Multi-Channel Talker-Independent Speaker Separation Through Location-Based Training

Hassan Taherian, *Graduate Student Member, IEEE*, Ke Tan, and DeLiang Wang, *Fellow, IEEE*

***Abstract*—Permutation ambiguity is a crucial issue for deep learning based talker-independent speaker separation. Deep clustering and permutation invariant training (PIT) have been widely used to address the permutation ambiguity problem in monaural scenarios. Although both approaches have been extended to multi-microphone scenarios, we believe that the permutation ambiguity problem can be naturally avoided by leveraging the spatial relations of multiple speakers. In this article, we present location-based training (LBT), a new approach to achieve talker independency in multi-channel speaker separation. Unlike PIT that examines all possible permutations, LBT assigns speakers according to their positions in physical space. With a linear training complexity to the number of concurrent speakers, LBT is computationally much more efficient than PIT with a factorial complexity, particularly when a large number of overlapping speakers needs to be separated. Specifically, we propose two training criteria: azimuth-based and distance-based training, using speaker azimuths and distances relative to a microphone array, respectively. Evaluation results show that LBT significantly outperforms PIT on two-speaker and three-speaker mixtures with different array geometries and in various acoustic conditions. In addition, we propose a joint training strategy to integrate azimuth-based and distance-based training, which further improves separation performance.**

***Index Terms*—Multi-channel speaker separation, location-based training, permutation invariant training, talker independence.**

I. Introduction

AS A FUNDAMENTAL fundamentalw task in speech processing, speaker separation aims to segregate multiple concurrent speakers. Solutions to speaker separation are important for human speech perception [1], as well as downstream speech processing systems such as speaker recognition, localization, diarization, and automatic speech recognition (ASR). In the past decade, deep learning has been the dominant approach to speaker separation, in which each output layer of a deep neural network (DNN) is associated with one distinct speaker in a multi-talker mixture [2]. Early DNN-based models for monaural separation are trained in a talker-dependent fashion, where the same underlying speakers are used for both training and testing [3], [4], [5]. For many practical applications, however, speaker separation needs to be talker-independent in order to handle untrained speakers. To achieve talker independency, a key challenge is the assignment of DNN output layers to the underlying speakers. Without proper output-speaker assignment, DNN training would not converge due to conflicting gradients. This is known as the permutation ambiguity problem, which has been addressed in the monaural setup in two main approaches: deep clustering [6] and permutation invariant training (PIT) [7]. Deep clustering generates an embedding vector for each time-frequency (T-F) unit using a DNN trained with an objective function invariant to speaker permutations. Using the K-means algorithm, these embeddings are clustered to estimate the ideal binary mask for each speaker. Different from deep clustering, PIT resolves permutation ambiguity by examining the losses from all possible output-speaker assignments without requiring an additional clustering step.

Recent studies have extended deep clustering and PIT to multi-microphone scenarios [8], [9], [10], [11], [12], [13], [14], [15], [16], [17]. Higuchi et al. [8] incorporated deep clustering into masking-based beamforming. In [9], a PIT-based separation model is trained with magnitude spectra and inter-microphone phase difference features. The T-F masks produced by the model are then used to formulate a masking-based beamformer for separating individual speaker signals. Another study [10] investigated masking-based beamforming with a PIT-based separation model trained in the time domain. Chen et al. [11] proposed a multi-channel separation model consisting of multiple DNNs, one for each frequency subband. These DNNs are jointly trained using PIT. In [16], a beam prediction network is used to select the best beam pattern for each speaker from a set of fixed beamformers. A PIT-based separation network is then used to process the output signals of the selected beamformers. Other works combine spectral and spatial information to differentiate signals from different directions. In [12], inter-microphone phase patterns are used as an additional input feature to a deep clustering network. Moreover, end-to-end multi-channel speech separation models with PIT have been developed, where spatial cues are learned directly from the multi-channel mixture in the time domain [14], [15] or frequency domain [17].

Despite these developments, previous multi-channel separation models address the permutation ambiguity by using a training criterion that relies only on spectral information. Leveraging

Manuscript received 8 March 2022; revised 8 July 2022; accepted 18 August 2022. Date of publication 26 August 2022; date of current version 3 September 2022. This work was supported in part by the National Science Foundation under Grants ECCS-1808932 and ECCS-2125074, in part by the Ohio Supercomputer Center and in part by the Pittsburgh Supercomputer Center under Grant NSF ACI-1928147. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiao-Lei Zhang. *(Corresponding author: Hassan Taherian.)*

Hassan Taherian and Ke Tan are with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210-1277 USA (e-mail: taherian.1@osu.edu; tan.650@osu.edu).

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.2022.3202129

spatial information afforded by microphone arrays, we study a new training approach for talker-independent multi-channel speaker separation. With the observation that multiple talkers cannot occupy the same location at the same time, we resolve the permutation ambiguity problem by utilizing spatial relations of multiple speakers. The new training approach, named location-based training (LBT), assigns DNN outputs according to speaker locations in the physical space. Specifically, we introduce two training criteria, i.e. azimuth-based and distance-based training, to label speakers consistently based on their azimuth angles and distances to a microphone array. With LBT, we train a multi-channel input single-channel output (MISO) system, to directly estimate the real and imaginary spectrograms of the speakers from those of a multi-channel mixture [17]. The idea of using speaker azimuths to resolve permutation ambiguity appears to be independently developed for speaker localization in [18] (see a later version in [18]) and for speaker separation in [19] (see a later version in [20]). In a preliminary study, we recently examined the performance of LBT with limited speaker positions for a single array geometry [21]. In the present study, we systematically investigate the new training criteria in different array configurations and acoustic environments. Our evaluation results show that LBT consistently outperforms PIT in both separation quality and automatic speech recognition accuracy.

In addition, we examine the impact of speaker placement on the performance of azimuth-based and distance-based training. Although both criteria outperform PIT, they become less effective in certain conditions. The performance of azimuth-based training significantly degrades when the azimuth differences between speakers become small. Similarly, the performance of distance-based training degrades when the speaker-array distances of different speakers are close. To boost the performance of LBT in such conditions, we introduce two methods to combine the relative advantages of azimuth-based and distance-based training. The first method is to dynamically select between the two criteria based on azimuth estimates of separated speakers, which are derived by speaker localization using mask-weighted generalized cross-correlation with phase transform (GCC-PHAT) [22]. In the second method, the azimuth-based and distance-based models are integrated and jointly trained. The resulting location-based model is more robust and performs significantly better than individual azimuth-based and distance-based models.

The rest of the paper is organized as follows. In Section II, we describe location-based training, the DNN architecture and the joint location model by integrating the azimuth and distance criteria. We then present the experimental setup and the evaluation results in Section III. Concluding remarks are provided in Section IV.

II. Algorithm Description

A. Signal Model and Permutation Invariant Training

A mixture speech signal with N concurrent speakers in a noisy and reverberant environment can be expressed as:

$$\boldsymbol{Y}(t,f) = \sum_{n=1}^{N} [\boldsymbol{S}_n(t,f) + \boldsymbol{H}_n(t,f)] + \boldsymbol{V}(t,f) \qquad (1)$$

where $\boldsymbol{Y}(t,f) = [Y^1(t,f), \ldots, Y^M(t,f)]^T \in \mathbb{C}^M$ denotes the short-time Fourier transform (STFT) vector received by an array with M microphones at time t and frequency f. $\boldsymbol{S}_n(t,f)$ and $\boldsymbol{H}_n(t,f) \in \mathbb{C}^M$ are the STFT vectors of the direct-path signal and reverberation for the n-th speaker, respectively. $\boldsymbol{V}(t,f) \in \mathbb{C}^M$ is the STFT vector of background noise. We assume that all speakers are still within the duration of a single mixture and the same array geometry is used for training and testing, i.e. fixed array geometry.

Given the mixture $\boldsymbol{Y}$, we formulate the estimation of clean speech S_n for speaker n at a reference microphone as a supervised learning problem [2]. For deep learning based speaker separation, DNN outputs should be properly assigned to speakers to avoid the permutation ambiguity problem. For example, the assignment can be made based on speaker identity and speaker gender, which leads to talker-dependent and gender-dependent separation models, respectively. To train talker-independent separation models, utterance-level PIT is widely utilized to address the permutation ambiguity problem [7], [14], [15], [17]. Using fixed output-speaker pairings for a whole utterance, utterance-level PIT selects the optimal permutation that minimizes the loss function from all possible speaker permutations [7]:

$$\mathcal{L}_{\text{PIT}} = \min_{\phi \in \Phi} \sum_{n=1}^{N} \mathcal{L}(\hat{S}_n, S_{\phi(n)}), \qquad (2)$$

where $\hat{S}_n$ is the estimated speech signal of speaker n. $\mathcal{L}$ denotes a loss function, symbol Φ the set of all permutations of N speakers with ϕ referring to one permutation.

B. Location-Based Training

The permutation ambiguity problem can be naturally avoided by exploiting the spatial information of speakers captured by a microphone array. We propose to utilize spatial relations of speakers to determine output-speaker assignments. Specifically, we propose two new training criteria based on speaker azimuth angles and speaker-array distances for multi-channel talker-independent separation. Assuming a polar coordinate system with the center of microphone array as the origin, we define the loss function for azimuth-based training as follows:

$$\mathcal{L}_{\text{Azimuth}} = \sum_{n=1}^{N} \mathcal{L}(\hat{S}_n, S_{\theta_n}), \qquad (3)$$

where $\theta_1, \theta_2, \ldots, \theta_N \in \{1, \ldots, N\}$ are the sorted speaker indices based on their azimuths relative to the microphone array. Fig. 1 illustrates LBT with 3 speakers. In azimuth-based training, output-speaker assignments follow the azimuth order, where the first output is tied to the speaker with the smallest azimuth, the second output to the speaker with the second smallest azimuth, and so on. By assigning speakers consistently based on the order of azimuth angles, such a criterion enforces the DNN to leverage spatial information to address the permutation ambiguity problem. Note that the azimuth range is dependent on the array geometry. Non-linear arrays cover the horizontal plane with the full azimuth range. However, the azimuth range for a linear array should be $[0, \pi)$, due to the well-documented front-back confusion of linear arrays.

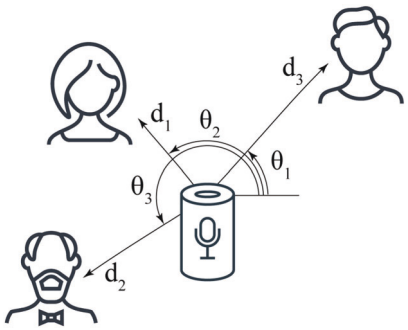


Fig. 1. Illustration of new training criteria based on speaker azimuths and distances relative to a microphone array.

Similarly, DNN outputs can be organized according to speaker-array distances. The loss function for distance-based training is defined as:

$$\mathcal{L}_{\text{Distance}} = \sum_{n=1}^{N} \mathcal{L}(\hat{S}_n, S_{d_n}), \tag{4}$$

where $d_1, d_2, \ldots, d_N \in \{1, \ldots, N\}$ are speaker indices sorted in ascending order based on speaker distances to the microphone array. With this criterion, we assign the nearest speaker to the first output layer, and the second nearest speaker to the second output layer, and so on. This training criterion is also illustrated in Fig. 1. Different from azimuth-based training, distance-based training resolves the permutation ambiguity problem through the consistent pairings of DNN output layers and speaker distances.

Leveraging spatial relations of speakers, both LBT criteria lead to a simple training procedure to achieve talker independence in speaker separation. Compared with PIT, LBT does not need to examine various speaker permutations. Hence LBT has linear training complexity, computationally a lot more efficient than PIT and its variants whose training complexity is factorial or polynomial to the number of speakers [23], [24]. Specifically, the computational complexity of (3) and (4) is $O(N)$ while that of (2) is $O(N!)$. With the lower complexity of location-based training, multi-channel separation models can be efficiently trained to accommodate mixtures with a large number of concurrent speakers. Moreover, LBT-based models produce outputs that are ordered according to speaker spatial locations, facilitating speaker localization. This property is useful for continuous speaker separation where an audio stream is processed in short sliding segments. In such scenarios, PIT-based separation requires a post-processing step to keep the output-speaker assignments consistent between the segments [25], [26]. With LBT, however, separated segments can be naturally organized based on common locations, a prominent grouping principle in human and computational auditory scene analysis [27], [28].

C. Multi-Channel Complex Ratio Masking

With the fixed array geometry assumption, a complex-domain MISO separation system can implicitly learn the spectral and spatial information within array signals [29], which achieves separation performance comparable to or better than masking-based beamforming [20]. In this study, we extend the Dense-UNet architecture proposed in [30] for multi-channel complex ratio masking (MC-CRM). The input to the MC-CRM is a stack of real and imaginary components of the mixture STFT at all microphones. For each speaker, MC-CRM outputs a complex ratio mask cRM_n, which is then multiplied by the mixture STFT at the reference microphone to estimate the separated sources in the complex domain [31]:

$$\hat{S}_n(t,f) = \text{cRM}_n(t,f) \otimes Y^{\text{ref}}(t,f) \tag{5}$$

where symbol $\otimes$ denotes point-wise complex multiplication. In the end, we perform inverse STFT to resynthesize the waveforms of the underlying speakers. The multi-channel Dense-UNet consists of 4 downsampling and 4 upsampling layers, and 9 densely-connected convolutional blocks. In the encoder part of the model, dense blocks and downsampling layers are interleaved to project the input feature map unto a higher level of abstraction. The encoded features are restored to the original resolution with alternated dense blocks and upsampling layers in the decoder part of the model. In addition, skip connections are used to link the dense blocks between the encoder and the decoder at the corresponding level. Each dense block contains 5 convolutional layers, each of which has $C = 64$ channels, a kernel size of 3×3 and a stride of 1×1. The middle layer in each dense block is replaced with a frequency mapping layer to deal with inconsistencies between different frequency bands [30].

We adopt the loss function in [29], which is based on ℓ_1 norm of the difference between the real and imaginary spectrograms of estimated and target speech with an additional magnitude loss term. For each output-speaker pair, the loss function is defined as:

$$\begin{aligned}\mathcal{L}(\hat{S}, S) = &\left\|\hat{S}^{(r)} - S^{(r)}\right\|_1 + \left\|\hat{S}^{(i)} - S^{(i)}\right\|_1 \\ &+ \left\||\hat{S}| - |S|\right\|_1,\end{aligned} \tag{6}$$

where superscripts r and i denote real and imaginary parts, $|S|$ and $|\hat{S}|$ represent the target and estimated magnitude spectrograms, and $|\hat{S}|$ is calculated from the estimated real and imaginary components $\hat{S}^{(r)}$ and $\hat{S}^{(i)}$:

$$|\hat{S}| = \sqrt{(\hat{S}^{(r)})^2 + (\hat{S}^{(i)})^2}. \tag{7}$$

D. Fusing Azimuth-Based and Distance-Based Training

With the proposed azimuth or distance criterion, the separation model learns to discriminate speakers based on a single dimension of the polar coordinate system. It is expected that such separation models will not perform well in conditions where speakers are located at nearby places along that dimension. In these conditions, azimuth and distance (or radius) dimensions can be used together to improve the separation performance.

A simple way to obtain a more general separation model is by utilizing azimuth-based and distance-based criteria simultaneously and selecting the outputs from the better-performing model. We determine the better model on the basis of the estimates of speaker azimuths. The azimuth of speaker k can

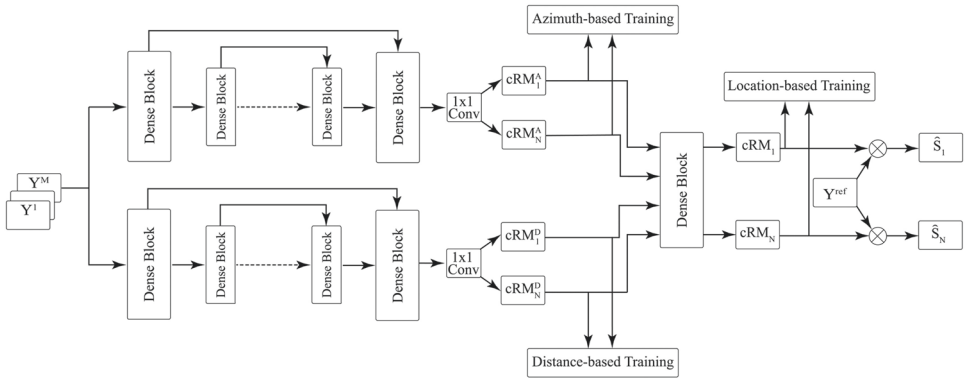


Fig. 2. Schematic diagram of the proposed joint training framework for location-based training. cRM_n^{A}, cRM_n^{D} and cRM_n refer to the estimated complex ratio masks of the n-th speaker by the azimuth branch, the distance branch and the fusion model, respectively.

be well estimated from a speech mixture using mask-weighted GCC-PHAT [22], [32]:

$$\hat{\tau}_k = \underset{\tau}{\text{argmax}} \sum_{(p,q)\in\Omega} \sum_{t,f} \lambda_k \text{GCC}_{p,q}(t,f,\tau), \tag{8}$$

where $\text{GCC}_{p,q}(t,f,\cdot)$ represents the GCC-PHAT function for microphone pair (p,q). Symbol τ denotes the time delay corresponding to a candidate azimuth, and Ω is the set of all microphone pairs. Moreover, λ_k is a ratio mask for speaker k, computed using the outputs of either azimuth- or distance-based models:

$$\lambda_k = \frac{|\hat{S}_k|^2}{|\hat{S}_k|^2 + |Y^{\text{ref}} - \hat{S}_k|^2}. \tag{9}$$

We select the model trained with the azimuth criterion if the estimated azimuth difference is larger than a predefined threshold and the model trained with the distance criterion otherwise. Selecting LBT models using GCC-PHAT is straightforward for two-speaker mixtures. However, applying this method to mixtures with more than two speakers becomes more complicated as more speakers crowd the azimuth dimension.

Another way is to fuse azimuth-based and distance-based models through joint training. Fig. 2 depicts the diagram of the proposed fusion model. The model contains two branches of the identical multi-channel Dense-UNet, each of which takes the stack of real and imaginary spectra of the microphones as input. We train the first branch with the azimuth criterion and the second branch with the distance criterion. The estimated masks from the two branches are concatenated and processed by a fusion dense block to produce the complex ratio masks for speaker separation. The fusion dense block can be regarded as a post-filter that combines and further improves the estimated masks from the azimuth and distance branches. In this study, we use the azimuth criterion to optimize the fusion dense block. The joint model is trained with the following loss function:

$$\mathcal{L}_{\text{Location}} = \sum_{n=1}^{N} \mathcal{L}(\hat{S}_n^{\text{A}}, S_{\theta_n}) + \sum_{n=1}^{N} \mathcal{L}(\hat{S}_n^{\text{D}}, S_{d_n}) + \sum_{n=1}^{N} \mathcal{L}(\hat{S}_n, S_{\theta_n}), \tag{10}$$

where $\hat{S}_n^{\text{A}}$, $\hat{S}_n^{\text{D}}$ and $\hat{S}_n$ are the estimated speech signal of speaker n from the azimuth branch, the distance branch, and the fusion dense block, respectively.

We emphasize that the azimuth-based, distance-based, and fusion models are trained together using the same loss function measured at both hidden and output layers of the separation network (see Fig. 2), resembling the neural cascade architecture [33]. This training strategy is different from commonly used sequential training, where preceding modules are pre-trained and then either fixed or fine-tuned in later training stages. Our training strategy allows the errors in the fusion model to directly influence the optimization of the azimuth-based and distance-based models.

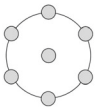
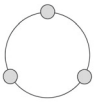
III. Evaluation Results and Comparisons

A. Experimental Setup

For evaluation, we simulate room impulse responses (RIRs) for two microphone array geometries using the image method [34], [35]. The first microphone array has a similar geometry to Amazon Echo, which has 6 microphones uniformly distributed on a circle with a radius of 4.25 cm and one microphone at the center of the circle. The second array has 3 microphones, i.e. the minimum number of microphones to exhibit the full azimuth range $[0, 2\pi)$. Specifically, we use a triangular microphone array where microphones are placed on the circle with a radius of 4.25 cm. We simulate rectangular rooms with random length, width, and height dimensions in the range of [4×4×3, 9×9×4] meters, with the microphone array placed in the center of the room.

The speech sources are placed in positions uniformly sampled from 360 candidate azimuth angles in the range of -180° to 180° with a 1° resolution using a uniform distribution. For a speaker pair (i,j), the source-array distances d_i and d_j are uniformly sampled such that $|d_i - d_j| \geqslant 0.2$ m. Speaker distances are selected in 0.05 m steps. Moreover, the minimum source-array

TABLE I
AVERAGE ESTOI (%), PESQ, SI-SNR (DB) AND SDR (DB) RESULTS OF DIFFERENT TRAINING CRITERIA ON REVERBERANT 2-SPEAKER AND 3-SPEAKER MIXTURES

	Model	#Parameters	Criterion	7-channel Circular Array ESTOI	PESQ	SI-SNR	SDR	3-channel Triangular Array ESTOI	PESQ	SI-SNR	SDR
2-speaker	Unprocessed	–	–	37.36	1.61	-8.15	-1.75	37.35	1.61	-8.15	-1.75
	SC-CRM	4.88M	PIT	62.34	2.20	-0.54	3.48	62.34	2.20	-0.54	3.48
	MC-CRM	4.91M	PIT	74.78	2.74	4.64	7.88	71.27	2.59	3.00	6.89
	MC-CRM	4.91M	Azimuth	80.98	3.03	6.66	9.74	77.65	2.86	4.55	8.12
	MC-CRM	4.91M	Distance	79.75	2.95	6.43	9.13	73.55	2.67	3.52	6.90
	MC-CRM Large	11.15M	PIT	76.17	2.79	5.29	8.44	74.27	2.70	3.65	7.06
	Model Selection with GCC-PHAT	9.82M	–	81.33	3.04	6.76	9.84	77.85	2.86	4.63	8.22
	Joint Model	10.22M	Location	**83.77**	**3.12**	**8.22**	**10.73**	**79.77**	**2.96**	**5.32**	**8.31**
3-speaker	Unprocessed	–	–	29.02	1.38	-8.97	-4.58	29.02	1.38	-8.97	-4.58
	SC-CRM	4.88M	PIT	48.36	1.78	-2.18	1.72	48.36	1.78	-2.18	1.72
	MC-CRM	4.91M	PIT	67.85	2.47	4.97	7.28	57.74	2.10	1.32	4.36
	MC-CRM	4.91M	Azimuth	70.96	2.64	5.55	8.33	63.20	2.33	2.77	6.18
	MC-CRM	4.91M	Distance	66.97	2.42	4.72	6.84	56.45	2.00	1.20	3.87
	Joint Model	11.17M	Location	**74.79**	**2.74**	**7.18**	**9.35**	**68.04**	**2.50**	**4.15**	**6.93**

distance is set to 0.3 m. We assume that speakers are placed at the same height as the microphone array.

We create speech mixtures with 2 and 3 speakers in both anechoic and reverberant conditions. The multi-channel mixtures are created by spatializing the WSJ0-2mix and WSJ0-3mix datasets [6] with the simulated RIRs, which include 20000, 5000, and 3000 mixtures in the training, validation, and test sets, respectively. We adopt the 'min' version of the datasets, where the longer speech signals are truncated to have the same length as the shortest speech signal in a mixture. For the reverberant mixtures, the reverberation time (T60) is randomly sampled between 0.15 and 0.6 seconds. For all speakers, the direct-path (anechoic) signal at the reference microphone is used as the target signal. Note that we treat the center microphone of the 7-channel circular array as the reference microphone. For the 3-channel array, the microphone positions are symmetric and we designate the first microphone as the reference microphone.

All signals are sampled at 16 kHz. For STFT, we use the square root of the Hanning window with a 32 ms frame length and an 8 ms frame shift. We extract 257-dimensional one-sided complex spectra using a 512-point discrete Fourier transform. All separation models are trained on 4-second segments using the Adam optimization algorithm with an initial learning rate of 0.00015. Learning rate adjustment and early stopping are adopted. The model with the lowest validation loss among different epochs is selected for testing. For the joint model, the number of kernels in each convolutional layer of the fusion dense block is $C = 64$ and $C = 128$ for two- and three-speaker mixtures, respectively.

We measure speaker separation performance using perceptual evaluation of speech quality (PESQ) [36], extended short-time objective intelligibility (ESTOI) [37], signal-to-distortion ratio (SDR) [38], and scale-invariant signal-to-noise ratio (SI-SNR) [39]. These are all standard metrics for speaker separation performance evaluation.

B. Results and Comparisons

Table I presents the evaluation results with different training criteria in reverberant conditions. As comparison baselines, we also report the results for PIT-based single-channel complex ratio masking (SC-CRM) and MC-CRM. Not surprisingly, PIT-based MC-CRM performs better than PIT-based SC-CRM (see also [20]). As shown in the table, for two-speaker mixtures, azimuth-based and distance-based training both outperform PIT with the 7-channel microphone array. With regard to the comparison between the two LBT criteria, azimuth-based training performs slightly better than distance-based training. The same performance trend is observed for the 3-channel microphone array. However, the performance gap is larger between azimuth-based and distance-based training.

For three-speaker mixtures, Table I shows that azimuth-based training performs better than PIT in both array geometries. Distance-based training underperforms azimuth-based training, but it yields comparable results to PIT.

Fig. 3 shows the average results for two-speaker reverberant mixtures with the 7-channel array, grouped into intervals based

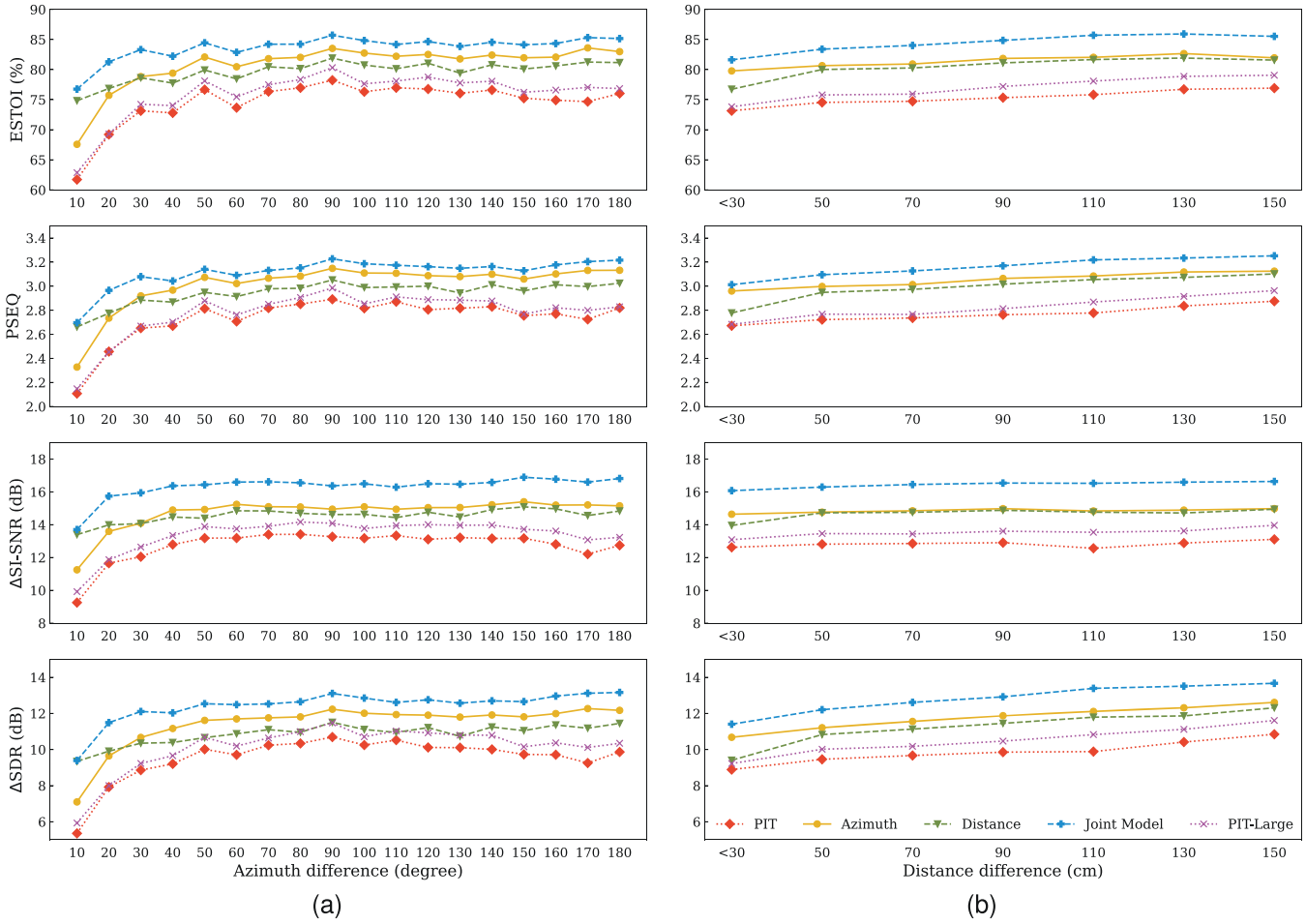


Fig. 3. Average results on reverberant 2-speaker mixtures with the 7-channel array, based on (a) azimuth difference with 10° intervals and (b) distance difference with 20 cm intervals. 'PIT-Large' refers to the PIT-based MC-CRM model with a larger number of parameters.

on speaker azimuth differences (Fig. 3(a)) and distance differences (Fig. 3(b)). From Fig. 3(a), we observe that both PIT and azimuth-based training degrade sharply when azimuth differences between speakers are less than 20°. In contrast, distance-based training is relatively insensitive to speaker azimuths, and thus outperforms the other two methods when azimuth differences are small. This also indicates that the PIT-based MC-CRM model learns to separate speakers by implicitly leveraging spatial information more correlated with azimuth than distance. By the same token, distance-based training results become worse when the difference between the source-array distances of speakers are small, as shown in Fig. 3(b). In addition, we observe that the performance of all separation models consistently improves as the distance differences between speakers increase.

By fusing the models trained with the azimuth and distance criteria, we achieve further improvements as shown in Table I. For model selection with GCC-PHAT, we use an empirical threshold of 20° [21]. With the model selection technique, the improvements over azimuth-based training are marginal, likely because only 12% of mixtures in this test set contain speakers with an azimuth difference less than 20°. However, we observe that the joint model significantly improves the performance for two-speaker mixtures in both array geometries. The joint model is even more effective for three-speaker mixtures, relative to the models trained with the azimuth and distance criteria. We should point out that the reason for the improvements provided by the joint model is not that it is a larger DNN model with more trainable parameters. To demonstrate this, we train a PIT-based MC-CRM model with a larger number of convolutional kernels ($C = 98$), which amounts to a similar number of parameters to the joint model. As shown in Table I, the joint model significantly outperforms the large MC-CRM model trained with PIT, although increasing the model size improves the performance of the PIT-based model to some extent.

Why does the joint model consistently outperform both azimuth-based and distance-based training? Unlike azimuth-based and distance-based training, the joint model leverages both azimuth and distance dimensions to separate the speaker from a particular direction and distance. The joint model improves the robustness of separation in conditions where a single criterion is not discriminative enough. This can be explained with the spherical intersection (SI) method for source localization in three-dimensional (3D) space. With the knowledge of array geometry, SI uses a set of time difference of arrival (TDoA) estimates

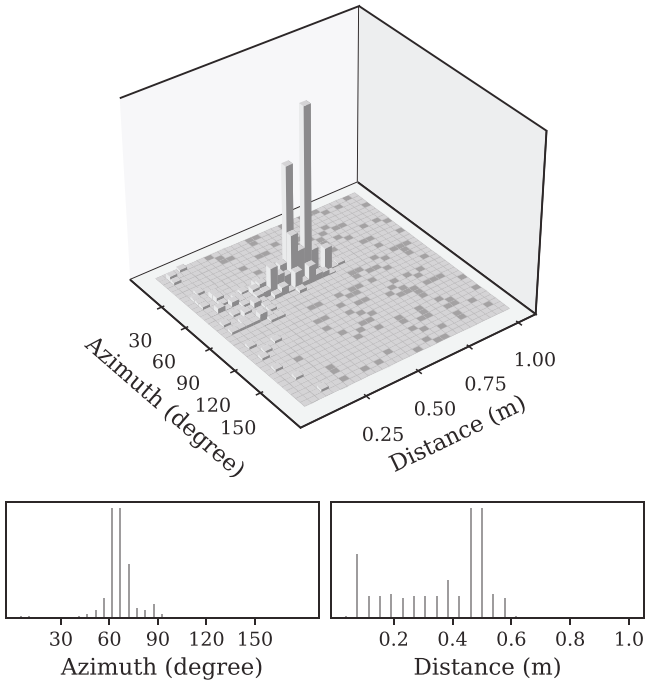


Fig. 4. Histograms of frame-wise estimated azimuths and distances for a 2-speaker mixture in a reverberant room with T60=640 ms. Speaker distances from a 7-channel microphone array are 0.43 m and 0.50 m and azimuths are 60° and 65°.

from microphone pairs to estimate the speaker location [40], [41]. Specifically, SI forms a set of quadratic equations with the pairs of microphone $m = 1, \ldots, M-1$ and the reference microphone:

$$c\tau_m = \|\boldsymbol{x}_m - \boldsymbol{x}_s\|_2 - \|\boldsymbol{x}_s\|_2 , \tag{11}$$

where c is the speed of sound and τ_m is the TDoA between the reference microphone and microphone m. Vectors $\boldsymbol{x}_s$ and $\boldsymbol{x}_m$ denote the speaker s and microphone m locations relative to the reference microphone, respectively. Geometrically, (11) represents a hyperboloid surface, and the speaker location lies on the intersection of all hyperboloids. An estimate of $\boldsymbol{x}_s$ is obtained using unconstrained or constrained least squares [41]. Fig. 4 shows histograms of frame-wise estimated azimuth and distance with SI for two closely-positioned speakers in a reverberant room. The joint azimuth and distance histogram exhibits two distinct peaks, corresponding to the azimuth and distance of the two speakers. However, the individual histograms of estimated azimuths and distances show only one peak occurring in two adjacent bins, which does not indicate two speakers in the mixture. This observation provides an explanation of the effectiveness of combining azimuth and distance information for speaker separation.

Table II shows the evaluation results in the anechoic condition. We see that the pattern of the results in this table is similar to that in Table I. Azimuth-based training achieves superior performance to PIT for two- and three-speaker mixtures with both array geometries. However, the MC-CRM model trained with the distance criterion significantly degrades in the anechoic condition, and its performance is closer to PIT-based SC-CRM than PIT-based MC-CRM. Comparing the results in the reverberant and anechoic conditions, it appears that distance-based training implicitly leverages direct-to-reverberant ratios (DRRs) of different speakers for speaker separation. The DRR is inversely proportional to the square of source-microphone distance in reverberant environments [42], [43]. As the source-microphone distance increases, the energy of the direct sound decreases while the energy of the reverberant sounds remains roughly constant. In the reverberant conditions, the model trained with the distance criterion may learn to assign the speaker with the highest DRR to the first output layer and the second highest DRR to the second output layer, and so on. In an anechoic room, the DRR is infinite and thus cannot serve as a discriminative cue for separating nearer and farther speakers. Note that we do not train a joint model in this case as azimuth-based training is clearly superior to distance-based training and should be employed.

To further investigate the effect of reverberation on the separation performance, we evaluate the models on test sets with different reverberation times (T60). We generate 4 additional test sets with fixed speaker positions, from short to long reverberation times. The evaluation results are reported in Table III. As the reverberation time increases, the performance difference between distance-based and azimuth-based training becomes smaller, demonstrating that reverberation plays an important role in the efficacy of distance-based training.

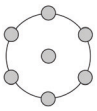
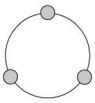
C. Evaluation on Noisy Reverberant Mixtures

In realistic acoustical environments, the speech signal of interest is contaminated by competing speakers and ambient noise simultaneously. The separation problem becomes substantially more challenging as the model needs to perform denoising, dereverberation and speaker separation. An interesting question is whether LBT can address multi-channel speaker separation in the presence of point-source noise, which could interfere with the received spatial patterns of speech signals. To answer this question, we now evaluate the performance of the LBT models trained on noisy reverberant mixtures. We create a spatialized version of the WHAM! dataset [44] using the 7-channel circular array. In the WHAM! dataset, each two-speaker mixture of the WSJ0-2mix dataset is paired with a nonspeech ambient noise, recorded in real environments such as coffee shops, restaurants and bars. Before convolving with RIRs, the speech sources and noise are scaled according to the WHAM! dataset. The same simulation procedure for three-speaker mixtures (see Section III-A) is used to generate two-speaker mixtures with a point-source noise. We randomly select T60 in the range of 0.15 to 0.25 s. The results are shown in Table IV. We observe that both azimuth-based and distance-based training outperform PIT. Moreover, the results suggest LBT generalizes well to noisy reverberant conditions.

D. Sensitivity to Microphone Spacing

This section investigates the sensitivity of models trained with the distance criterion to microphone spacing. We employ three 2-channel linear arrays with different inter-microphone

TABLE II
COMPARISON OF DIFFERENT TRAINING CRITERIA ON 2-SPEAKER AND 3-SPEAKER MIXTURES IN THE ANECHOIC CONDITION

				7-channel Circular Array				3-channel Triangular Array			
	Model	#Parameters	Criterion	ESTOI	PESQ	SI-SNR	SDR	ESTOI	PESQ	SI-SNR	SDR
2-speaker	Unprocessed	–	–	56.13	1.89	-0.01	0.18	56.13	1.89	-0.01	0.18
	SC-CRM	4.88M	PIT	82.93	2.89	11.63	12.06	82.93	2.89	11.63	12.06
	MC-CRM	4.91M	PIT	98.12	4.09	25.80	26.30	98.04	4.06	25.92	26.40
	MC-CRM	4.91M	Azimuth	**98.82**	**4.22**	**27.96**	**28.35**	**98.64**	**4.16**	**27.00**	**27.44**
	MC-CRM	4.91M	Distance	88.02	3.28	14.20	14.62	86.25	3.18	12.68	13.11
3-speaker	Unprocessed	–	–	38.55	1.48	-4.53	-4.19	38.55	1.48	-4.53	-4.19
	SC-CRM	4.88M	PIT	62.50	2.10	4.82	5.57	62.50	2.10	4.82	5.57
	MC-CRM	4.91M	PIT	84.18	3.17	13.53	14.14	82.43	3.05	12.98	13.53
	MC-CRM	4.91M	Azimuth	**91.15**	**3.57**	**17.86**	**18.36**	**88.10**	**3.36**	**15.76**	**16.28**
	MC-CRM	4.91M	Distance	65.39	2.19	5.55	6.24	64.92	2.24	5.61	6.32

TABLE III
EVALUATION OF DIFFERENT TRAINING CRITERIA ON 2-SPEAKER MIXTURES WITH THE 7-CHANNEL ARRAY AT DIFFERENT REVERBERATION TIMES

T60	Model	Criterion	ESTOI	PESQ	SI-SNR	SDR
160 ms	Unprocessed	–	53.30	1.86	-1.64	0.14
	MC-CRM	PIT	86.81	3.31	10.97	13.78
	MC-CRM	Azimuth	**91.00**	**3.56**	**13.06**	**16.13**
	MC-CRM	Distance	87.96	3.35	11.24	14.13
360 ms	Unprocessed	–	40.34	1.64	-6.67	-1.29
	MC-CRM	PIT	77.81	2.85	6.13	9.15
	MC-CRM	Azimuth	**83.39**	**3.11**	**8.08**	**10.97**
	MC-CRM	Distance	82.18	3.05	7.75	10.45
610 ms	Unprocessed	–	30.62	1.48	-9.39	-3.21
	MC-CRM	PIT	68.41	2.50	3.51	6.33
	MC-CRM	Azimuth	**75.42**	**2.77**	**5.51**	**8.16**
	MC-CRM	Distance	74.44	2.72	5.30	7.82
900 ms	Unprocessed	–	23.66	1.40	-11.34	-4.85
	MC-CRM	PIT	60.19	2.24	1.60	4.37
	MC-CRM	Azimuth	**67.81**	**2.51**	**3.63**	**6.17**
	MC-CRM	Distance	67.11	2.46	3.44	5.9

TABLE IV
COMPARISON OF DIFFERENT TRAINING CRITERIA ON 2-SPEAKER MIXTURES WITH THE 7-CHANNEL ARRAY IN NOISY REVERBERANT CONDITION

	Criterion	ESTOI	PESQ	SI-SNR	SDR
Unprocessed	–	31.48	1.52	-7.36	-4.91
MC-CRM	PIT	68.47	2.49	5.99	7.83
MC-CRM	Azimuth	**73.94**	**2.80**	**7.52**	**10.28**
MC-CRM	Distance	70.11	2.55	6.59	8.39

TABLE V
EVALUATION OF DISTANCE-BASED TRAINING ON 2-CHANNEL LINEAR ARRAY WITH DIFFERENT INTER-MICROPHONE DISTANCES

Inter-microphone Distance	ESTOI	PESQ	SI-SNR	SDR
Unprocessed	37.45	1.61	-8.12	-1.72
4.25 cm	67.38	2.41	1.34	5.22
8 cm	68.26	2.44	1.71	5.57
24 cm	**71.06**	**2.58**	**2.31**	**6.30**

distances, i.e. 4.25 cm, 8 cm and 24 cm. The same simulation procedure for generating reverberant two-speaker mixtures is used, except that the azimuth range of speakers is limited to $[0, \pi)$. The left-sided microphone is treated as the reference microphone. Separation results are shown in Table V, which suggests that the performance of distance-based training improves as the inter-microphone distance increases. This effect of microphone spacing could be explained by the coherence of microphone signals, which is widely used for DRR estimation [43]. The magnitude of the coherence indicates the strength of correlation between the signals received by a microphone pair. With larger inter-microphone distances, the inter-channel coherence decreases especially for high frequencies in an isotropic sound field [45]. This would also produce multi-microphone signals with more diverse distances and DRR cues, improving the performance of distance-based training.

E. Evaluation on SMS-WSJ Dataset

In this section, we further evaluate LBT on the SMS-WSJ dataset [46], which is a speaker separation and ASR task. This dataset includes reverberant two-speaker mixtures with a sampling rate of 8 kHz. The numbers of training, validation and testing mixtures in this datast are 33561, 982, and 1332, respectively. A circular array geometry with a radius of 10 cm is used for RIR simulation with T60 in the range of 0.2 to

TABLE VI
SPEAKER SEPARATION AND WER (IN %) RESULTS OF COMPARISON SYSTEMS EVALUATED WITH THE 6-CHANNEL ARRAY ON SMS-WSJ

	Criterion	ESTOI	PESQ	SI-SNR	SDR	WER (%)
Unprocessed	–	44.07	1.50	-5.46	-0.38	78.42
MC-CSM	PIT	88.03	3.12	11.25	13.10	11.83
MC-CSM	Azimuth	90.12	3.29	12.56	14.33	9.99
MC-CSM	Distance	88.83	3.15	11.63	13.44	10.53
Joint model (MC-CSM)	Location	90.96	3.33	13.22	14.82	9.62
cACGMM [47]	–	–	–	–	–	39.00
cACGMM with MVDR [47]	–	–	–	–	–	18.70
FaSNet + TAC + joint + 4ms [48]	PIT	77.10	2.37	8.60	–	29.80
Multi-channel Conv-TasNet [49]	PIT	84.40	2.78	10.80	–	23.05
MISO_1 [17]	PIT	86.20	3.06	10.20	–	13.92

0.5 s. The speaker azimuth angles and source-array distances are uniformly sampled in the range of [-180°, 180°] and [1.0, 2.0] m, respectively. An artificially generated white noise is added to the mixtures to simulate sensor noise. The default backend acoustic model associated with the SMS-WSJ dataset is used for ASR evaluation. As opposed to the official SMS-WSJ setup, we simultaneously perform dereverberation and separation by using direct sound as the training target.

Following [17], we modify the Dense-UNet architecture to perform multi-channel complex spectral mapping (MC-CSM) which has been shown to achieve better ASR performance. For the joint model, we use estimated real and imaginary components from the two branches of the fusion dense block to directly produce estimated speech signals. The number of convolutional kernels is set to $C = 76$. We also include the spectral magnitude of the first microphone as an additional input feature.

Table VI compares LBT models and other competitive talker-independent multi-channel speaker separation methods on SMS-WSJ in separation metrics as well as word error rate (WER). For all methods, we list the best reported results, and leave unreported fields blank. The complex angular central GMM (cACGMM) with or without minimum variance distortionless response (MVDR) beamforming corresponds to spatial clusting methods provided as the baselines of SMS-WSJ. The multi-channel Conv-TasNet [15] and multi-channel FaSNet with TAC modules [47] are time-domain end-to-end separation systems. We also compare our methods with a strong MISO_1 [17] system. We can see the same trend with the SMS-WSJ dataset that LBT models outperform PIT in both separation quality and ASR accuracy. Our LBT models obtain substantially better results than FaSNet with TAC modules, multichannel Conv-TasNet, and MISO_1.

IV. CONCLUDING REMARKS

In this study, we have proposed two novel training criteria to address the permutation ambiguity problem for multi-channel talker-independent speaker separation. Different from widely-used PIT, the new criteria organize DNN outputs on the basis of speaker azimuths and distances relative to a microphone array. Using MC-CRM, we have investigated the performance of LBT with different array geometries in various acoustic conditions. Our experimental results demonstrate that LBT yields significantly better separation performance than PIT on two- and three-speaker mixtures. Evaluation results show that the presence of room reverberation is essential for distance-based training. We also develop a joint training strategy to fuse azimuth-based and distance-based training. The joint model produces superior separation performance compared to the models trained with either of the criteria alone. In addition, we show that LBT outperforms PIT for multi-channel speaker separation in the presence of a point-source noise.

Location-based training is applicable to arbitrary array geometries, as long as the same geometry is used in training and testing. The assumption of a fixed geometry is not very constraining, and it is satisfied in most of array-based real-world applications such as Amazon Echo, Google Home, and hearing aids. Another spatial dimension that can be considered in future work is speaker elevation relative to a sensor array, which can be analogously leveraged for output-speaker assignment. In this scenario, a 3D location-based model can be trained by fusing the azimuth, distance, and elevation criteria to further improve speaker separation performance.

When multiple microphones are available, the proposed training criteria can be applied to other tasks susceptible to the permutation ambiguity, such as speaker-independent multi-pitch tracking [48], DNN-based speaker diarization [49] and multi-source speaker localization [18]. In summary, location-based training leverages distinct spatial locations of multiple speakers that exist naturally in physical space, and produces superior talker-independent speaker separation results free of permutation ambiguity.

ACKNOWLEDGMENT

The authors would like to thank Dr. Zhong-Qiu Wang for helpful discussions.

REFERENCES

[1] E. W. Healy, H. Taherian, E. M. Johnson, and D. L. Wang, "A causal and talker-independent speaker separation/dereverberation deep learning algorithm: Cost associated with conversion to real-time capable operation," *J. Acoustical Soc. Amer.*, vol. 150, pp. 3976–3986, 2021.

[2] 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.

[3] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 1562–1566.

[4] J. Du, Y. Tu, Y. Xu, L. Dai, and C.-H. Lee, "Speech separation of a target speaker based on deep neural networks," in *Proc. IEEE 12th Int. Conf. Signal Process.*, 2014, pp. 473–477.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

[10] T. Ochiai, M. Delcroix, R. Ikeshita, K. Kinoshita, T. Nakatani, and S. Araki, "Beam-tasnet: Time-domain audio separation network meets frequency-domain beamformer," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6384–6388.

[11] L. Chen, M. Yu, D. Su, and D. Yu, "Multi-band PIT and model integration for improved multi-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 705–709.

[12] 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.

[13] 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. 2019.

[14] R. Gu et al., "Enhancing end-to-end multi-channel speech separation via spatial feature learning," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 7319–7323.

[15] J. Zhang, C. Zorilă, R. Doddipatla, and J. Barker, "On end-to-end multi-channel time domain speech separation in reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6389–6393.

[16] Z. Chen, T. Yoshioka, X. Xiao, L. 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.

[17] 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, 2021.

[18] A. S. Subramanian, C. Weng, S. Watanabe, M. Yu, and D. Yu, "Deep learning based multi-source localization with source splitting and its effectiveness in multi-talker speech recognition," *Comput. Speech Lang.*, vol. 75, 2022, Art. no. 101360.

[19] K. Tan, "Convolutional and recurrent neural networks for real-time speech separation in the complex domain," Ph.D. dissertation, Ohio State Univ. Dept. Comput. Sci. Eng., 2021.

[20] K. Tan, Z.-Q. Wang, and D. L. Wang, "Neural spectrospatial filtering," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 605–621, 2022.

[21] H. Taherian, K. Tan, and D. L. Wang, "Location-based training for multi-channel talker-independent speaker separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 696–700.

[22] Z.-Q. Wang, X. Zhang, and D. L. Wang, "Robust speaker localization guided by deep learning-based time-frequency masking," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 178–188, Jan. 2019.

[23] S. Dovrat, E. Nachmani, and L. Wolf, "Many-speakers single channel speech separation with optimal permutation training," in *Proc. Interspeech*, 2021, pp. 3890–3894.

[24] H. Tachibana, "Towards listening to 10 people simultaneously: An efficient permutation invariant training of audio source separation using sinkhorn's algorithm," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 491–495.

[25] T. Yoshioka et al., "Advances in online audio-visual meeting transcription," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*., 2019, pp. 276–283.

[26] H. Taherian and D. L. Wang, "Time-domain loss modulation based on overlap ratio for monaural conversational speaker separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 5744–5748.

[27] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. London, U.K.: MIT Press, 1994.

[28] D. L. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.

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

[30] Y. Liu and D. L. Wang, "Divide and conquer: A deep CASA approach to talker-independent monaural speaker separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2092–2102, Dec. 2019.

[31] 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.

[32] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[33] H. Wang and D. L. Wang, "Neural cascade architecture with triple-domain loss for speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 734–743, 2022.

[34] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoustical Soc. Amer.*, vol. 65, pp. 943–950, 1979.

[35] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 351–355.

[36] *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for end-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs*, Rec. ITU-T P. 862, International Telecommunication Union Radiocommunication, 2001.

[37] 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.

[38] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[39] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[40] H. Schau and A. Robinson, "Passive source localization employing intersecting spherical surfaces from time-of-arrival differences," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. 35, no. 8, pp. 1223–1225, Aug. 1987.

[41] P. Stoica and J. Li, "Source localization from range-difference measurements," *IEEE Signal Process. Mag.*, vol. 23, no. 6, pp. 63–66, Nov. 2006.

[42] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA, USA: MIT Press, 1997.

[43] M. Zohourian and R. Martin, "Binaural direct-to-reverberant energy ratio and speaker distance estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 92–104, 2020.

[44] G. Wichern et al., "WHAM!: Extending speech separation to noisy environments," in *Proc. Interspeech*, 2019, pp. 1368–1372.

[45] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. Hoboken, NJ, USA: Wiley, 2006.

[46] L. Drude, J. Heitkaemper, C. Boeddeker, and R. Haeb-Umbach, "SMS-WSJ: Database, performance measures, and baseline recipe for multi-channel source separation and recognition," 2019, *arXiv:1910.13934*.

[47] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6394–6398.

[48] Y. Liu and D. L. Wang, "Permutation invariant training for speaker-independent multi-pitch tracking," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5594–5598.

[49] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with permutation-free objectives," in *Proc. Interspeech*, 2019, pp. 4300–4304.