




Robust frame-level speaker localization guided by multi-channel speech enhancement and inter-channel phase-difference losses

Shanmukha Srinivas Battula,^{1,a)} Hassan Taherian,^{1,b)}  Ashutosh Pandey,² Daniel Wong,² Buye Xu,²  and DeLiang Wang^{1,3,c)} 

¹Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio 43210, USA

²Meta Reality Labs, Redmond, Washington 98052, USA

³Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, Ohio 43210, USA

ABSTRACT:

In the presence of room reverberation and background noise, the performance of frame-level speaker localization is severely limited. To address this challenge, this study performs multi-channel speech enhancement based on complex spectral mapping (CSM), followed by direction-of-arrival (DOA) estimation using weighted generalized cross-correlation with phase transform (GCC-PHAT). The proposed approach differs from prevailing deep learning methods that operate on multi-channel inputs directly for speaker localization. This study initially investigates multi-input single-output (MISO) based speech enhancement models with two loss functions and then extends to multi-input multi-output (MIMO) based modeling for conceptual and computational efficiency. The results demonstrate that the phase estimates obtained from CSM models are reliable for frame-level DOA estimation and MIMO systems outperform MISO systems. In addition, the study proposes new multi-channel loss functions for MIMO systems that incorporate phase differences in order to better preserve inter-channel phase relations, which is key to accurate sound localization. Systematic evaluations with multiple microphone array geometries using both simulated and recorded room impulse responses, as well as real recordings, demonstrate that the proposed model yields excellent frame-level speaker localization results in reverberant and noisy environments and outperforms related methods by a large margin, even surpassing their utterance-level results. © 2025 Acoustical Society of America.

<https://doi.org/10.1121/10.0038627>

(Received 18 February 2025; revised 10 July 2025; accepted 10 July 2025; published online 28 July 2025)

[Editor: Siu-Kit Lau]

Pages: 790–800

I. INTRODUCTION

Robust speaker localization aims to find the spatial location of a target source relative to a microphone array. Localizing sound sources in an acoustic environment has numerous applications, including auditory scene analysis, teleconferencing, immersive virtual or augmented reality, and voice-activated human-machine interaction. Direction-of-arrival (DOA) information is essential for beamforming, widely used in signal processing or deep learning based systems for speech enhancement (Gannot *et al.*, 2017; Vincent *et al.*, 2018). DOA estimation relies on spatial features such as inter-channel time differences, inter-channel phase differences (IPDs), inter-channel level differences, and relative transfer functions (RTFs). Standard DOA methods include generalized cross correlation with phase transform (GCC-PHAT) (Knapp and Carter, 1976), steered response power with phase transform (SRP-PHAT) (DiBiase *et al.*, 2001), and multiple signal classification (MUSIC) (Schmidt, 1986). For example, Kabzinski and Habets proposed a weighted least-squares optimization solution for DOA estimation by minimizing the difference between observed and expected

IPDs (Kabzinski and Habets, 2019). However, the DOA performance of these algorithms severely degrades in reverberant and noisy environments.

Recent studies on robust speaker localization employ a deep neural network (DNN) to directly estimate DOA through supervised training (Grumiaux *et al.*, 2022). Typically, these methods use spatial features, GCC-PHAT, or input spectrograms and employ convolutional neural networks (CNNs), convolutional recurrent networks (CRNs), or attention based networks as model architectures. Chakrabarty and Habets train a CNN using phase spectrograms as input to generate posterior probabilities for each DOA class on white noise signals (Chakrabarty and Habets, 2017). Subsequent studies demonstrate the benefits of training on speech signals and incorporating recurrent connections (Vargas *et al.*, 2021; Bohlender *et al.*, 2021). Some studies utilize IPD (Pang *et al.*, 2019; Shimada *et al.*, 2021; Phokhinanan *et al.*, 2023), RTF (Baek *et al.*, 2023), gammatone filter bank (Goli and van de Par, 2023), and GCC-PHAT responses (Varzandeh *et al.*, 2024) as inputs to estimate the DOA probabilities. Other studies employ a DNN to estimate spatial features, such as direct-path RTF (Yang *et al.*, 2021) and IPD (Pak and Shin, 2019; Wang *et al.*, 2024).

An alternative approach first enhances reverberant and noisy signals and then uses the enhanced signals for DOA

^{a)}Email: battula.12@osu.edu

^{b)}Email: taherian.1@osu.edu

^{c)}Email: dwang@cse.ohio-state.edu

estimation. It is worth noting that psychoacoustic evidence suggests that localization depends on source separation, rather than preceding separation (Hartmann, 1999; Darwin, 2008). Typically, these two-stage methods employ a DNN to estimate a real-valued mask, which is then used in either traditional signal processing or DNN based DOA estimation (Pertilä and Cakir, 2017; Wang *et al.*, 2019; Zhang *et al.*, 2019; Mack *et al.*, 2022). Wang *et al.* (2019) utilize a bi-directional recurrent network to estimate monaural oracle time-frequency (T-F) masks, such as the ideal ratio mask (IRM) or phase sensitive mask (PSM), for speech enhancement. These estimated masks serve as weights to amplify the T-F units dominated by the target source in subsequent GCC-PHAT based localization. In Mack *et al.* (2022), an estimated real-valued mask from the first DNN is utilized within a CNN as a weighting mechanism, referred to as feature masking, for estimating DOA. However, these algorithms rely on the phase of the mixture signals for localization, which limits their frame-level DOA performance.

Recent speech enhancement methods use complex spectral mapping (CSM) (Wang and Wang, 2020; Tan *et al.*, 2022), which can estimate clean phase, a capability not shared by mask-based methods operating in the magnitude domain. Utilizing clean phase estimates overcomes a main limitation in two-stage approaches that rely on noisy phases for robust localization. This enables combining the phase estimation capability of strong speech enhancement to be leveraged by DOA algorithms to achieve robustness. Furthermore, this approach potentially provides more interpretability for robust frame-level localization in terms of delineating enhancement and localization errors.

Most of the above methods perform DOA estimation at the utterance level. Compared to utterance-level speaker localization, frame-level DOA estimation is required in moving source scenarios. In addition, utterance-level DOA is not conducive to real-time applications, which need frame-level direction estimates. However, frame-level localization presents greater challenges due to the lack of temporal pooling, which is a proven technique to improve localization accuracy at the utterance level. Therefore, accurately estimating and weighting IPDs is crucial for achieving robust frame-level accuracy.

In this context, we propose a two-stage approach for robust frame-level speaker localization, where the first stage employs multi-channel CSM for speech enhancement and the second stage performs weighted generalized cross correlation with phase transform (WGCC) for localization. Initially, we train multi-input single-output (MISO) based CSM systems with two loss functions for the first stage. Our results show that the phase estimates from the trained MISO-based CSM models are reliable and significantly improve the DOA estimation of WGCC compared to noisy phases. To further improve the efficiency of training and inference, we propose to replace MISO-based CSM by multi-input multi-output (MIMO) based CSM (Wang and Wang, 2020; Taherian *et al.*, 2023). We demonstrate that

the phase estimates from the trained MIMO-based CSM models are more reliable and achieve superior DOA performance compared to MISO systems. Additionally, MIMO modeling allows us to explore inter-channel relations in training losses, which is a distinct advantage over MISO modeling.

Prior research on DNN-based localization, as reviewed by Grumiaux *et al.* (2022), has incorporated IPDs as input features or attempted to predict IPDs as the output (Pang *et al.*, 2019; Pak and Shin, 2019; Shimada *et al.*, 2021; Wang *et al.*, 2024). Other studies have incorporated an IPD loss term to preserve inter-channel relations for speech enhancement (Wang and Wang, 2020; Hwang *et al.*, 2022; Thaleiser and Enzner, 2023; Olivan *et al.*, 2024; Tokala *et al.*, 2024). Motivated by the importance of phase differences and different from the previous methods of utilizing IPDs for localization, we propose to train MIMO based speech enhancement with a novel IPD term in loss functions for accurate localization. The inclusion of the IPD term ensures that inter-channel phase relations are maintained in the enhanced speech and is found to clearly improve the localization performance across microphone arrays and reverberant-noisy conditions in both simulated and real environments.

A preliminary conference version has been presented (Battula *et al.*, 2025), and it performs frame-level speaker localization using a two-microphone array. The present study goes far beyond the preliminary study by systematically investigating frame-level localization and phase-difference losses using eight-microphone linear and seven-microphone circular arrays in reverberant and noisy acoustic environments. The eight-microphone linear array allows us to investigate the proposed approach on a larger linear array, and the circular array, free from the front-and-back ambiguity seen in linear arrays, enables the evaluation across the full DOA range. This paper also examines MISO modeling and demonstrates that MIMO outperforms MISO consistently. Additionally, we examine the impacts of source-to-microphone distance, reverberation time, and T-F weighting in WGCC on frame-level localization performance. Finally, the present study demonstrates the generalization of proposed speaker localization to real recordings by evaluating on the LOCATA evaluation dataset (Löllmann *et al.*, 2018).

The rest of the paper is organized as follows. Section II presents the problem formulation for DOA, CSM based speech enhancement and training objectives, and WGCC. Section III describes the experimental setup. Section IV presents evaluation results and comparisons, followed by a conclusion in Sec. V.

II. SYSTEM DESCRIPTION

The signal received by a P -channel microphone array in a reverberant and noisy environment can be modeled as

$$\mathbf{Y}(t, f) = \mathbf{S}(t, f) + \mathbf{N}(t, f), \quad (1)$$

where $\mathbf{S} \in \mathbb{C}^{P \times 1}$ denotes the short-time Fourier transform (STFT) of direct-path speech signal and $\mathbf{N} \in \mathbb{C}^{P \times 1}$ denotes

the STFT of all non-target signals, including speech reverberation and reverberant noise. Symbols t and f denote the time frame and frequency bin, respectively. Given the reverberant and noisy microphone array observation \mathbf{Y} , our goal is to first estimate the direct-path speech \mathbf{S} , i.e., $\hat{\mathbf{S}} = \mathcal{F}(\mathbf{Y})$, where \mathcal{F} denotes the estimator, and subsequently its direction θ .

A. Multi-channel CSM

The standard formulation of multi-channel CSM is to estimate the complex spectrogram of the target speech at a reference microphone m , $\hat{S}_m = \mathcal{F}(\mathbf{Y}; \Theta)$, where Θ denotes the set of all model parameters, from a noisy mixture (Wang and Wang, 2020). Specifically, the mixture's real and imaginary spectrograms at all microphones are concatenated and passed into a DNN as input and the DNN outputs an estimate of the target's real and imaginary spectrograms at the reference microphone. To estimate the target source at all microphones for second-stage localization requires designating every microphone as the reference microphone, leading to P enhancement modules, as illustrated in Fig. 1(a). To simplify training and inference, the DNN can be extended to simultaneously estimate the target's real and imaginary spectrograms at all the microphones, corresponding to the MIMO system, as shown in Fig. 1(b). In this study, we propose to use the large version of SpatialNet (Quan and Li, 2024) specifically designed to perform multi-channel CSM for speech separation. We adapt SpatialNet to a MIMO system by adding as many final linear layers as the number of microphones to estimate multiple target outputs, i.e., $\hat{\mathbf{S}} = \mathcal{F}(\mathbf{Y}; \Theta)$. As illustrated in Fig. 1, after first-stage speech enhancement, speaker localization is performed for a P -channel setup as the second stage.

B. Training objectives

For the MISO setup, following earlier studies (Williamson *et al.*, 2016; Fu *et al.*, 2017; Tan and Wang, 2020; Quan and Li, 2024), the loss function is calculated by comparing the real and imaginary spectrograms of the separated (\hat{S}_p) and target speech (S_p) at reference microphone m :

$$\mathcal{L}_{\text{RI+Mag}} = \mathcal{L}_{\text{RI}} + \mathcal{L}_{\text{Mag}}, \quad (2)$$

$$\mathcal{L}_{\text{RI}} = \|\hat{S}_m^{(r)} - S_m^{(r)}\|_1 + \|\hat{S}_m^{(i)} - S_m^{(i)}\|_1, \quad (3)$$

$$\mathcal{L}_{\text{Mag}} = \|\|\hat{S}_m\| - \|S_m\|\|_1, \quad (4)$$

where superscripts (r) and (i) denote real and imaginary parts, respectively, and $\|\cdot\|_1$ the l_1 norm. From now on, we refer to the monaural output loss functions as $\mathcal{L}_{\text{RI}}^{\text{MISO}}$ and $\mathcal{L}_{\text{RI+Mag}}^{\text{MISO}}$.

For the MIMO setup with multiple outputs, the loss functions are calculated by averaging the individual loss functions for each output across all microphones. The corresponding MIMO loss functions are

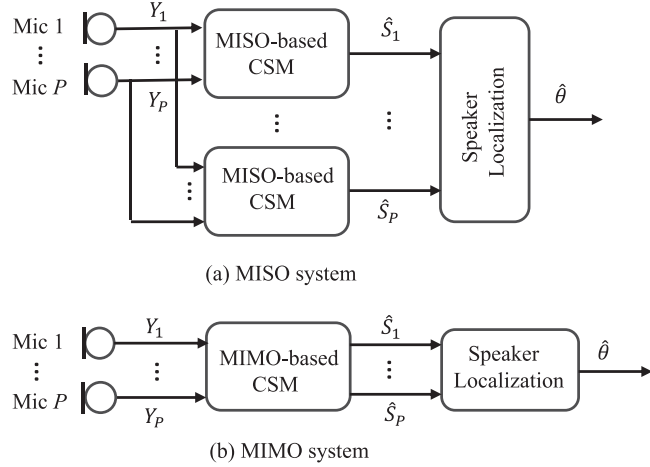


FIG. 1. Block diagram of the proposed system with multi-channel CSM for speech enhancement followed by speaker localization. (a) A MISO system with P modules. (b) A MIMO system for speech enhancement. *Mic*, microphone.

$$\mathcal{L}_{\text{RI}} = \frac{1}{P} \sum_{p=1}^P \mathcal{L}_{\text{RI}}(p), \quad (5)$$

$$\mathcal{L}_{\text{RI+Mag}} = \frac{1}{P} \sum_{p=1}^P \mathcal{L}_{\text{RI+Mag}}(p), \quad (6)$$

where p indicates a microphone.

Although training with the aforementioned loss functions implicitly yields phase estimates, guiding speech enhancement strategically can potentially improve localization accuracy. This is conceptually akin to guiding speech enhancement to improve automatic speech recognition with a magnitude term in the loss function (Wang and Wang, 2020). The MIMO modeling allows us to enforce IPDs between the estimated and target spectrograms via a loss function. To mitigate phase wrapping problems (Kabzinski and Habets, 2019), the loss function for a microphone pair considers the distance between the estimated and target IPDs in a complex plane. We define:

$$\begin{aligned} \mathcal{L}_{\text{IPD}} &= \frac{1}{N_{\Omega}} \sum_{(p,q) \in \Omega} W_p W_q \|\exp(j\hat{S}_{p,q}) - \exp(jS_{p,q})\|_1 \\ &= \frac{1}{N_{\Omega}} \sum_{(p,q) \in \Omega} W_p W_q (\|\cos \hat{S}_{p,q} - \cos S_{p,q}\|_1 \\ &\quad + \|\sin \hat{S}_{p,q} - \sin S_{p,q}\|_1), \end{aligned} \quad (7)$$

where j denotes the imaginary unit and $\hat{S}_{p,q}, S_{p,q}$ represent the IPD between a microphone pair (p, q) of the estimated and target speech, respectively: $S_{p,q} = \angle S_p - \angle S_q$. Ω denotes the set of all possible microphone pairs with cardinality $N_{\Omega} = P(P-1)/2$. The weights W_p and W_q correspond to the target speech IRM at the respective microphones. This weighting mechanism serves to accentuate the T-F units where the target speech dominates (Wang *et al.*, 2019).

Incorporating \mathcal{L}_{IPD} imposes a penalty on IPD estimation errors, reflecting the primary role of IPDs in sound

localization. Considering different ways of including an IPD term, we investigate the following two reasonable alternatives:

$$\mathcal{L}_{\text{RI+IPD}} = \mathcal{L}_{\text{RI}} + \mathcal{L}_{\text{IPD}}, \quad (8)$$

$$\mathcal{L}_{\text{RI+Mag+IPD}} = \mathcal{L}_{\text{RI+Mag}} + \mathcal{L}_{\text{IPD}}. \quad (9)$$

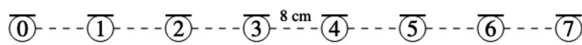
C. WGCC for localization

For microphone pair p and q , the GCC-PHAT method (Knapp and Carter, 1976) computes their generalized cross correlation coefficients with a weighting mechanism based on phase transform to estimate the time delay,

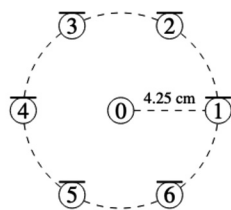
$$GCC_{p,q}(t, f, \theta) = \mathcal{R} \left\{ \frac{Y_p(t, f) Y_q(t, f)^H}{|Y_p(t, f)| |Y_q(t, f)|} e^{-j2\pi(f/N)\tau_{p,q}(\theta)} \right\}, \quad (10)$$

where $\mathcal{R}\{\cdot\}$ extracts the real part, $(\cdot)^H$ represents conjugate transpose, $|\cdot|$ computes the magnitude, N is the number of discrete Fourier transform frequencies, and f_s is the sampling rate in Hz. Term $\tau_{p,q}(\theta) = (d_{\theta q} - d_{\theta p})/c_s$ denotes the time delay of a candidate direction or location θ , where c_s is the speed of sound in the air and $d_{\theta q}$ and $d_{\theta p}$ represent the distance between the hypothesized sound source and microphones p and q , respectively.

The PHAT weighting mechanism, i.e., magnitude normalization in Eq. (10), ensures that each T-F unit has an equal contribution and prevents the dominance of large magnitude T-F units. In Wang *et al.* (2019), mask-weighted generalized cross correlation with phase transform (MGCC) was introduced to emphasize the T-F units dominated by target speech over the T-F units contaminated by noise and reverberation. This method first calculates the GCC coefficients using the input noisy spectrogram and then applies a weighting mask to them. However, our approach relies on complex spectral mapping and the GCC function in our study is calculated directly from the enhanced speech signals (see Fig. 1). For this critical distinction, we introduce WGCC to improve the performance of GCC-PHAT in the presence of reverberation and noise,



(a) 8-microphone linear array



(b) 7-microphone circular array

FIG. 2. Illustration of two microphone arrays.

$$WGCC_{p,q}(t, f, \theta) = W_p(t, f) W_q(t, f) GCC_{p,q}(t, f, \theta), \quad (11)$$

where the weights W_p and W_q are estimated IRMs computed from the estimated target complex spectrograms at the corresponding microphones.

Finally, frame-level DOA is estimated as

$$\hat{\theta}_t = \arg \max_{\theta} \sum_{(p,q) \in \Omega} \sum_{f=1}^{N/2} WGCC_{p,q}(t, f, \theta). \quad (12)$$

III. EXPERIMENTAL SETUP

We use the LibriSpeech corpus (Panayotov *et al.*, 2015) for target speech signals, the DEMAND dataset (Thiemann *et al.*, 2013) for point-source noise simulations, and the TIMIT corpus (Garofolo *et al.*, 1993) to create diffuse noises. Specifically, diffuse noise signals are produced by convolving 72 randomly chosen TIMIT utterances corresponding to 72 speaker positions at 5° intervals from θ to $\theta + 355^\circ$, where θ denotes the angle of the target source, as described in Zhang and Wang (2017) and Tan *et al.* (2022). LibriSpeech train-clean-100, dev-clean, and test-clean utterances are used for training, validation, and testing, respectively. We use the street noise category from the DEMAND dataset for validation and testing and the remaining categories for training. We use the TIMIT training utterances to create diffuse noises for training and validation, and the TIMIT test utterances for testing. Reverberation time (T60) and signal-to-noise ratio (SNR) are randomly sampled between $[0, 1]$ s and $[-5, 5]$ dB, respectively. The length, width, and height dimensions of a room are randomly sampled between $[5, 10]$ m, $[5, 10]$ m, and $[3, 4]$ m, respectively. As shown in Fig. 2, we use an eight-microphone linear array with 8 cm inter-microphone distance and a seven-microphone circular array with a 4.25 cm radius. We pick the center two microphones from the linear array for two-microphone experiments. The microphone array is placed around the room center; more specifically, the array is randomly displaced between $[-0.5, 0.5]$ m along the length and width dimensions and rotated between $[-60, 60]^\circ$ relative to the room center.

Target and noise signals are sampled at 16 kHz and are placed around the microphone array on a circle of radius sampled between $[1, 3]$ m. Room impulse responses (RIRs) are generated using the image source method (Allen and Berkley, 1979). Speech and noise are convolved with the simulated multi-channel RIRs, and SNR is computed with respect to reverberant speech and reverberant noise. The detailed simulation procedure is provided in Algorithm 1. We normalize the mixture signals such that the root mean square of the mixture waveform is 1. The same scaling factor is applied to the target speech signal. Train and validation dataset sizes are 100 000 and 10 000, respectively.

For the diffuse noise test setup with simulated RIRs, room size is picked from three rooms: $[6, 6, 2.4]$ m³,

ALGORITHM 1. Dataset simulation process.

For each sample in training and validation do

1. Randomly sample acoustic scene parameters
 - 1.1 Room length, width, and height dimensions from $[5, 5, 3]$ to $[10, 10, 4]$ m;
 - 1.2 Microphone array displacement from the room center in length and width dimensions from $[-0.5, 0.5]$ m;
 - 1.3 Microphone array rotation from $[-60, 60]^\circ$;
 - 1.4 Source-to-microphone distance from $[1, 3]$ m;
 - 1.5 Source DOA from $[0, 180]^\circ$ for linear array and $[0, 360]^\circ$ for circular array;
 - 1.6 Reverberation T60 from $[0, 1.0]$ s;
 - 1.7 SNR from $[-5, 5]$ dB;
2. Generate multi-channel room impulse responses using the image source method and convolve with source signals to obtain reverberant sources;
3. Generate multi-channel diffuse noise or point-source noise signals and scale them to target SNR levels;
4. Add the scaled diffuse noise or point-source noise to the reverberant speech to obtain a reverberant and noisy multi-channel mixture;

$[8, 8, 3]$ m³, and $[10, 10, 4]$ m³. A microphone array is placed at the room center. Target and noise signals are placed around the microphone array on a circle with a radius chosen from $\{1, 2, 3\}$ m. The DOA grid is sampled between $[0, 180]^\circ$ and $[0, 360]^\circ$ in 5° intervals, resulting in totals of 37 and 72 possible DOA directions for the linear and circular array respectively. T60 and SNR are picked from $\{0.2, 0.4, 0.6, 0.8, 1.0\}$ s and $\{-5, 0, 5\}$ dB, respectively. All possible speaker DOAs are considered for every T60, SNR, and source-to-microphone distance resulting in $37 \times 5 \times 3 \times 3 = 1665$ and $72 \times 5 \times 3 \times 3 = 3240$ configurations, respectively, for the linear and circular array in each room.

For the point-source noise test setup with simulated RIRs, room size is fixed to $[6, 6, 2.4]$ m³. A microphone array is placed at the room center. Target and noise signals are placed around the microphone array on a circle with a radius chosen from $\{1, 2\}$ m. We use the same DOA grid, T60, and SNR values as in the diffuse noise test setup. All possible DOA pairs of speaker and noise (not from the same direction) are considered for every T60, SNR, and source-to-microphone distance, resulting in $37 \times 36 \times 5 \times 3 \times 2 = 39960$ and $72 \times 71 \times 5 \times 3 \times 2 = 153360$ configurations, respectively, for the linear array and circular array.

For the test setup with recorded RIRs, we pick the eight-microphone linear array with 8 cm inter-microphone distance from Hadad *et al.* (2014) to match our training configuration. We pick the center two microphones for a two-microphone evaluation. The recorded RIRs are provided for 1 m and 2 m distances sampled from a DOA grid between $[0, 180]^\circ$ in 15° intervals, resulting in a total of 13 RIRs for each distance. Diffuse noises for real recordings are generated using TIMIT utterances convolved with 26 recorded RIRs from all possible directions and distances. Note that recorded diffuse noises are generated differently from simulated diffuse noises used in training. SNR, T60, and source-

to-microphone distance are picked from $\{-5, 0, 5\}$ dB, $\{0.16, 0.36, 0.61\}$ s, and $\{1, 2\}$ m, respectively.

For the STFT analysis, frame length and frame shift are 32 ms and 16 ms, respectively. For model training, we use the mini-batch size of 8 and an initial learning rate of 0.001, which decays by 0.98 for every 2 epochs. The maximum number of training epochs is set to 50, and mixed precision is used for training and testing. Note that the models are trained separately on 4-s segments for each experiment. We train our models for each loss function, noise type, and microphone array, resulting in $4 \times 2 \times 3 = 24$ MIMO models. For MISO systems, we train a model for each microphone designated as the reference microphone separately, resulting in $2 \times 2 \times 8 = 32$ and $2 \times 2 \times 2 = 8$ models for the linear arrays with eight and two microphones, respectively. To improve computational efficiency, we use the trained MISO model parameters from reference microphone 0 (see Fig. 2) to initialize the parameters for other microphones and fine-tune them for three epochs. We train $2 \times 2 = 4$ MISO models for the circular array, where a single MISO model is trained with reference microphone 1 and then the circular shift trick is applied to get the estimates for the other microphones on the circle (Wang and Wang, 2020).

IV. EVALUATIONS AND COMPARISONS

For DOA evaluation metrics, we consider frame-level accuracy (ACC) and mean absolute error (MAE) as defined below:

$$\text{ACC}(\%) = \frac{\sum_t D_t C_t}{\sum_t D_t} \times 100, \quad (13)$$

$$\text{MAE}(\circ) = \frac{\sum_t D_t |\theta_t - \hat{\theta}_t|}{\sum_t D_t}, \quad (14)$$

where θ_t represents the ground truth DOA at frame t . D_t is a binary number indicating whether frame t has speech signal, as determined by a voice activity detector (Wiseman, 2019). $C_t = 1$ if and only if $|\theta_t - \hat{\theta}_t| \leq \theta_{th}$, where $\theta_{th} = 5^\circ$ is commonly chosen as the tolerance threshold (Mack *et al.*, 2022).

For speaker localization in reverberant and noisy environments, we compare with standard GCC-PHAT, oracle MGCC-PHAT, and three DNN-based methods. The oracle MGCC method uses the IRM, not its estimate, for computing weights, and it serves as an upper bound for the two-stage systems that combine real-valued masks and MGCC. Mask_{BLSTM} (Wang *et al.*, 2019) uses a two layer recurrent network with bi-directional long short-term memory (BLSTM) to estimate the real-valued PSM using a log power spectrogram across all microphones and then performs MGCC for utterance-level localization. This model has 13.1×10^6 (M) parameters. To obtain frame-level results for Mask_{BLSTM}, we adapt MGCC, similar to Eq. (12). SADOA (Mack *et al.*, 2022) employs cascaded DNNs with 12.5 M parameters, and is trained end-to-end, for utterance-level DOA estimation. The first DNN estimates a

real-valued mask, which is then utilized within the second DNN. We modify the first DNN to BLSTM for a fair comparison. Although the model is trained at the utterance level, it inherently outputs DOA estimates at the frame level. For frame-level results, we pick the DOA as the angle with the highest value. Phase_{CNN} (Vargas *et al.*, 2021) uses a CNN with 8.7 M parameters for DOA classification with frame-level mixture phase spectrograms as inputs.

A. DOA results with simulated RIRs

In Table I, we evaluate the frame-level DOA of the proposed model with different loss functions and the comparison methods using simulated RIRs in the presence of diffuse noise for three microphone array geometries. The arrays comprise two linear ones with two and eight microphones, denoted as linear-2 and linear-8, respectively, and a circular array with 7 microphones (circular-7). The DOA performance of GCC-PHAT directly applied on microphone signals is very poor across microphone arrays, reflecting the challenging nature of noisy and reverberant test conditions. Incorporating the ideal weighting mechanism in oracle MGCC improves localization results compared to GCC-PHAT. This represents the upper bound of DOA performance achievable with masking-based enhancement and MGCC. Mask_{BLSTM} uses estimated PSMs from a DNN as a weighting mechanism, and its frame-level DOA performance is better than that of GCC-PHAT but worse than that of the oracle MGCC. SADOA employs a DNN for localization along with a weighting mechanism and outperforms GCC-PHAT and Mask_{BLSTM}. Phase_{CNN} employs a DNN for

localization using noisy phases without a weighting mechanism and has better performance than GCC-PHAT and Mask_{BLSTM}. The utterance-level DOA results, provided in parentheses, of Mask_{BLSTM} and SADOA are improved, but their frame-level performance is poor. For instance, even at the high SNR of 5 dB with eight-microphone linear and seven-microphone circular arrays, the oracle MGCC method yields only approximately 70% and 60% of frames with DOA estimates within the acceptable tolerance across utterances. Relying on noisy phases significantly limits the frame-level performance of the baseline methods.

The proposed approach utilizes the clean phase estimated from CSM-based speech enhancement in the WGCC for speaker localization. As demonstrated by the results in Table I, the phase estimates from the enhancement models are reliable and the DOA performance of our systems surpasses the oracle MGCC and is much better than the DNN baselines of Mask_{BLSTM}, SADOA, and Phase_{CNN}. The frame-level results of our model exceed even the utterance-level results of Mask_{BLSTM} and SADOA.

We now compare the DOA performance of our systems with different loss functions and varying numbers of outputs. From Table I, we observe that the MISO systems trained with $\mathcal{L}_{\text{RI}}^{\text{MISO}}$ and $\mathcal{L}_{\text{RI+Mag}}^{\text{MISO}}$ already achieve significant improvements in DOA performance over the baselines. Among the two MISO systems, $\mathcal{L}_{\text{RI+Mag}}^{\text{MISO}}$ yields better DOA performance. The proposed MIMO based model with WGCC outperforms the corresponding MISO systems. The results demonstrate that extending MISO to MIMO systems, i.e., turning $\mathcal{L}_{\text{RI}}^{\text{MISO}}$ to \mathcal{L}_{RI} , and $\mathcal{L}_{\text{RI+Mag}}^{\text{MISO}}$ to $\mathcal{L}_{\text{RI+Mag}}$, elevates the DOA performance. Similar to the MISO results, $\mathcal{L}_{\text{RI+Mag}}$

TABLE I. DOA results with simulated RIRs under diffuse noise conditions at three SNR levels. The second column indicates whether T-F weighting is used. Utterance-level DOA results are presented in parentheses, and bold type indicates the best result.

| Method/Loss | T-F Weighting | Linear-2 | | | | | | Linear-8 | | | | | | Circular-7 | | | | | |
|---|---------------|-------------|------------|-------------|------------|-------------|------------|-------------|------------|-------------|------------|-------------|------------|-------------|------------|-------------|------------|-------------|------------|
| | | -5 dB | | 0 dB | | 5 dB | | -5 dB | | 0 dB | | 5 dB | | -5 dB | | 0 dB | | 5 dB | |
| | | ACC | MAE | ACC | MAE | ACC | MAE | ACC | MAE | ACC | MAE | ACC | MAE | ACC | MAE | ACC | MAE | ACC | MAE |
| GCC-PHAT | ✗ | 12.0 | 48.2 | 15.4 | 44.7 | 20.1 | 40.2 | 20.9 | 45.8 | 28.4 | 39.7 | 37.6 | 33.1 | 14.6 | 71.2 | 21.5 | 61.8 | 30.4 | 51.3 |
| Oracle MGCC | ✓ | 24.1 | 37.8 | 30.5 | 32.7 | 38.1 | 27.6 | 52.9 | 23.3 | 61.7 | 18.5 | 69.8 | 14.3 | 40.6 | 39.1 | 50.9 | 30.9 | 60.8 | 24.2 |
| Mask _{BLSTM} | ✓ | 15.1 | 43.2 | 19.9 | 38.6 | 25.4 | 34.3 | 29.3 | 34.2 | 38.3 | 28.5 | 46.5 | 23.9 | 22.5 | 58.4 | 32.5 | 47.6 | 42.7 | 38.7 |
| | ✓ | (55.0) | (11.2) | (62.5) | (7.8) | (68.2) | (6.0) | (75.8) | (6.7) | (82.9) | (4.4) | (85.9) | (3.3) | (90.8) | (4.2) | (98.6) | (1.1) | (99.5) | (1.5) |
| SADOA | ✓ | 17.9 | 47.7 | 23.2 | 42.4 | 28.3 | 38.4 | 35.0 | 33.4 | 44.4 | 27.3 | 52.3 | 22.9 | 33.7 | 49.3 | 44.9 | 40.8 | 53.6 | 35.1 |
| | ✓ | (58.9) | (11.5) | (69.2) | (8.1) | (75.2) | (6.3) | (88.0) | (4.4) | (90.4) | (3.7) | (92.0) | (3.3) | (95.8) | (5.3) | (98.6) | (5.4) | (99.0) | (5.0) |
| Phase _{CNN} | ✗ | 16.6 | 44.7 | 21.3 | 40.1 | 27.1 | 35.4 | 32.6 | 34.7 | 43.0 | 27.7 | 52.5 | 22.1 | 32 | 50.9 | 45.2 | 39.0 | 56.8 | 30.0 |
| $\mathcal{L}_{\text{RI}}^{\text{MISO}}$ | ✗ | 37.5 | 25.4 | 48.8 | 19.9 | 58.3 | 15.8 | 79.3 | 9.3 | 87.2 | 6.0 | 91.8 | 3.9 | 73.3 | 11.0 | 83.1 | 6.2 | 88.7 | 3.9 |
| | ✓ | 52.0 | 18.5 | 66.3 | 12.5 | 73.7 | 9.4 | 85.1 | 4.9 | 90.5 | 3.3 | 93.6 | 2.7 | 79.8 | 10.5 | 87.6 | 6.9 | 91.6 | 4.7 |
| $\mathcal{L}_{\text{RI+Mag}}^{\text{MISO}}$ | ✗ | 53.3 | 17.1 | 60.8 | 14.0 | 68.0 | 11.3 | 83.5 | 7.2 | 89.5 | 4.9 | 93.2 | 3.3 | 83.2 | 4.3 | 89.8 | 2.6 | 93.4 | 1.8 |
| | ✓ | 61.1 | 12.4 | 71.5 | 8.8 | 77.5 | 6.9 | 91.7 | 2.7 | 95.4 | 1.7 | 97.2 | 1.2 | 87.7 | 5.8 | 93.2 | 3.4 | 95.3 | 2.4 |
| \mathcal{L}_{RI} | ✗ | 40.7 | 22.9 | 49.9 | 18.7 | 58.7 | 15.2 | 74.9 | 11.8 | 83.5 | 8.0 | 89.0 | 5.4 | 70.5 | 10.8 | 80.9 | 6.4 | 86.3 | 4.5 |
| | ✓ | 63.7 | 9.8 | 72.7 | 7.7 | 77.4 | 6.7 | 87.0 | 4.0 | 92.0 | 2.6 | 94.8 | 1.8 | 83.2 | 7.9 | 88.6 | 5.2 | 91.2 | 4.0 |
| $\mathcal{L}_{\text{RI+Mag}}$ | ✗ | 52.5 | 17.1 | 60.8 | 14.0 | 68.0 | 11.3 | 83.8 | 6.8 | 89.4 | 4.8 | 92.5 | 3.4 | 76.3 | 6.1 | 85.1 | 3.8 | 90.1 | 2.6 |
| | ✓ | 67.8 | 8.6 | 77.1 | 6.5 | 81.4 | 5.6 | 92.7 | 2.2 | 95.7 | 1.4 | 96.8 | 1.1 | 89.4 | 3.8 | 93.7 | 2.4 | 95.5 | 1.7 |
| $\mathcal{L}_{\text{RI+IPD}}$ | ✗ | 75.7 | 5.6 | 82.2 | 4.6 | 86.0 | 3.8 | 87.4 | 5.3 | 92.5 | 3.5 | 95.2 | 2.3 | 93.2 | 1.9 | 95.5 | 1.4 | 96.6 | 1.0 |
| | ✓ | 79.7 | 3.9 | 86.4 | 2.9 | 89.3 | 2.5 | 92.8 | 2.1 | 96.1 | 1.2 | 97.9 | 0.8 | 96.1 | 1.7 | 97.4 | 1.2 | 97.8 | 1.0 |
| $\mathcal{L}_{\text{RI+Mag+IPD}}$ | ✗ | 74.9 | 6.7 | 81.7 | 5.3 | 86.1 | 4.1 | 90.7 | 3.6 | 94.1 | 2.6 | 96.1 | 1.8 | 88.3 | 2.6 | 93.7 | 1.6 | 95.9 | 1.1 |
| | ✓ | 80.9 | 3.9 | 87.5 | 2.9 | 90.4 | 2.4 | 95.1 | 1.5 | 97.4 | 0.9 | 98.3 | 0.7 | 94.8 | 1.8 | 97.2 | 1.1 | 98.0 | 0.8 |

yields better DOA performance compared to \mathcal{L}_{RI} . Although \mathcal{L}_{RI} and $\mathcal{L}_{\text{RI+Mag}}$ achieve high DOA performance, incorporating IPDs into the loss functions, i.e., turning \mathcal{L}_{RI} to $\mathcal{L}_{\text{RI+IPD}}$, and $\mathcal{L}_{\text{RI+Mag}}$ to $\mathcal{L}_{\text{RI+Mag+IPD}}$, further increases DOA accuracy and lowers MAE. In addition, incorporating the weighting mechanism in WGCC [see Eq. (11)] further improves localization results, especially at lower SNRs.

Figure 3 presents the DOA results of the proposed approach with different losses and the oracle MGCC for different arrays. It plots the average DOA accuracies of the oracle MGCC and the proposed system across all SNR, T60, and source-to-microphone distance values for different microphone arrays. It is evident from the figure that utilizing clean phase estimates consistently leads to large improvements compared to the oracle MGCC. The bar plots in Fig. 3 show that the MIMO systems outperform the MISO system, and among the loss functions, $\mathcal{L}_{\text{RI+Mag+IPD}}$ consistently performs the best.

Figure 4 presents the DOA results of the proposed approach and the oracle MGCC for the two-microphone array with respect to different source-to-microphone distances and T60 values. It is worth noting that $\mathcal{L}_{\text{RI+Mag+IPD}}$ achieves better or comparable results at the longest distance of 3 m and highest T60 of 1 s compared to other loss functions at the shortest distance of 1 m and lowest T60 of 0.2 s.

Table II evaluates the frame-level DOA of the proposed approach with different losses and the baseline methods using simulated RIRs in the presence of point-source noises for different microphone arrays. Similar to the findings under diffuse noise conditions, the results in Table II demonstrate the superior performance of the proposed method over all the baselines across microphone arrays. We observe the same trend that MIMO models outperform MISO models. A comparison between the last two rows clearly shows that incorporating the \mathcal{L}_{IPD} loss further enhances the localization performance, demonstrating its utility across different array geometries. Among notable differences between Tables I and II, the frame-level DOA performance of the proposed method in point-source noise conditions is better at lower SNRs, likely because a diffuse noise is more

difficult to separate than a point-source noise. In addition, the magnitude loss [see Eq. (4)] is not helpful, particularly for the circular array.

Figure 5 illustrates the DOA results of the different loss functions for both diffuse and point-source noise conditions for the two-microphone array. As clear from the figure, the frame-level localization results from $\mathcal{L}_{\text{RI+Mag+IPD}}$ have a higher number of DOA estimates within the tolerance range and fluctuate less compared to $\mathcal{L}_{\text{RI+Mag}}$. The same observation holds when comparing $\mathcal{L}_{\text{RI+IPD}}$ and \mathcal{L}_{RI} .

B. DOA results with recorded RIRs and real recordings

Table III provides the DOA results, averaged over all SNR, T60, and source-to-microphone distance values (see Sect. III), with the recorded RIRs from Hadad *et al.* (2014) that match the training array geometry for the eight-microphone linear array. We pick the center two microphones for the two-microphone evaluation. Our analysis for each speaker direction indicates that the endfire directions, i.e., 0° and 180° , contribute to the majority of errors and are excluded from the evaluation, as done in Wang *et al.* (2019). In line with the simulated RIR results in Tables I and II, the results in Table III demonstrate that our approach works well on recorded RIRs and produces strong frame-level DOA results. Our systems yield much higher localization accuracies and lower MAEs than the baselines for both diffuse and point-source noise conditions using two- and eight-microphone linear arrays.

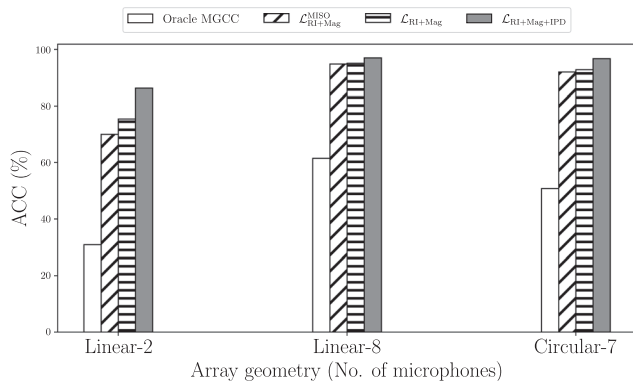


FIG. 3. Frame-level DOA accuracies of the oracle MGCC and the proposed method with different loss functions under diffuse noise conditions across different array geometries.

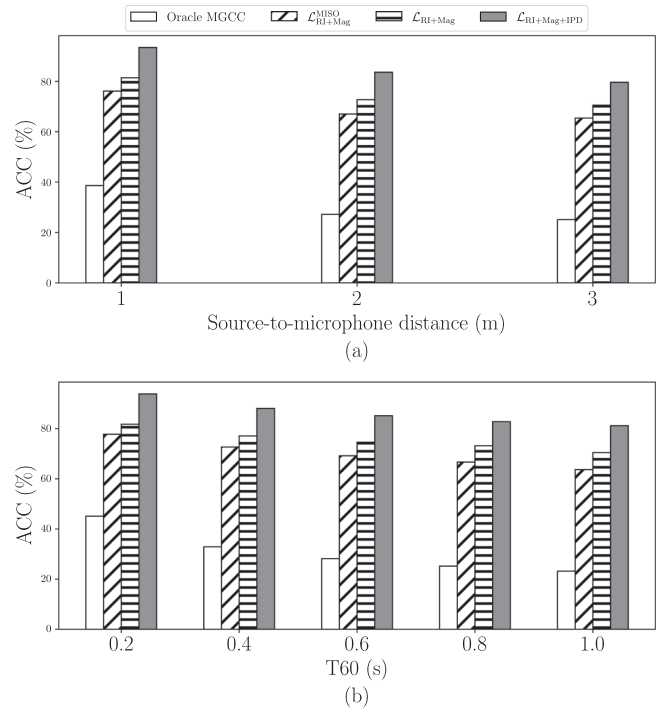


FIG. 4. Frame-level DOA accuracies of the oracle MGCC and the proposed method for the two-microphone array with different loss functions under diffuse noise conditions. (a) DOA accuracies with different source-to-microphone distances. (b) DOA accuracies with different T60 values.

TABLE II. DOA results with simulated RIRs under point-source noise conditions at three SNR levels. Utterance-level DOA results are presented in parentheses, and bold type indicates the best result.

| Method/Loss | Linear-2 | | | | | | Linear-8 | | | | | | Circular-7 | | | | | |
|--|-------------|------------|-------------|------------|-------------|------------|-------------|------------|-------------|------------|-------------|------------|-------------|------------|-------------|------------|-------------|------------|
| | -5 dB | | 0 dB | | 5 dB | | -5 dB | | 0 dB | | 5 dB | | -5 dB | | 0 dB | | 5 dB | |
| | ACC | MAE | ACC | MAE | ACC | MAE | ACC | MAE | ACC | MAE | ACC | MAE | ACC | MAE | ACC | MAE | ACC | MAE |
| GCC-PHAT | 11.6 | 49.9 | 15.1 | 46.4 | 19.1 | 42.2 | 16.7 | 51.4 | 24.4 | 45.4 | 33.0 | 38.7 | 13.5 | 76.7 | 21.8 | 67.4 | 30.0 | 56.1 |
| Oracle MGCC | 28.2 | 35.1 | 34.1 | 30.7 | 39.6 | 27.0 | 57.7 | 22.3 | 65.3 | 18.0 | 71.4 | 14.6 | 47.0 | 37.2 | 57.8 | 29.6 | 64.1 | 23.9 |
| Mask _{BLSTM} | 15.7 | 44.7 | 19.9 | 40.4 | 24.1 | 36.5 | 28.6 | 39.9 | 36.7 | 33.6 | 43.8 | 28.4 | 23.3 | 63.8 | 32.7 | 53.1 | 41.7 | 43.9 |
| | (48.7) | (19.0) | (57.9) | (11.4) | (61.9) | (8.4) | (68.3) | (13.6) | (78.8) | (6.1) | (83.7) | (3.9) | (75.7) | (26.0) | (92.5) | (7.8) | (98.1) | (2.9) |
| SADOA | 16.8 | 63.4 | 21.1 | 57.4 | 25.5 | 51.7 | 35.5 | 39.2 | 43.4 | 33.4 | 50.0 | 28.8 | 30.0 | 59.5 | 38.9 | 51.2 | 46.5 | 44.7 |
| | (59.4) | (16.5) | (68.5) | (9.8) | (73.0) | (7.4) | (80.6) | (10.5) | (89.4) | (4.9) | (93.0) | (3.0) | (88.5) | (13.9) | (96.9) | (5.1) | (99.1) | (2.6) |
| Phase _{CNN} | 15.7 | 51.0 | 20.1 | 46.6 | 24.8 | 42.0 | 30.7 | 41.4 | 37.4 | 36.4 | 43.6 | 32.3 | 29.7 | 57.4 | 38.1 | 49.3 | 45.1 | 43.2 |
| $\mathcal{L}_{\text{RI}}^{\text{MISO}}$ | 79.1 | 6.6 | 84.1 | 5.1 | 86.7 | 4.4 | 97.1 | 1.4 | 97.9 | 1.0 | 98.3 | 0.9 | 93.5 | 4.0 | 95.2 | 3.0 | 96.1 | 2.4 |
| $\mathcal{L}_{\text{RI}+\text{Mag}}^{\text{MISO}}$ | 79.1 | 6.0 | 84.2 | 4.4 | 87.0 | 3.8 | 97.1 | 1.3 | 98.1 | 0.9 | 98.5 | 0.8 | 95.4 | 2.3 | 96.8 | 1.7 | 97.4 | 1.4 |
| \mathcal{L}_{RI} | 82.8 | 5.1 | 86.5 | 4.0 | 88.4 | 3.5 | 95.6 | 1.9 | 97.2 | 1.3 | 97.8 | 1.1 | 97.1 | 1.7 | 97.9 | 1.2 | 98.2 | 1.1 |
| $\mathcal{L}_{\text{RI}+\text{Mag}}$ | 83.2 | 4.7 | 86.9 | 3.7 | 89.0 | 3.2 | 96.6 | 1.4 | 97.9 | 0.9 | 98.4 | 0.8 | 96.0 | 2.2 | 97.4 | 1.5 | 97.6 | 1.4 |
| $\mathcal{L}_{\text{RI}+\text{IPD}}$ | 86.5 | 3.7 | 89.4 | 3.0 | 90.7 | 2.5 | 96.9 | 1.3 | 98.3 | 0.8 | 98.8 | 0.6 | 97.4 | 1.6 | 98.3 | 1.1 | 98.7 | 0.9 |
| $\mathcal{L}_{\text{RI}+\text{Mag}+\text{IPD}}$ | 86.5 | 3.4 | 90.0 | 2.6 | 91.7 | 2.2 | 95.6 | 1.9 | 97.2 | 1.3 | 97.8 | 1.1 | 96.3 | 2.0 | 97.6 | 1.3 | 97.9 | 1.1 |

We further evaluate on the real recordings from Task 1 of the LOCATA evaluation dataset (Löllmann *et al.*, 2018), and the results are provided in Table III. This task involves localizing a single static speaker using static microphone arrays in a room with the dimensions [7.1, 9.8, 3] m³. The dataset uses speech sentences from the CSTR VCTK1 (Veaux *et al.*, 2017) database and provides the ground truth positional data of all microphones and sources. The acoustic environment contains reverberation with T60 = 0.55 s, and the source-to-microphone distance varies between 1 and 3.5 m. We have picked microphones 6 and 9 from the DICIT array to match the trained two-channel array geometry. Note that the microphone array and the source are not

on the same horizontal plane. Consistent with the findings using simulated and recorded RIRs, the results in the LOCATA column in Table III demonstrate that our approach generalizes well to real recordings. The results clearly show that our proposed systems outperforms baselines, with significantly higher localization accuracies and lower MAEs, although the amounts of improvement are not as high as for frame-level DOA estimation using simulated and recorded RIRs. In this evaluation, the inclusion of the IPD loss term yields slightly better performance compared to the counterparts without the term. This is probably due to the large differences between training and test conditions in both array geometries and acoustic environments.

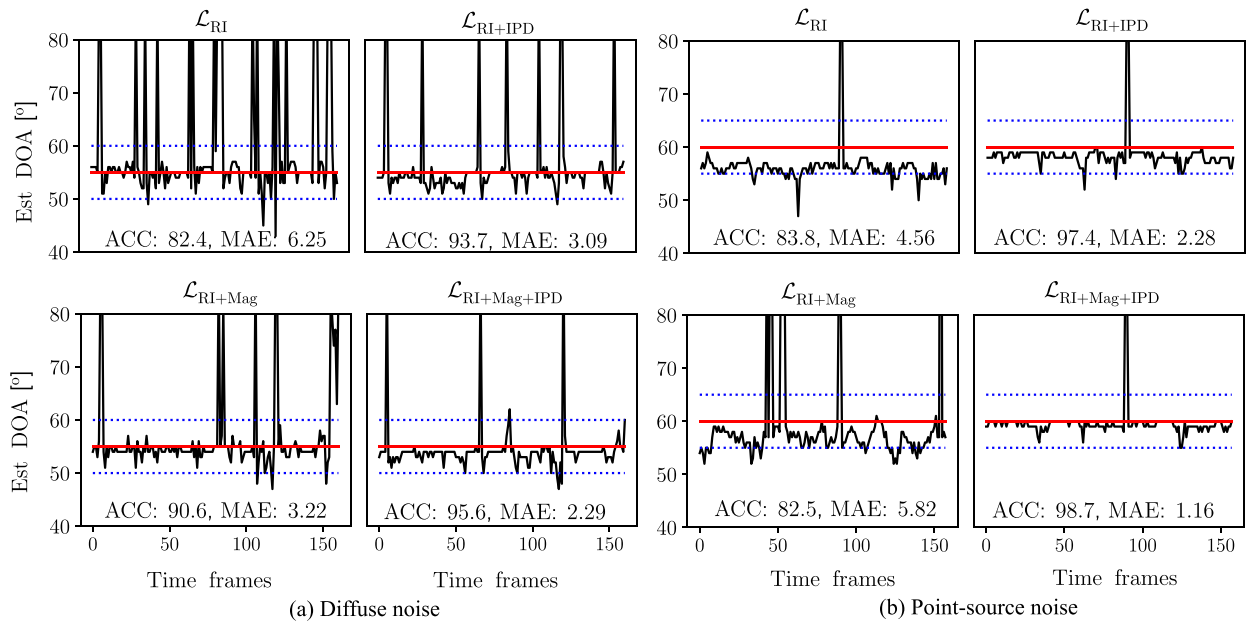


FIG. 5. Frame-level DOA estimation comparisons of different loss functions in a simulated environment with SNR = -5 dB and T60 = 1 s. (a) Diffuse noise conditions. (b) Point-source noise conditions. The red solid and blue dotted lines represent the ground truth DOA and tolerance range, respectively. The corresponding ACC and MAE results are shown in each case. Non-speech frames are excluded from the plot.

TABLE III. DOA results with recorded RIRs in diffuse and point-source noise conditions and on LOCATA task 1. Most of the Linear-2 results are already given in Battula *et al.* (2025), and included here for completeness. Bold type indicates the best result.

| Method/Loss | Diffuse noise | | | | Point-source noise | | | | LOCATA | |
|-------------------------------|---------------|-------------|-------------|------------|--------------------|------------|-------------|------------|-------------|------------|
| | Linear-2 | | Linear-8 | | Linear-2 | | Linear-8 | | Linear-2 | |
| | ACC | MAE | ACC | MAE | ACC | MAE | ACC | MAE | ACC | MAE |
| GCC-PHAT | 23.7 | 40.8 | 38.1 | 40.8 | 11.9 | 54.6 | 14.7 | 55.3 | 46.9 | 13.3 |
| Mask _{BLSTM} | 27.6 | 36.6 | 48.3 | 30.0 | 20.6 | 44.5 | 33.6 | 41.9 | 58.4 | 10.6 |
| SADOA | 23.8 | 49.6 | 43.2 | 36.7 | 21.9 | 60.6 | 49.6 | 29.8 | 51.9 | 27 |
| Phase _{CNN} | 20.1 | 59.1 | 41.6 | 38.0 | 18.2 | 49.9 | 42.5 | 34.8 | 51.7 | 14.5 |
| $\mathcal{L}_{\text{MISO}}$ | 45.6 | 26.8 | 86.5 | 6.2 | 83.4 | 4.1 | 99.1 | 1.5 | 70.3 | 5.3 |
| \mathcal{L}_{RI} | 59.2 | 19.5 | 92.3 | 3.5 | 84.7 | 3.6 | 99.0 | 1.4 | 73.8 | 4.6 |
| $\mathcal{L}_{\text{RI+Mag}}$ | 42.6 | 30.4 | 84.9 | 8.0 | 86.1 | 3.2 | 99.0 | 1.4 | 72.1 | 5.0 |
| \mathcal{L}_{RI} | 59.0 | 17.1 | 91.9 | 3.9 | 86.1 | 3.1 | 99.5 | 1.2 | 72.3 | 4.7 |
| $\mathcal{L}_{\text{RI+Mag}}$ | 60.5 | 18.8 | 86.0 | 8.6 | 85.5 | 3.2 | 99.8 | 1.2 | 73.5 | 4.5 |
| $\mathcal{L}_{\text{RI+IPD}}$ | 64.9 | 16.6 | 92.3 | 4.8 | 86.2 | 3.0 | 98.8 | 1.5 | 72.4 | 4.8 |

C. Speech enhancement results

As the localization results in the tables are based on multi-channel speech enhancement, their corresponding enhancement results are given in Table IV to provide insights into relative advantages of different loss functions. In the table, scores are averaged across the three SNR levels. Enhancement performance is measured in terms of the commonly used metrics of short-time objective intelligibility (STOI) (Taal *et al.*, 2011), perceptual evaluation of speech quality (PESQ) (Rix *et al.*, 2001), and scale-invariant signal-to-distortion ratio (SI-SDR) (LeRoux *et al.*, 2019). We compute the enhancement scores based on reference microphone 0. From Table IV, we observe that the inclusion of the IPD term in a loss function yields small amounts of PESQ improvement in speech enhancement results across microphone arrays and noise conditions and comparable STOI and SI-SDR scores. For instance, $\mathcal{L}_{\text{RI+IPD}}$ shows approximately a 0.1 improvement in PESQ for both the linear 2-microphone and circular 7-microphone arrays under diffuse noise conditions compared to \mathcal{L}_{RI} .

Figure 6 illustrates the role of IPD loss in the MIMO systems. As evident in the figure, the inclusion of the IPD

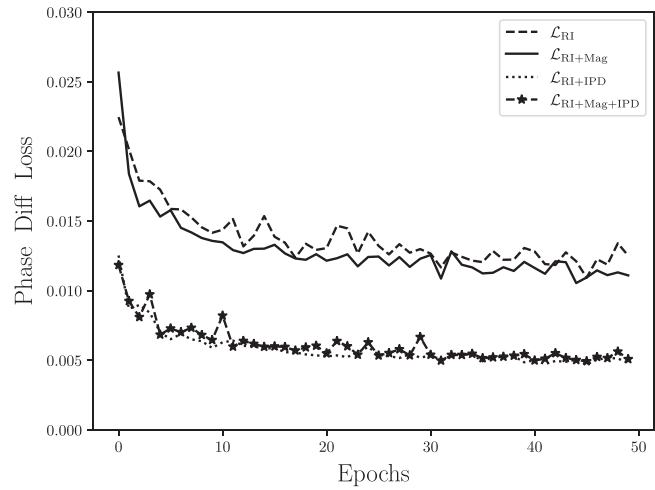


FIG. 6. Comparison of phase-difference loss term for different loss functions on the validation dataset for the two-microphone array in diffuse noise condition.

term results in lower phase-difference loss values for $\mathcal{L}_{\text{RI+Mag+IPD}}$ and $\mathcal{L}_{\text{RI+IPD}}$ than the counterparts of $\mathcal{L}_{\text{RI+Mag}}$ and \mathcal{L}_{RI} . The improvement seems to be magnified in the DOA results in Table I. This demonstrates that enhancement and localization are distinct tasks. Overall, $\mathcal{L}_{\text{RI+Mag+IPD}}$ yields the best localization and speech enhancement performance.

V. CONCLUSION

In this study, we have systematically investigated the proposed two-stage approach that combines multi-channel CSM for speech enhancement and WGCC for robust frame-level speaker localization. Our approach leverages the intrinsic ability of CSM to estimate phase and improves by a large margin the frame-level localization accuracy compared to conventional DOA techniques, magnitude-domain masking, and other DNN methods in reverberant and noisy environments. Our study shows that phase estimates from CSM models are reliable, and two-stage modeling provides a measure of explainability for frame-level localization. Additionally, we show that MIMO based systems consistently outperform MISO based systems. In the MIMO

TABLE IV. Speech enhancement results in diffuse and point-source noise conditions. Most of the Linear-2 results are already given in Battula *et al.* (2025) and are included here for completeness. Bold type indicates the best result.

| Loss | Diffuse noise | | | | | | | | | Point-source noise | | | | | | | | |
|-------------------------------|---------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|--------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | Linear-2 | | | Linear-8 | | | Circular-7 | | | Linear-2 | | | Linear-8 | | | Circular-7 | | |
| | STOI | PESQ | SI-SDR | STOI | PESQ | SI-SDR | STOI | PESQ | SI-SDR | STOI | PESQ | SI-SDR | STOI | PESQ | SI-SDR | STOI | PESQ | SI-SDR |
| Unprocessed | 54.1 | 1.31 | −9.1 | 54.1 | 1.31 | −9.1 | 54.1 | 1.31 | −9.1 | 55.4 | 1.36 | −11.3 | 55.8 | 1.36 | −11.3 | 55.4 | 1.36 | −11.3 |
| $\mathcal{L}_{\text{MISO}}$ | 87.9 | 2.24 | 10.2 | 92.5 | 2.66 | 12.9 | 92.7 | 2.71 | 13.5 | 96.6 | 3.46 | 14.0 | 97.6 | 3.78 | 14.7 | 97.6 | 3.77 | 15.5 |
| \mathcal{L}_{RI} | 88.9 | 2.61 | 10.1 | 93.0 | 3.05 | 13.0 | 93.3 | 3.07 | 13.5 | 96.7 | 3.70 | 13.2 | 97.4 | 3.87 | 14.0 | 97.7 | 3.90 | 15.4 |
| $\mathcal{L}_{\text{RI+Mag}}$ | 88.9 | 2.23 | 10.6 | 92.4 | 2.66 | 12.8 | 93 | 2.72 | 13.9 | 96.9 | 3.60 | 14.7 | 97.2 | 3.70 | 13.9 | 97.6 | 3.79 | 15.6 |
| \mathcal{L}_{RI} | 89.6 | 2.68 | 10.3 | 93.0 | 3.05 | 12.9 | 93.6 | 3.10 | 13.8 | 97.0 | 3.75 | 14.4 | 97.4 | 3.85 | 13.4 | 97.6 | 3.91 | 15.5 |
| $\mathcal{L}_{\text{RI+Mag}}$ | 88.9 | 2.33 | 10.7 | 92.5 | 2.68 | 13.0 | 93.4 | 2.81 | 13.9 | 96.0 | 3.53 | 13.5 | 97.2 | 3.71 | 13.7 | 97.7 | 3.80 | 15.7 |
| $\mathcal{L}_{\text{RI+IPD}}$ | 89.8 | 2.70 | 10.7 | 93.1 | 3.10 | 12.8 | 93.7 | 3.16 | 13.8 | 97.0 | 3.76 | 14.4 | 97.4 | 3.86 | 13.9 | 97.6 | 3.91 | 15.4 |

training framework, the proposed loss functions incorporating IPDs further enhance DOA performance. As a result, the proposed approach yields accurate speaker localization across microphone arrays and reverberant-noisy conditions in simulated and real environments.

In future research, we plan to extend the proposed approach to address moving speaker scenarios and multi-talker environments, as well as binaural sound localization [see, e.g., Phokhinanan *et al.* (2023)]. Additionally, we plan to incorporate a DNN in the second, DOA estimation, stage to further enhance frame-level localization performance. We will also assess whether performing enhancement and localization as multi-task learning can lead to better DOA estimation than the two-stage approach investigated in this paper.

ACKNOWLEDGMENTS

This research was supported in part by a research contract from Meta Reality Labs, the Ohio Supercomputer Center, and the Pittsburgh Supercomputer Center (NSF ACI-1928147). All research and experiments were conducted at the Ohio State University.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

- Allen, J. B., and Berkley, D. A. (1979). "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.* **65**, 943–950.
- Baek, M. S., Yang, J. Y., and Chang, J. H. (2023). "Deeply supervised curriculum learning for deep neural network-based sound source localization," in *Interspeech 2023*, Dublin, Ireland (ISCA, Stockholm), pp. 3744–3748.
- Battula, S., Taherian, H., Pandey, A., Wong, D., Xu, B., and Wang, D. L. (2025). "Robust frame-level speaker localization in reverberant environments by exploiting phase difference losses," in *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India (IEEE, New York), pp. 1–5.
- Bohlender, A., Spriet, A., Tirry, W., and Madhu, N. (2021). "Exploiting temporal context in CNN based multisource DOA estimation," *IEEE/ACM Trans. Audio Speech Lang. Process.* **29**, 1594–1608.
- Chakrabarty, S., and Habets, E. A. P. (2017). "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *2017 Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY (IEEE, New York), pp. 136–140.
- Darwin, C. J. (2008). "Listening to speech in the presence of other sounds," *Phil. Trans. R Soc. B* **363**, 1011–1021.
- DiBiase, J., Silverman, H., and Brandstein, M. (2001). "Robust localization in reverberant rooms," in *Microphone Arrays*, edited by M. Brandstein and D. Ward (Springer, Berlin), pp. 157–180.
- Fu, W., Hu, T. Y., Tsao, Y., and Lu, X. (2017). "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, Tokyo, Japan (IEEE, New York), pp. 1–6.
- Gannot, S., Vincent, E., Markovich-Golan, S., and Ozerov, A. (2017). "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio Speech Lang. Process.* **25**, 692–730.
- Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. (1993). *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus*, National Institute of Standards and Technology, Gaithersburg, MD.
- Goli, P., and van de Par, S. (2023). "Deep learning-based speech specific source localization by using binaural and monaural microphone arrays in hearing aids," *IEEE/ACM Trans. Audio Speech Lang. Process.* **31**, 1652–1666.
- Grumiaux, P.-A., Kitić, S., Girin, L., and Guérin, A. (2022). "A survey of sound source localization with deep learning methods," *J. Acoust. Soc. Am.* **152**, 107–151.
- Hadad, E., Heese, F., Vary, P., and Gannot, S. (2014). "Multichannel audio database in various acoustic environments," in *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Juan-les-Pins, France (IEEE, New York), pp. 313–317.
- Hartmann, W. M. (1999). "How we localize sound," *Phys. Today* **52**(11), 24–29.
- Hwang, S., Kim, M., and Shin, J. W. (2022). "Dual microphone speech enhancement based on statistical modeling of interchannel phase difference," *IEEE/ACM Trans. Audio Speech Lang. Process.* **30**, 2865–2874.
- Kabzinski, T., and Habets, E. A. P. (2019). "A least squares narrowband DOA estimator with robustness against phase wrapping," in *2019 27th European Signal Processing Conference (EUSIPCO)*, A Coruña, Spain (IEEE, New York), pp. 1–5.
- Knapp, C., and Carter, G. (1976). "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust. Speech Signal Process.* **24**, 320–327.
- LeRoux, J., Wisdom, S., Erdogan, H., and Hershey, J. R. (2019). "SDR—Half-baked or well done?," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK (IEEE, New York), pp. 626–630.
- Löllmann, H. W., Evers, C., Schmidt, A., Mellmann, H., Barfuss, H., Naylor, P. A., and Kellermann, W. (2018). "The LOCATA challenge data corpus for acoustic source localization and tracking," in *2018 IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, Sheffield, UK (IEEE, New York), pp. 410–414.
- Mack, W., Wechsler, J., and Habets, E. (2022). "Signal-aware direction-of-arrival estimation using attention mechanisms," *Comput. Speech Lang.* **75**, 101363.
- Olivan, C. H., Delcroix, M., Ochiai, T., Tawara, N., Nakatani, T., and Araki, S. (2024). "Interaural time difference loss for binaural target sound extraction," in *2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Aalborg, Denmark (IEEE, New York), pp. 210–214.
- Pak, J., and Shin, J. W. (2019). "Sound localization based on phase difference enhancement using deep neural networks," *IEEE/ACM Trans. Audio Speech Lang. Process.* **27**, 1335–1345.
- Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). "LibriSpeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, Australia (IEEE, New York), pp. 5206–5210.
- Pang, C., Liu, H., and Li, X. (2019). "Multitask learning of time-frequency CNN for sound source localization," *IEEE Access* **7**, 40725–40737.
- Pertilä, P., and Cakir, E. (2017). "Robust direction estimation with convolutional neural networks based steered response power," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA (IEEE, New York), pp. 6125–6129.
- Phokhinanan, W., Obin, N., and Argentieri, S. (2023). "Binaural sound localization in noisy environments using frequency-based audio vision transformer (FAViT)," in *Interspeech 2023*, Dublin Ireland (ISCA, Stockholm), pp. 3704–3708.
- Quan, C., and Li, X. (2024). "SpatialNet: Extensively learning spatial information for multichannel joint speech separation, denoising, and dereverberation," *IEEE/ACM Trans. Audio Speech Lang. Process.* **32**, 1310–1323.
- Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. (2001). "Perceptual evaluation of speech quality (PESQ)—a new method for

- speech quality assessment of telephone networks and codecs,” in *2001 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Salt Lake City, UT (IEEE, New York), pp. 749–752.
- Schmidt, R. (1986). “Multiple emitter location and signal parameter estimation,” *IEEE Trans. Antennas Propagat.* **34**, 276–280.
- Shimada, K., Koyama, Y., Takahashi, N., Takahashi, S., and Mitsufuji, Y. (2021). “ACCDQA: Activity-coupled cartesian direction of arrival representation for sound event localization and detection,” in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada (IEEE, New York), pp. 915–919.
- Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2011). “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” *IEEE Trans. Audio Speech Lang. Process.* **19**, 2125–2136.
- Taherian, H., Pandey, A., Wong, D. E., Xu, B., and Wang, D. L. (2023). “Multi-input multi-output complex spectral mapping for speaker separation,” in *Interspeech 2023* (ISCA, Stockholm), pp. 1070–1074.
- Tan, K., and Wang, D. L. (2020). “Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement,” *IEEE/ACM Trans. Audio Speech Lang. Process.* **28**, 380–390.
- Tan, K., Wang, Z. Q., and Wang, D. L. (2022). “Neural spectrospatial filtering,” *IEEE/ACM Trans. Audio Speech Lang. Process.* **30**, 605–621.
- Thaleiser, S., and Enzner, G. (2023). “Binaural-projection multichannel Wiener filter for cue-preserving binaural speech enhancement,” *IEEE/ACM Trans. Audio Speech Lang. Process.* **31**, 3730–3745.
- Thiemann, J., Ito, N., and Vincent, E. (2013). “The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings,” *Proc. Mtgs. Acoust.* **19**, 035081.
- Tokala, V., Grinstein, E., Brookes, M., Doclo, S., Jensen, J., and Naylor, P. A. (2024). “Binaural speech enhancement using deep complex convolutional transformer networks,” in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, South Korea (IEEE, New York), pp. 681–685.
- Vargas, E., Hopgood, J. R., Brown, K., and Subr, K. (2021). “On improved training of CNN for acoustic source localisation,” *IEEE/ACM Trans. Audio Speech Lang. Process.* **29**, 720–732.
- Varzandeh, R., Doclo, S., and Hohmann, V. (2024). “Speech-aware binaural DOA estimation utilizing periodicity and spatial features in convolutional neural networks,” *IEEE/ACM Trans. Audio Speech Lang. Process.* **32**, 1198–1213.
- Veaux, C., Yamagishi, J., and MacDonald, K. (2017). “English multi-speaker corpus for CSTR voice cloning toolkit,” <https://datashare.ed.ac.uk/handle/10283/3443> (Last viewed 02/10/2025).
- Vincent, E., Virtanen, T., and Gannot, S., eds. (2018). *Audio Source Separation and Speech Enhancement* (Wiley, Hoboken, NJ).
- Wang, Y., Yang, B., and Li, X. (2024). “IPDnet: A universal direct-path IPD estimation network for sound source localization,” *IEEE/ACM Trans. Audio Speech Lang. Process.* **32**, 5051–5064.
- Wang, Z. Q., and Wang, D. L. (2020). “Deep learning based target cancellation for speech dereverberation,” *IEEE/ACM Trans. Audio Speech Lang. Process.* **28**, 941–950.
- Wang, Z. Q., Zhang, X., and Wang, D. L. (2019). “Robust speaker localization guided by deep learning-based time-frequency masking,” *IEEE/ACM Trans. Audio Speech Lang. Process.* **27**, 178–188.
- Williamson, D. S., Wang, Y., and Wang, D. L. (2016). “Complex ratio masking for monaural speech separation,” *IEEE/ACM Trans. Audio Speech Lang. Process.* **24**, 483–492.
- Wiseman, J. (2019). “Wiseman/py-webrtcvad,” GitHub, <https://github.com/wiseman/py-webrtcvad>.
- Yang, B., Li, X., and Liu, H. (2021). “Supervised direct-path relative transfer function learning for binaural sound source localization,” in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada (IEEE, New York), pp. 825–829.
- Zhang, W., Zhou, Y., and Qian, Y. (2019). “Robust DOA estimation based on convolutional neural network and time-frequency masking,” in *Interspeech 2019*, Graz, Austria (ISCA, Stockholm), pp. 2703–2707.
- Zhang, X., and Wang, D. L. (2017). “Deep learning based binaural speech separation in reverberant environments,” *IEEE/ACM Trans. Audio Speech Lang. Process.* **25**, 1075–1084.