture. Such helix structure could be exploited by a strong learning machine like deep neural networks to better model inter-frequency structures and achieve better separation. Indeed, in conventional spectral clustering, which significantly motivated the design of deep clustering [10], [1], it is suggested that spectral clustering has the capability of modeling such a distribution for clustering [35]. The distribution of an alternative representation, IPD, is depicted in Fig. 3(a). Clearly, the wrapped lines are not continuous across frequencies because of phase wrapping. Such abrupt discontinuity could make it harder for the neural network to exploit the inter-frequency structures. As a workaround, the distribution of cosIPD is depicted in Fig. 3(b). Although the continuity improves, without sinIPD, the number of crossings among the wrapped lines significantly increases. Such crossings, also observed in Fig. 3(a) and Fig. 3(c), are mostly resulted from spatial aliasing and phase wrapping, indicating that the interchannel phase patterns are indistinguishable even though the sources are spatially separated with different time delays and therefore posing fundamental difficulties for conventional BSS techniques that only utilize spatial information. In such cases, spectral information would be the only cue to rely on for separation. Our study hence also incorporates spectral features $\log(|Y_p|)$ for model training, and leverages the recently proposed chimera++ networks [3], which have been shown to produce state-of-the-art monaural separation, although only tested in anechoic conditions. Another advantage of including spectral features is that IPD itself is ambiguous across frequencies when the microphone spacing is large, meaning that there does not exist a one-to-one mapping between IPDs and ideal mask values. The incorporation of spectral features could help at resolving this ambiguity, as is suggested in our recent study [21]. Note that the chimera++ network naturally models all the frequencies simultaneously to exploit inter-frequency structures, hence avoiding an error-prone second-stage frequency alignment step that is necessary in conventional narrowband spatial clustering. In addition, the BLSTM better models temporal structures than complex GMMs and their variants, which typically make strong independence assumptions along the temporal axis.

We also incorporate ILDs, computed as in Eq. (10), to train chimera++ networks, as they become indicative about target directions especially when the microphone spacing is large and in setups like the binaural setup [11], [36].

B. Multi-Channel Speech Enhancement

To extend the proposed two-channel approach to multi-channel cases, one straightforward way is to concatenate the interchannel phase patterns and spectral features of all the microphone pairs as the input features for model training, as is done in [37]. However, this makes the input dimension dependent on the number of microphones and could make the trained model accustomed to one particular microphone geometry. Our recent study [21] proposes an ad-hoc approach to extend two-channel deep clustering to multi-channel cases by performing run-time K-means clustering on a super-vector obtained by concatenating the embeddings computed from each microphone pair. However, it only performs model training using pairwise microphone information, hence incapable of exploiting the geometrical constraints and the spatial information contained in all the microphones.

To build a model that is directly applicable to arrays with any number of microphones arranged in diverse layouts, we think that it is necessary to constructively combine all the microphones into a fixed-dimensional representation. Under this guideline, we propose two fixed-dimensional directional features, one based on compensating ambiguous IPDs using estimated phase differences and the other based on T-F masking based beamforming, as additional inputs to train an enhancement network to improve the mask estimation of each source at the reference microphone. See Fig. 1 for an illustration of the overall pipeline of our proposed approach. Note that at run time, we need to run the enhancement network once for each source for separation.

• *Compensated IPD:* More specifically, for the $P(\geq 2)$ microphones, we first apply the trained two-channel chimera++ network to each of the P pairs consisting of one pair $\langle p, q\rangle$ between the reference microphone p and a randomly-chosen non-reference microphone q, and $P-1$ pairs $\langle q', p\rangle$ for any non-reference microphone $q'(\neq p)$. The motivation of using this set of pairs is that we try to obtain an estimated mask for each source at each microphone. Note that for any non-reference microphone q', we can indeed randomly select another microphone to make a pair, but here we simply pair it and the reference microphone p. After obtaining the estimated masks $\hat{Q}_1^{(c)}, \ldots, \hat{Q}_P^{(c)}$ of all the P pairs from the two-channel chimera++ network, we permute the C masks at each microphone to create for each source c a new set of masks $\hat{M}_1^{(c)}, \ldots, \hat{M}_P^{(c)}$ such that they are all aligned to source c. At training time, such an alignment is readily available from Eq. (5), i.e., $\hat{M}_1^{(c)} = \hat{Q}_1^{\hat{\varphi}_1(c)}, \ldots, \hat{M}_P^{(c)} = \hat{Q}_P^{\hat{\varphi}_P(c)}$. At run time, we align the masks using Algorithm 1, where an average mask is maintained for each source in the alignment procedure to determine the best permutation for each non-reference microphone. We then compute the speech covariance matrix of each source using the aligned estimated masks, following recent developments of T-F masking based beamforming [38]–[40].

$$\hat{\Phi}^{(c)}(f) = \frac{1}{T}\sum_t \eta^{(c)}(t,f)\,\boldsymbol{Y}(t,f)\,\boldsymbol{Y}(t,f)^H, \tag{11}$$

Algorithm 1: Mask Alignment Procedure At Run Time. Binary Weight Matrix W Used In Step (4) Indicates T-F Units With Energy Larger Than -40 dB Of The Mixture's Maximum Energy.

Input: $\hat{Q}_1^{(c)}, \ldots, \hat{Q}_P^{(c)}$, for $c = 1, \ldots, C$, and reference microphone p.
Output: Aligned masks $\hat{M}_1^{(c)}, \ldots, \hat{M}_P^{(c)}$, for $c = 1, \ldots, C$;
(1) $\hat{M}_p^{(c)} = \hat{Q}_p^{(c)}$, for $c = 1, \ldots, C$;
(2) $\hat{M}_{avg}^{(c)} = \hat{M}_p^{(c)}$, for $c = 1, \ldots, C$;
(3) $counter = 1$;
For non-reference microphone q' in $\{1, \ldots, p-1, p+1, \ldots, P\}$ **do**
(4) $\varphi^* = \arg\min_{\varphi \epsilon \Psi} \sum_{c=1}^{C} \|W(\hat{M}_{avg}^{(c)} - \hat{Q}_{q'}^{\varphi(c)})\|_1$;
(5) $\hat{M}_{q'}^{(c)} = \hat{Q}_{q'}^{\varphi^*(c)}$, for $c = 1, \ldots, C$;
(6) $\hat{M}_{avg}^{(c)} = (\hat{M}_{avg}^{(c)} * counter + \hat{M}_{q'}^{(c)})/(counter + 1)$, for $c = 1, \ldots, C$;
(7) $counter += 1$;
End

where $(\cdot)^H$ computes Hermitian transposition, T is the number of frames, and $\eta^{(c)}(t,f)$ is the median [39] of the aligned estimated masks:

$$\eta^{(c)}(t,f) = \text{median}\left(\hat{M}_1^{(c)}(t,f), \ldots, \hat{M}_P^{(c)}(t,f)\right) \tag{12}$$

The key idea here is to only use the T-F units dominated by source c for the estimation of its covariance matrix. The steering vector for each source $\hat{\boldsymbol{r}}^{(c)}(f)$ is then computed as:

$$\hat{\boldsymbol{r}}^{(c)}(f) = \mathcal{P}\left\{\hat{\Phi}^{(c)}(f)\right\}, \tag{13}$$

where $\mathcal{P}\{\cdot\}$ compute the principal eigenvector. The motivation is that if $\hat{\Phi}^{(c)}(f)$ is well-estimated, it would be close to a rank-one matrix for a directional speaker source [38], [40], [13]. Its principal eigenvector is hence a reasonable estimate of the steering vector. This way of estimating steering vectors [38], [40] has been demonstrated to be very effective in recent CHiME challenges [18]. Note that this steering vector estimation step is essentially similar to direction of arrival (DOA) estimation.

Following our recent study [41], the directional features are then compensated in the following way:

$$DF_p^{(c)}(t,f) = \frac{1}{P-1}\sum_{\langle q',p\rangle \epsilon \Omega} \cos\Big\{\angle Y_{q'}(t,f) - \angle Y_p(t,f) - \left(\angle \hat{r}_{q'}^{(c)}(f) - \angle \hat{r}_p^{(c)}(f)\right)\Big\}, \tag{14}$$

where Ω contains all the $P-1$ pairs between each non-reference microphone q' and the reference microphone p. Here, $\angle Y_{q'}(t,f) - \angle Y_p(t,f)$ represents the observed phase difference and $\angle \hat{r}_{q'}^{(c)}(f) - \angle \hat{r}_p^{(c)}(f)$ the estimated phase difference (or the phase compensation term for source c). The motivation is that if a T-F unit is dominated by source c, the observed phase difference is expected to be aligned with its estimated phase difference. The phase compensation term is used to establish the consistency of the directional features along frequency such that

at any frequency and no matter which direction source c arrives from, a value close to one in $DF_p^{(c)}(t,f)$ would indicate that the T-F unit is likely dominated by the source c, while dominated by other sources if much smaller than one, only if the steering vector can be estimated accurately. This property makes the directional features highly discriminative for DNN based T-F masking to enhance the signal from a specific direction. In addition, by establishing the consistency along frequency, the phase compensation term alleviates the ambiguity of IPDs, which could be problematic when directly used for the training of the two-channel chimera++ networks in Section II.C. When there are more than two microphones, we simply average the compensated IPDs together. This makes the trained models directly applicable to microphone arrays with various numbers of microphones arranged in diverse geometry. The phase compensation term is designed to combine all the microphone pairs constructively.

There were previous studies [28], [42], [43], [29] utilizing spatial features for deep learning based speech enhancement (i.e., speech vs. noise). The spatial features in those studies are only designed for binaural speech enhancement, where only two sensors are considered and the target is right in the front direction. However, in more general cases, the target speaker may originate in any directions and the spatial features used in those studies would no longer work well. There was one speech enhancement study [43] considering compensating cosIPDs. However, it needs a separate DOA module that requires microphone geometry, and does not address DOA estimation in a robust way. Diffuseness features have also been applied in deep learning and T-F masking based beamforming for speech enhancement [41], [44]. However, such features are incapable of suppressing directional interferences, which we aim to suppress in this study. On the other hand, directional features are capable of suppressing diffuse noises.

• *T-F Masking Based Beamforming*: Another alternative directional feature is derived using beamforming, as beamforming can constructively combine target signals captured by different microphones and destructively for non-target signals, only if the signal statistics or target directions critical for beamforming can be accurately determined. Recent development in the CHiME challenges has suggested that deep learning based T-F masking can be utilized to compute such signal statistics accurately [18], demonstrating state-of-the-art robust ASR performance. Here, we leverage this recent development to construct a multi-channel Wiener filter [13]:

$$\hat{\boldsymbol{w}}_p^{(c)}(f) = \left(\hat{\Phi}^{(y)}(f)\right)^{-1} \hat{\Phi}^{(c)}(f)\boldsymbol{u}, \tag{15}$$

where $\hat{\Phi}^{(y)}(f) = \frac{1}{T}\sum_t \boldsymbol{Y}(t,f)\boldsymbol{Y}(t,f)^H$ is the mixture covariance matrix and $\boldsymbol{u}$ is a one-hot vector with u_p being one. Clearly, this way of constructing beamformers is blind to microphone geometry and the number of microphones. The directional feature is then computed as:

$$DF_p^{(c)}(t,f) = \log\left(\left|\hat{\boldsymbol{w}}_p^{(c)}(f)^H \boldsymbol{Y}(t,f)\right|\right) \tag{16}$$

• *Enhancement Network 1:* Clearly, using the spatial features alone for enhancement network training is not sufficient enough for accurate separation, as the sources could be spatially close and the reverberation components of other sources could also arrive from the estimated direction. We hence combine $DF_p^{(c)}$ with spectral features $\log(|Y_p|)$, and the initial mask estimates $\hat{M}_p^{(c)}$ obtained from the two-channel chimera++ network to train an enhancement network to estimate the phase-sensitve spectrum of source c at microphone p. This way, the neural network can take in both spectral and spatial information, and learn to enhance the signals with particular spectral characteristics and arriving from a particular direction. The objective function for training the enhancement network (denoted as $\mathbf{Enh}_1$) is:

$$\mathcal{L}_{Enh_1} = \left\| \hat{R}_p^{(c)} |Y_p| - T_0^{|Y_p|}\left(\left|S_p^{(c)}\right| \cos\left(\angle S_p^{(c)} - \angle Y_p\right)\right)\right\|_1, \tag{17}$$

where $\hat{R}_p^{(c)}$ denotes the estimated mask from the Enh_1 network. Following [27], the L_1 loss is used to compute the objective function. At run time, we execute the enhancement network once for each source, and the separated source c is obtained as $\hat{S}_p^{(c)} = \hat{R}_p^{(c)} Y_p$. Note that here the mixture phase is used for re-resynthesis.

• *Enhancement Network 2*: The above approach however cannot utilize the enhanced phase provided by beamforming. When the number of microphones is large, the enhanced phase $\hat{\theta}_p^{(c)}(t,f) = \angle(\hat{\boldsymbol{w}}_p^{(c)}(f)^H \boldsymbol{Y}(t,f))$ is expected to be better than $\angle Y_p$, if the speech distortion introduced by beamforming is minimal. We hence use the former as the phase estimate of source c. To obtain a good magnitude estimate, we train an enhancement network (denoted as $\mathbf{Enh}_2$) to predict the phase-sensitive spectrum of source c with respect to $|Y_p|e^{j\hat{\theta}_p^{(c)}}$, based on the same features used in Enh_1, i.e., $DF_p^{(c)}$, $\log(|Y_p|)$ and $\hat{M}_p^{(c)}$. The loss function used for training is:

$$\mathcal{L}_{Enh_2} = \left\| \hat{Z}_p^{(c)} |Y_p| - T_0^{|Y_p|}\left(\left|S_p^{(c)}\right| \cos\left(\angle S_p^{(c)} - \hat{\theta}_p^{(c)}\right)\right)\right\|_1, \tag{18}$$

where $\hat{Z}_p^{(c)}$ denotes the estimated mask of the Enh_2 network. At run time, the separated source c is obtained as $\hat{S}_p^{(c)} = \hat{Z}_p^{(c)}|Y_p|e^{j\hat{\theta}_p^{(c)}}$.

Different from the above two ways of integrating beamforming, another alternative is to extract spectral features from the beamformed mixture, train an enhancement network to predict the ideal masks computed from the beamformed sources, and at run time apply the estimated masks to the beamformed mixture [29]. In contrast, our approach uses beamforming results as directional features to improve the mask estimation at the reference microphone p, with or without using the phase of the beamformed mixture, since $S_p^{(c)}$, rather than beamformed sources $\boldsymbol{w}^{(c)}(f)^H \boldsymbol{S}^{(c)}(t,f)$, is considered as the reference for metric computation. This way, we can systematically compare the performance of single- and multi-channel processing, as well as the effects of various algorithms for reverberant source

separation. Note that we do not use beamformed sources as the reference signals for metric computation, as they usually contain speech distortions in reverberant environments, and are sensitive to the number of microphones, microphone geometry, and the type of beamformer used to obtain $\boldsymbol{w}^{(c)}(f)$. In addition, for BSS algorithms that do not involve any beamforming, such as spatial clustering or independent component analysis (ICA), it is not reasonable to use beamformed sources as the reference signals for evaluation. We will leave this alternative for future research on de-reverberation and multi-speaker ASR.

We emphasize again that our models, once trained, can be directly applied to arrays with any numbers of microphones arranged in various layouts. At run time, we can first apply the trained two-channel chimera++ network on each microphone pair of interest, then use Eq. (14) or (16) to constructively combine the spatial information contained in all the microphones, and finally apply the well-trained Enh_1 or Enh_2 networks for further separation. Note that the two-channel chimera++ network essentially functions as a DOA module to estimate target directions and signal statistics for spatial feature computation and beamforming. Indeed, it can be replaced by a monaural chimera++ network, while the two-channel one produces much better initial mask estimation because of the effective exploitation of spatial information, although in a very straightforward way.

C. Run-Time Iterative Mask Refinement

In Eq. (12), $\eta^{(c)}$ is computed from the estimated masks $\hat{M}_p^{(c)}$ produced by the chimera++ network that only exploits two-channel information. Such masks are expected to be not as accurate as $\hat{R}_p^{(c)}$ produced by Enh_1, which can utilize the spatial information from all the microphones and suffers less from IPD ambiguity. Using $\hat{R}_p^{(c)}$ for T-F masking based beamforming would hence likely leads to better beamforming results, which can in turn benefit the enhancement networks.

More specifically, at run time, after obtaining $\hat{R}_p^{(c)}$ using Enh_1, we use it in Eq. (12) to recompute a multi-channel Wiener filter $\widehat{\boldsymbol{w}}_p^{(c)}$ and feed the combination of $\log(|\widehat{\boldsymbol{w}}_p^{(c)}(f)^H \boldsymbol{Y}(t,f)|)$, $\log(|Y_p|)$ and $\hat{R}_p^{(c)}$ directly to Enh_2 to get $\widehat{\widehat{Z}}_p^{(c)}$. The separated source is then obtained as $\hat{S}_p^{(c)} = \widehat{\widehat{Z}}_p^{(c)} |Y_p| e^{j\widehat{\widehat{\theta}}_p^{(c)}}$, where $\widehat{\widehat{\theta}}_p^{(c)}(t,f) = \angle(\widehat{\boldsymbol{w}}_p^{(c)}(f)^H \boldsymbol{Y}(t,f))$. We denote this iterative mask estimation approach as $\textbf{Enh}_1+\textbf{Enh}_2$. We emphasize this approach is performed at run time and does not require any model training. Note that $\hat{R}_p^{(c)}$ can be improved with more iterations, but here we only do one iteration due to computation considerations.

V. Experimental Setup

We train our models using only simulated RIRs, while test on simulated as well as real-recorded RIRs. The RIRs are convolved with the anechoic two-speaker and three-speaker mixtures in the recently proposed wsj0-2mix and wsj0-3mix corpus[1] [1], each of which contains 20,000, 5,000 and 3,000 anechoic monaural speaker mixtures in its 30-hour training, 10-hour validation and 5-hour test data. Note that the speakers in the training set and test set are not overlapped. The task is hence speaker-independent. The signal to interference ratio (SIR) for wsj0-2mix mixtures are randomly drawn from -5 dB to 5 dB. For wsj0-3mix, the third speaker is added such that its energy is the same as that of the first two speakers combined. The sampling rate is 8 kHz.

Algorithm 2: Data Spatialization Process (Simulated RIRs).

Input: wsj0-3mix;
Output: spatialized reverberant wsj0-3mix;
For each source *s1*, source *s2*, source *s3* in wsj0-3mix **do**
  Sample room length r_x and width r_y from $[5, 10]$ m;
  Sample room height r_z from $[3, 4]$ m;
  Sample mic array height a_z from $[1, 2]$ m;
  Sample displacement n_x and n_y of mic array from $[-0.2, 0.2]$ m;
  Place array center at $[\frac{r_x}{2} + n_x, \frac{r_y}{2} + n_y, a_z]$ m;
  Sample microphone spacing a_r from $[0.02, 0.09]$ m;
  For $p = 1 : P(= 8)$ **do**
    Place mic p at $[\frac{r_x}{2} + n_x - \frac{P-1}{2}a_r + (p-1)a_r, \frac{r_y}{2} + n_y, a_z]$ m;
  End
  Sample speaker locations in the frontal plane:

$$s_x^{(1)}, s_y^{(1)}, s_z^{(1)} = a_z;$$
$$s_x^{(2)}, s_y^{(2)}, s_z^{(2)} = a_z;$$
$$s_x^{(3)}, s_y^{(3)}, s_z^{(3)} = a_z;$$

  such that any two speakers are at least 15° apart from each other with respect to the array center, and the distance from each speaker to the array center is in between $[0.75, 2]$ m;
  Sample T60 from $[0.2, 0.7]$ s;
  Generate impulse responses using RIR generator and convolve them with *s1*, *s2* and *s3*;
  Concatenate channels of reverberated *s1*, *s2* and *s3*, scale them to match SIR among original *s1*, *s2* and *s3*, and add them to obtain reverberated mixture;
End

The data spatialization process using simulated RIRs for wsj0-3mix is detailed in Algorithm 2. The RIR generator[2] is employed to generate the simulated RIRs. The general guideline is to make the setup as random as possible while still subject to realistic constraints. For each wsj0-3mix mixture, we randomly generate a room with random room characteristic, speaker locations, and microphone spacing. Our study considers a linear array setup, where the target speakers are placed in the frontal plane and are at least 15° apart from each other. We generate 20,000, 5,000, and 3,000 eight-channel mixtures for training,

[1] Available at http://www.merl.com/demos/deep-clustering
[2] Available at https://github.com/ehabets/RIR-Generator

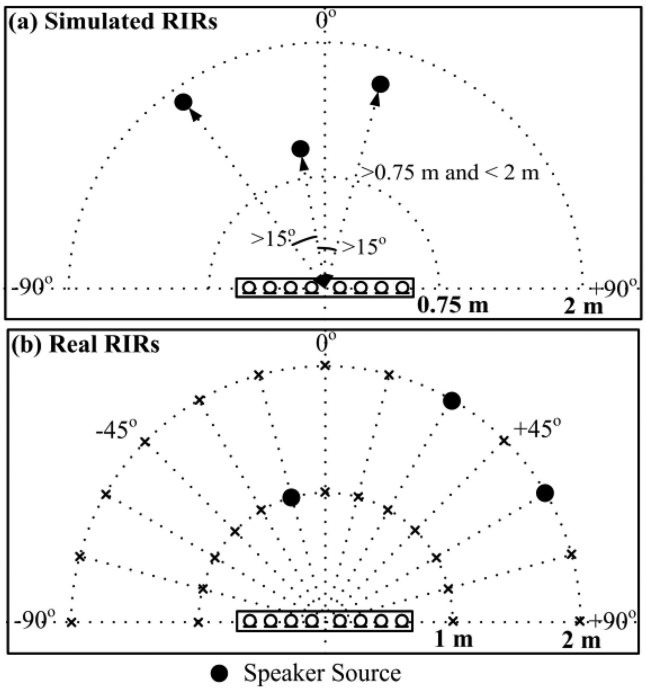

Fig. 4. Illustration of experimental setup.

validation and testing, respectively. A T60 value for each mixture is randomly drawn in the range [0.2, 0.7] s. See Fig. 4(a) for an illustration of this setup. The spatialization of wsj0-2mix is performed in a similar way. The average speaker-to-microphone distance is 1.38 m with 0.37 m standard deviation and the average direct-to-reverberant energy ratio (DRR) is 0.49 dB with 3.92 dB standard deviation.

We also generate another 3,000 eight-channel mixtures using the Multi-Channel Impulse Responses Database[3] [45], which is recorded at Bar-Ilan University using eight-microphone linear arrays with three different inter-microphone spacing, including 3-3-3-8-3-3-3, 4-4-4-8-4-4-4, 8-8-8-8-8-8-8 cm, under three reverberant time (0.16, 0.36, 0.61 s) created by using a number of covering panels on the walls. The RIRs are measured in steps of 15° from $-90°$ to 90° and at a distance of 1 m and 2 m to the array center, in a room with size approximately at $6 \times 6 \times 2.4$ m. See Fig. 4(b) for an illustration of this setup. For each mixture, we place each speaker in a random direction and at a random distance, using a randomly-chosen linear array and a randomly-chosen reverberation time among 0.16, 0.36 and 0.61 s. Note that for any two speakers, they are at least $15°$ apart with respect to the array center. The average DRR is 2.8 dB with 3.8 dB standard derivation in this case. We emphasize that this is a very realistic setup, as it is speaker-independent and more importantly, we use simulated RIRs for training and real RIRs for testing.

At run time, we randomly pick a subset of microphones for each utterance for testing. The aperture size can be 2 cm at minimum and 63 cm at maximum for the simulated RIRs, and 3 cm and 56 cm for the real RIRs.

[3]Available at http://www.eng.biu.ac.il/∼gannot/RIR_DATABASE/

TABLE I
SDR (dB) RESULTS ON SPATIALIZED REVERBERANT WSJ0-2MIX USING UP TO TWO MICROPHONES

Approaches	Input Features	Simu RIRs	Real RIRs
Unprocessed	-	0.0	0.0
1ch PIT	$\log(\lvert Y_p \rvert)$	7.5	7.3
1ch deep clustering	$\log(\lvert Y_p \rvert)$	7.3	7.4
1ch chimera++	$\log(\lvert Y_p \rvert)$	8.4	8.4
2ch chimera++	$\log(\lvert Y_p \rvert)$,IPD	10.2	9.8
2ch chimera++	$\log(\lvert Y_p \rvert)$,cosIPD	9.7	10.0
2ch chimera++	$\log(\lvert Y_p \rvert)$,cosIPD,sinIPD	10.4	10.1
+ Enh_1	$\log(\lvert Y_p \rvert)$,$\widehat{M}_p^{(c)}$	10.7	10.5
+ Enh_1	$\log(\lvert Y_p \rvert)$,$DF_p^{(c)}$(Eq. (14)),$\widehat{M}_p^{(c)}$	10.8	10.7
+ Enh_1	$\log(\lvert Y_p \rvert)$,$DF_p^{(c)}$(Eq. (16)),$\widehat{M}_p^{(c)}$	11.1	11.1
2ch chimera++	$\log(\lvert Y_p \rvert)$,cosIPD,sinIPD,ILD	10.4	10.1

The chimera++ and enhancement network respectively contains four and three BLSTM layers, each with 600 units in each direction. We cut each mixture into 400-frame segments and use these segments to train our models. The Adam algorithm is utilized for optimization. A dropout rate of 0.3 is applied to the output of each BLSTM layer. The window size is 32 ms and the hop size is 8 ms. A 256-point DFT is applied to extract 129-dimensional log magnitude features after square-root Hann window is applied to the signal. The α in Eq. (6) is empirically set to 0.975 and the embedding dimension D set to 20, following [3]. We emphasize that the enhancement network is trained using the directional features computed from various numbers of microphones, as the quality of the directional features varies with the number of microphones. For all the input features, we apply global mean-variance normalization before feed-forwarding.

Following the SiSEC challenges [46], average signal-to-distortion ratio (SDR) computed using the *bss_eval_images* software is used as the major evaluation metric. We also report average perceptual estimation of speech quality (PESQ) and extended short-time objective intelligibility (eSTOI) [47] scores to measure speech quality and intelligibility. Note that we consider the reverberant image of each source at the reference microphone, i.e., $s_p^{(c)}$, as the reference signal for metric computation.

VI. EVALUATION RESULTS

We first report the results on the reverberant wsj0-2mix spatialized using the simulated RIRs in the second last column of Table I. Clearly, the chimera++ network shows clear improvements over the individual models (8.4 vs. 7.5 and 7.3 dB), which align with the findings in [3]. Even with random microphone spacing, incorporating interchannel phase patterns for model training produces large improvement compared with only using monaural spectral information. This is likely because interchannel phase patterns naturally form clusters within each frequency regardless of microphone spacing, and we use a clustering-based DNN model to exploit such information for separation. Among various forms of IPD features, the combination of cosIPD and

TABLE II
SDR (DB) RESULTS ON SPATIALIZED REVERBERANT WSJ0-3MIX USING UP TO TWO MICROPHONES

Approaches	Input Features	Simu RIRs	Real RIRs
Unprocessed	-	-3.3	-3.2
1ch chimera++	$\log(\lvert Y_p \rvert)$	4.0	4.0
2ch chimera++	$\log(\lvert Y_p \rvert)$,IPD	7.1	6.1
2ch chimera++	$\log(\lvert Y_p \rvert)$,cosIPD	5.8	5.9
2ch chimera++	$\log(\lvert Y_p \rvert)$,cosIPD,sinIPD	7.3	6.3
+ Enh_1	$\log(\lvert Y_p \rvert)$,$\widehat{M}_p^{(c)}$	7.6	6.7
+ Enh_1	$\log(\lvert Y_p \rvert)$,$DF_p^{(c)}$(Eq. (14)),$\widehat{M}_p^{(c)}$	7.8	6.9
+ Enh_1	$\log(\lvert Y_p \rvert)$,$DF_p^{(c)}$(Eq. (16)),$\widehat{M}_p^{(c)}$	7.9	7.1

sinIPD leads to consistently better performance over using IPD or cosIPD (10.4 vs. 10.2 and 9.7 dB), likely because this combination naturally maintains the helix structures that can be exploited by the network. Further including the ILD features for training does not lead to clear improvement (10.4 vs. 10.4 dB), likely because level differences are very small in far-field conditions. Using the Enh_1 network brings further improvement as it provides better magnitude estimates. Compensating IPDs (i.e., Eq. (14)) using estimated phase differences to reduce the ambiguity and using beamforming results (i.e., Eq. (16)) as directional features push the performance from 10.4 to 10.8 and 11.1 dB, respectively. The former feature is worse than the latter one, likely because the former is mathematically similar to the delay-and-sum beamformer, which is known to be less powerful than the multi-channel Wiener filter. In the following experiments, we use Eq. (16) to compute the directional feature if not specified. The last column of Table I presents the results on the real RIRs. The performance is as comparably good as on the simulated RIRs, although the model is trained only on the simulated RIRs.

Table II presents the results obtained on the spatialized wsj0-3mix using the simulated RIRs and real RIRs, with up to two microphones. Similar trends as in Table I are observed.

Table III and Table IV compare the proposed algorithms with other systems along with the oracle performance of various ideal masks, using up to eight microphones, and in terms of SDR, PESQ and eSTOI. Because of utilizing the phase provided by beamforming, Enh_2 shows consistent improvement over Enh_1, especially when more microphones are available. This justifies the proposed way of integrating beamforming for separation. Performing run-time iterative mask refinement using Enh_1+Enh_2 leads to slight improvement over Enh_2 in the two-speaker case, while clear improvement is observed in the three-speaker case, especially when more microphones are available. This indicates the effectiveness of using $\hat{R}_p^{(c)}$ for T-F masking based beamforming, especially when $\hat{M}_p^{(c)}$ is not good enough.

Recent studies [17] apply monaural deep clustering on each microphone signal to derive a T-F masking based beamformer for each frequency for separation. To compare with their algorithms, we use the truncated PSM (tPSM), computed as $T_0^{1.0}(|S_p^{(c)}|\cos(\angle S_p^{(c)} - \angle Y_p)/|Y_p|)$, in Eq. (12) to compute oracle $\hat{\Phi}^{(c)}$ and report oracle MCWF results (denoted as tPSM-MCWF). We also report the estimated MCWF (eMCWF) performance obtained using $\hat{M}_p^{(c)}$ computed from the two-channel chimera++ network. Clearly, the beamforming approach requires relatively large number of microphones to produce reasonable separation. Although using estimated masks, the eMCWF is comparable to tPSM-MCWF. As can be observed, both of them are not as good as Enh_2, which combines beamforming with spectral masking. We also compare the proposed algorithms with MESSL[4] [23], a popular wideband GMM based spatial clustering algorithm proposed for two-microphone arrays, and GCC-NMF[5] [24], a location based stereo BSS algorithm, where dictionary atoms obtained from non-negative matrix factorization (NMF) are assigned to individual sources over time according to their time difference of arrival estimates obtained from GCC-PHAT. Note that oracle microphone spacing information is supplied to MESSL and GCC-NMF for the enumeration of time delays. Independent low-rank matrix analysis (ILRMA)[6] [25], originated from the ICA stream of research, is a strong and representative algorithm for determined and over-determined BSS. It unifies independent vector analysis (IVA) and multi-channel NMF by exploiting NMF decomposition to capture the spectral characteristics of each source as the generative source model in IVA. The recently proposed multi-channel deep clustering (MCDC) [21] integrates conventional spatial clustering with deep clustering by including interchannel phase patterns to train deep clustering networks. Its extension to multi-channel cases is achieved by first applying a well-trained two-channel deep clustering model on every microphone pair, then stacking the embeddings obtained from all the pairs, and finally performing K-means on the stacked embeddings to obtain an estimated binary mask for separation. Following the suggestions by an anonymous reviewer, we evaluate two extensions of MCDC as alternative ways of exploiting multi-channel spatial information. The first one, denoted as MC-Chimera++, concatenates the embeddings provided by our two-channel chimera++ network for K-means clustering, and the second one uses the median mask produced in Eq. (12) for separation, i.e., $\hat{S}_p^{(c)} = \eta^{(c)} Y_p$. Clearly, the proposed algorithms are consistently better than the MCDC approach and the two extensions, likely because the proposed algorithm is more end-to-end and better exploits spatial information contained in more than two microphones.

The performance of various oracle masks is presented in the last columns of Table III and Table IV. The ideal binary mask (IBM) is computed based on which source is dominant at each T-F unit. The ideal ratio mask (IRM) is calculated as the magnitude of each source over the sum of all the magnitudes. Compared with such monaural ideal masks that use mixture phase for re-synthesis, the multi-channel tPSM (MC-tPSM), calculated as $T_0^{1.0}(|S_p^{(c)}|\cos(\angle S_p^{(c)} - \hat{\theta}_p^{(c)})/|Y_p|)$ where $\hat{\theta}_p^{(c)}$ here is computed from tPSM-MCWF and used as the phase for re-synthesis, is clearly better and becomes even better when more microphones are available. Note that MC-tPSM represents the

[4]Available at https://github.com/mim/messl
[5]Available at https://github.com/seanwood/GCC-nmf
[6]Available at http://d-kitamura.net/programs/ILRMA_release20180411.zip

TABLE III
PERFORMANCE COMPARISON WITH OTHER APPROACHES ON **REAL RIRS** USING VARIOUS NUMBERS OF MICROPHONES ON SPATIALIZED REVERBERANT WSJ0-2MIX

Metrics	#mics	Mixture	MESSL [23]	GCC-NMF [24]	ILRMA [25]	MCDC [21]	MC-Chimera++	Using $\eta^{(c)}$ in Eq. (12)	eMCWF	Enh_1	Enh_2	Enh_1+Enh_2	tPSM-MCWF	Oracle Masks			
														IRM	IBM	tPSM	MC-tPSM
SDR (dB)	2	0.0	4.1	5.0	8.9	9.2	9.4	10.2	6.7	11.1	11.1	11.2	7.1	12.1	13.0	14.1	14.1
	3		-	-	9.5	9.6	9.8	10.4	8.1	11.5	11.9	12.1	8.6				14.8
	4		-	-	9.5	9.8	9.9	10.5	9.0	11.7	12.5	12.7	9.6				15.3
	5		-	-	9.7	9.9	10.0	10.6	9.7	11.8	13.0	13.2	10.4				15.8
	6		-	-	9.8	10.0	10.0	10.6	10.3	11.9	13.3	13.6	11.0				16.2
	7		-	-	9.8	10.0	10.0	10.6	10.7	12.0	13.6	13.9	11.5				16.5
	8		-	-	9.7	10.0	10.1	10.6	11.0	12.0	13.8	14.2	11.9				16.7
PESQ	2	2.06	2.27	2.16	2.73	2.19	2.20	2.98	2.51	3.12	3.21	3.24	2.53	3.79	3.29	3.83	3.92
	3		-	-	2.80	2.24	2.24	2.98	2.66	3.23	3.35	3.40	2.69				3.97
	4		-	-	2.82	2.26	2.26	3.01	2.75	3.29	3.43	3.48	2.79				4.00
	5		-	-	2.83	2.27	2.27	3.01	2.81	3.33	3.49	3.54	2.86				4.02
	6		-	-	2.84	2.27	2.27	3.03	2.86	3.35	3.52	3.58	2.91				4.04
	7		-	-	2.84	2.27	2.27	3.02	2.90	3.37	3.55	3.60	2.96				4.06
	8		-	-	2.84	2.27	2.27	3.03	2.93	3.38	3.57	3.63	2.99				4.07
eSTOI (%)	2	54.8	58.9	56.7	73.8	71.8	72.5	79.0	65.8	82.1	83.4	84.1	66.7	92.1	87.7	92.7	94.0
	3		-	-	75.6	73.5	74.0	79.2	70.5	83.7	85.6	86.4	71.8				94.6
	4		-	-	76.0	74.2	74.5	79.9	73.4	84.7	87.0	87.8	74.9				95.1
	5		-	-	76.5	74.6	74.8	80.0	75.6	85.3	87.9	88.7	77.2				95.4
	6		-	-	76.7	74.8	74.9	80.2	77.2	85.8	88.5	89.3	79.0				95.7
	7		-	-	76.7	74.9	75.0	80.2	78.5	86.1	89.0	89.8	80.4				95.9
	8		-	-	76.7	74.9	75.0	80.3	79.4	86.3	89.4	90.2	81.4				96.1

TABLE IV
PERFORMANCE COMPARISON WITH OTHER APPROACHES ON **REAL RIRS** USING VARIOUS NUMBERS OF MICROPHONES ON SPATIALIZED REVERBERANT WSJ0-3MIX

Metrics	#mics	Mixture	MESSL [23]	GCC-NMF [24]	ILRMA [25]	MCDC [21]	MC-Chimera++	Using $\eta^{(c)}$ in Eq. (12)	eMCWF	Enh_1	Enh_2	Enh_1+Enh_2	tPSM-MCWF	Oracle Masks			
														IRM	IBM	tPSM	MC-tPSM
SDR (dB)	2	-3.2	2.0	2.6	-	5.6	5.5	6.6	3.9	7.1	7.3	7.4	4.5	9.2	10.1	11.3	11.6
	3		-	-	4.6	6.1	5.9	6.7	4.9	7.5	7.9	8.2	5.7				12.1
	4		-	-	5.0	6.3	6.2	7.0	5.7	7.8	8.4	8.8	6.5				12.5
	5		-	-	5.1	6.4	6.3	7.2	6.3	8.0	8.9	9.4	7.2				12.9
	6		-	-	5.2	6.5	6.4	7.3	6.7	8.2	9.3	9.8	7.7				13.2
	7		-	-	5.2	6.5	6.4	7.3	7.0	8.3	9.6	10.1	8.2				13.5
	8		-	-	5.3	6.5	6.4	7.3	7.3	8.4	9.8	10.4	8.5				13.7
PESQ	2	1.67	1.87	1.68	-	1.49	1.48	2.45	2.10	2.48	2.55	2.59	2.14	3.60	2.87	3.64	3.73
	3		-	-	2.22	1.55	1.54	2.46	2.26	2.64	2.74	2.81	2.30				3.79
	4		-	-	2.26	1.57	1.56	2.53	2.35	2.73	2.85	2.94	2.41				3.83
	5		-	-	2.28	1.58	1.57	2.54	2.43	2.81	2.95	3.05	2.48				3.85
	6		-	-	2.29	1.59	1.58	2.56	2.48	2.84	3.00	3.12	2.54				3.87
	7		-	-	2.30	1.59	1.59	2.56	2.52	2.88	3.05	3.17	2.59				3.89
	8		-	-	2.31	1.59	1.59	2.57	2.55	2.90	3.09	3.21	2.63				3.91
eSTOI (%)	2	37.5	43.3	37.9	-	53.0	52.4	62.5	47.5	65.4	66.9	68.2	49.4	87.6	80.4	88.5	90.2
	3		-	-	54.3	55.5	55.0	62.9	53.2	68.5	70.7	72.5	55.9				91.2
	4		-	-	56.3	56.7	56.4	64.9	57.2	70.7	73.4	75.5	60.0				91.8
	5		-	-	57.0	57.3	56.9	65.2	60.1	72.4	75.5	77.8	63.1				92.3
	6		-	-	57.5	57.6	57.3	65.9	62.2	73.4	76.8	79.2	65.4				92.7
	7		-	-	57.8	57.7	57.4	65.8	63.9	74.2	77.9	80.3	67.4				93.0
	8		-	-	58.0	57.6	57.6	66.2	65.2	74.7	78.6	81.1	69.0				93.3

upper bound performance of Enh_2. The results clearly show the effectiveness of using $\hat{\theta}_p^{(c)}$ as the phase estimate.

By exploiting spatial information, we improve the performance of monaural chimera++ network from 8.4 to 11.2 dB when using two microphones and to 14.2 dB when using eight microphones on the spatialized wsj0-2mix corpus, and from 4.0 to 7.4 and 10.4 dB on the spatialized wsj0-3mix corpus. These results are comparable to the oracle performance of the monaural IBM, IRM and tPSM in terms of the SDR metric, confirming the effectiveness of multi-channel processing.

VII. CONCLUDING REMARKS

We have proposed a novel approach that combines complementary spectral and spatial features for deep learning based multi-channel speaker separation in reverberant environments.

This spatial feature approach is found to be very effective for improving the magnitude estimate of the target speaker in an estimated direction and with particular spectral structures. In addition, leveraging the enhanced phase provided by masking based beamforming driven by a two-channel chimera++ network produces further improvements. Future research will consider simultaneous separation and de-reverberation, which can be simply approached by using direct sound as the target in the PIT branch of the chimera++ network and in the outputs of the enhancement network, as well as applications to multi-speaker ASR. We shall also consider combining the proposed approach with end-to-end optimization [4].

Before closing, we point out that our current study has several limitations that need to be addressed in future work. First, similar to many deep learning based monaural speaker separation studies, our approach assumes that the number of speakers is known in advance. Second, our current system is focused on offline processing to push performance boundaries. To built an online low-latency system, one should consider replacing BLSTMs with uni-directional LSTMs, and accumulating the signal statistics, such as $\hat{\Phi}^{(y)}(f)$ and $\hat{\Phi}^{(c)}(f)$, used in beamforming in an online fashion. Third, our current system deals with reverberant speaker separation and no environmental noise is considered. Future research will need to consider de-noising as well, perhaps by extending our recent work in [41] and [48]. We shall also consider algorithms and experiments on conditions with shorter utterances, moving speakers, and even stronger reverberations, as they appear to pose challenges for masking based beafmorming in some ASR applications [49], [50].

ACKNOWLEDGMENT

We would like to thank Dr. J. Le Roux and Dr. J. R. Hershey for helpful discussions, and the anonymous reviewers for their constructive comments.

REFERENCES

[1] 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.

[2] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. Interspeech*, 2016, pp. 545–549.

[3] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 686–690.

[4] Z.-Q. Wang, J. Le Roux, D. L. Wang, and J. R. Hershey, "End-to-end speech separation with unfolded iterative phase reconstruction," in *Proc. Interspeech*, 2018, pp. 2708–2712.

[5] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 4, pp. 787–796, Apr. 2018.

[6] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 241–245.

[7] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multi-talker 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.

[8] 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.

[9] Y.-M. Qian, C. Weng, X. Chang, S. Wang, and D. Yu, "Past review, current progress, and challenges ahead on the cocktail party problem," *Frontiers Inf. Technol. Electron. Eng.*, vol. 19, pp. 40–63, 2018.

[10] F. Bach and M. Jordan, "Learning spectral clustering, with application to speech separation," *J. Mach. Learn. Res.*, vol. 7, pp. 1963–2001, 2006.

[11] D. L. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ, USA: Wiley, 2006.

[12] X.-L. Zhang and D. L. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 5, pp. 967–977, May 2016.

[13] 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.

[14] M. I. Mandel and J. P. Barker, "Multichannel spatial clustering using model-based source separation," *New Era Robust Speech Recognit. Exploiting Deep Learn.*, pp. 51–78, 2017.

[15] N. Ito, S. Araki, and T. Nakatani, "Recent advances in multichannel source separation and denoising based on source sparseness," *Audio Source Separation*, pp. 279–300, 2018.

[16] L. Drude and R. Haeb-Umbach, "Tight integration of spatial and spectral features for BSS with deep clustering embeddings," in *Proc. Interspeech*, 2017, pp. 2650–2654.

[17] T. Higuchi, K. Kinoshita, M. Delcroix, K. Zmolkova, and T. Nakatani, "Deep clustering-based beamforming for separation with unknown number of sources," in *Proc. Interspeech*, 2017, pp. 1183–1187.

[18] 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.

[19] Z. Chen, J. Li, X. Xiao, T. Yoshioka, H. Wang, Z. Wang, and Y. Gong, "Cracking the cocktail party problem by multi-beam deep attractor network," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2017, pp. 437–444.

[20] Z. Chen, T. Yoshioka, X. Xiao, J. Li, M. L. Seltzer, and Y. Gong, "Efficient integration of fixed beamformers and speech separation networks for multi-channel far-field speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5384–5388.

[21] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 1–5.

[22] C. Darwin, "Listening to speech in the presence of other sounds," *Philosophical Trans. Roy. Soc. B, Biol. Sci.*, vol. 363, no. 1493, pp. 1011–1021, 2008.

[23] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 18, no. 2, pp. 382–394, Feb. 2010.

[24] S. U. N. Wood, J. Rouat, S. Dupont, and G. Pironkov, "Blind speech separation and enhancement with GCC-NMF," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 745–755, Apr. 2017.

[25] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation with independent low-rank matrix analysis," *Audio Source Separation*, pp. 125–155, 2018.

[26] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 708–712.

[27] Z.-Q. Wang and D. L. Wang, "Recurrent deep stacking networks for supervised speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 71–75.

[28] Y. Jiang, D. L. Wang, R. Liu, and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 2112–2121, Dec. 2014.

[29] 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.

[30] Z.-Q. Wang and D. L. Wang, "Mask-weighted STFT ratios for relative transfer function estimation and its application to robust ASR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5619–5623.

[31] H. Sawada, S. Araki, and S. Makino, "A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2007, pp. 139–142.

[32] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1830–1840, Sep. 2010.
[33] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 516–527, Mar. 2011.
[34] T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, and T. Nakatani, "Online MVDR beamformer based on complex Gaussian mixture model with spatial prior for noise robust ASR," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 780–793, Apr. 2017.
[35] U. Shaham, K. Stanton, H. Li, B. Nadler, R. Basri, and Y. Kluger, "SpectralNet: Spectral clustering using deep neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2018.
[36] J. Traa, M. Kim, and P. Smaragdis, "Phase and level difference fusion for robust multichannel source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 6687–6691.
[37] 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.
[38] 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.
[39] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2015, pp. 444–451.
[40] 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.
[41] 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.
[42] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani, "Exploring multi-channel features for denoising-autoencoder-based speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 116–120.
[43] P. Pertilä and J. Nikunen, "Distant speech separation using predicted time-frequency masks from spatial features," *Speech Commun.*, vol. 68, pp. 97–106, 2015.
[44] Y. Liu, A. Ganguly, K. Kamath, and T. Kristjansson, "Neural network based time-frequency masking and steering vector estimation for two-channel MVDR beamforming," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 6717–6721.
[45] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. Int. Workshop Acoust. Signal Enhancement*, 2014, pp. 313–317.
[46] F.-R. Stöter, A. Liutkus, and N. Ito, "The 2018 signal separation evaluation campaign," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, 2018, pp. 293–305.
[47] 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.
[48] Z.-Q. Wang and D. L. Wang, "All-neural multi-channel speech enhancement," in *Proc. Interspeech*, 2018, pp. 3234–3238.
[49] J. Heymann, M. Bacchiani, and T. Sainath, "Performance of mask based statistical beamforming in a smart home scenario," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 6722–6726.
[50] C. Boeddeker, H. Erdogan, T. Yoshioka, and R. Haeb-Umbach, "Exploring practical aspects of neural mask-based beamforming for far-field speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 6697–6701.

Zhong-Qiu Wang (S'16) received the B.E. degree in computer science and technology from the Harbin Institute of Technology, Harbin, China, in 2013, and the M.S degree in computer science and engineering from The Ohio State University, Columbus, OH, USA, in 2017. He is currently working toward the Ph.D degree with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, USA. His research interests are microphone array processing, robust automatic speech recognition, speech enhancement and speaker separation, machine learning, and deep learning.

DeLiang Wang, photograph and biography not available at the time of publication.