ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10890098

Robust Frame-level Speaker Localization in Reverberant and Noisy Environments by Exploiting Phase Difference Losses

Shanmukha Srinivas Battula[1], Hassan Taherian[1], Ashutosh Pandey[3], Daniel Wong[3], Buye Xu[3], and DeLiang Wang[1,2]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Center for Cognitive and Brain Sciences, The Ohio State University, USA
[3]Meta Reality Labs, USA

***Abstract*—This paper investigates robust speaker localization at the frame level on the basis of complex spectral mapping, which is capable of learning both the magnitude and phase of the target signal. Unlike prevailing deep learning methods for speaker localization, we perform MIMO (multi-input multi-output) based multi-channel speech enhancement first and then localize the enhanced speaker using weighted generalized cross correlation. In addition, we propose new multi-channel loss functions that incorporate phase differences in order to preserve inter-channel phase relations, which is key to accurate sound localization. Systematic evaluations using simulated and recorded room impulse responses 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.**

***Index Terms*—MIMO speech enhancement, robust speaker localization, complex spectral mapping, inter-channel phase difference loss.**

I. Introduction

Direction of arrival (DOA) estimation aims to find the azimuth angle of a target source originating from a specific spatial location relative to a microphone array. This estimation relies on spatial features such as inter-channel time differences, inter-channel phase differences (IPD), inter-channel level differences, and the relative transfer function (RTF). Standard DOA estimation methods include generalized cross correlation with phase transform (GCC-PHAT) [1] and steered response power with phase transform [2]. However, the performance of these algorithms severely degrades in reverberant and noisy environments.

Most recent works on robust speaker localization employ a deep neural network (DNN) to directly estimate DOA [3]. Typically these methods use spatial features, GCC-PHAT, or spectrograms as inputs. For instance, the studies in [4] and [5] utilize IPDs, and those in [6]–[8] use phase spectrograms as input to generate posterior probabilities for each DOA class. Other studies estimate the spatial features such as direct path RTF [9] and IPD [10].

An alternative approach involves enhancing reverberant and noisy signals and subsequently using the enhanced signals for DOA estimation. 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 [11]–[13]. Wang et al. [12] 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 a weighting mechanism to amplify the T-F units dominated by the target source while utilizing the phase of mixture signals in GCC-PHAT calculations. Likewise, an estimated real-valued mask is utilized within a CNN (convolutional neural network) as a weighting mechanism for estimating DOA in [13]. However, these algorithms rely on the phase of the mixture signals in localization, which limits their frame-level DOA performance.

Recently, speech enhancement methods use complex spectral mapping (CSM) [14], [15] which can estimate clean phase, a capability not shared by mask-based methods in the magnitude domain. Utilizing clean phase estimates overcomes the limitations of using noisy phases in two-stage approaches for robust localization. Furthermore, simultaneously estimating a target source in all microphone outputs is computationally and conceptually more efficient than estimating the target in each of the microphones individually.

In this context, we propose a two-stage approach for robust frame-level speaker localization where the first stage employs MIMO (multi-input multi-output) based complex spectral mapping for speech enhancement and the second stage applies weighted GCC-PHAT (WGCC) for localization. MIMO modeling allows us to explore inter-channel relations in training losses, which is a distinct advantage over MISO (multi-input single-output) based methods. Motivated by the use of inter-channel constraints in loss functions [16]–[18], 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 in both simulated and real environments.

Section II describes the problem formulation for DOA, MIMO based CSM for speech enhancement, training objectives, and weighted GCC-PHAT for localization. Section III provides the experimental setup. Section IV presents evaluation results and comparisons, followed by a conclusion in Section V.

II. System Description

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

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

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 estimate the anechoic speech $\mathbf{S}$, i.e., $\widehat{\mathbf{S}} = \mathcal{F}(\mathbf{Y})$, where $\mathcal{F}$ denotes the estimator, and consequently its direction θ.

A. Multi-Channel Complex Spectral Mapping

The standard formulation of multi-channel complex spectral mapping is to estimate the complex spectrogram of the target speech at a reference microphone m, $\widehat{\mathbf{S}}_m = \mathcal{F}(\mathbf{Y};\Theta)$, where Θ denotes the set of all model parameters, from the noisy mixture [14]. Specifically,

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).

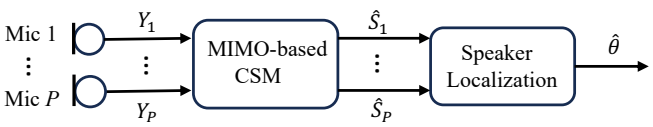

Fig. 1. Block diagram of the proposed two-stage system using MIMO-based speech enhancement and speaker localization with a *P*-channel microphone array.

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. In this study, we propose to use the large version of SpatialNet described in [19] for multi-channel speech enhancement. We extend SpatialNet to a MIMO system by adding as many final linear layers as the number of microphones to estimate multiple target outputs, i.e., $\widehat{\mathbf{S}} = \mathcal{F}(\mathbf{Y}; \Theta)$. Fig. 1 illustrates a *P*-channel setup.

B. Training Objectives

Following earlier studies [19]–[22], the multi-output loss function is calculated by comparing the real and imaginary spectrograms of the separated ($\widehat{S}$) and target speech (S) at each microphone p:

$$\mathcal{L}_{\text{RI+Mag}} = \mathcal{L}_{\text{RI}} + \mathcal{L}_{\text{Mag}} \tag{2}$$

$$\mathcal{L}_{\text{RI}} = \frac{1}{P}\sum_{p=1}^{P}(\|\widehat{S}_p^{(r)} - S_p^{(r)}\|_1 + \|\widehat{S}_p^{(i)} - S_p^{(i)}\|_1) \tag{3}$$

$$\mathcal{L}_{\text{Mag}} = \frac{1}{P}\sum_{p=1}^{P}\||\widehat{S}_p| - |S_p|\|_1 \tag{4}$$

where superscripts (r) and (i) denote real and imaginary parts respectively, and $\|.\|_1$ the l_1 norm.

The MIMO modeling allows us to compare inter-channel phase differences between the estimated and target spectrograms via a loss function. We define:

$$\mathcal{L}_{\text{IPD}} = \frac{1}{N_\Omega}\sum_{(p,q)\in\Omega} W_{p,q} \begin{matrix}(\|\cos\widehat{S}_{\text{IPD}}^{(p,q)} - \cos S_{\text{IPD}}^{(p,q)}\|_1 + \\ \|\sin\widehat{S}_{\text{IPD}}^{(p,q)} - \sin S_{\text{IPD}}^{(p,q)}\|_1)\end{matrix} \tag{5}$$

where $\widehat{S}_{\text{IPD}}^{(p,q)}, S_{\text{IPD}}^{(p,q)}$ represent the IPD between a microphone pair (*p, q*) of the estimated and target speech respectively. Ω is the set of all possible microphone pairs with cardinality $N_\Omega = \frac{P(P-1)}{2}$. The weighting, $W_{p,q}$, is chosen to be a product of the corresponding target speech IRM values. This weighting mechanism serves to accentuate the T-F units where the target speech dominates.

Incorporating $\mathcal{L}_{\text{IPD}}$ imposes a penalty on IPD estimation errors, reflecting the primary role of IPD in sound localization. Considering different ways of including an IPD term, we investigate the following two reasonable alternatives:

$$\mathcal{L}_{\text{RI+IPD}} = \mathcal{L}_{\text{RI}} + \mathcal{L}_{\text{IPD}} \tag{6}$$

$$\mathcal{L}_{\text{RI+Mag+IPD}} = \mathcal{L}_{\text{RI+Mag}} + \mathcal{L}_{\text{IPD}} \tag{7}$$

C. Weighted GCC-PHAT for Localization

For a microphone pair (*p, q*), the GCC-PHAT algorithm [1] computes their generalized cross-correlation coefficients with a weighting mechanism based on phase transform to estimate the time delay:

$$GCC_{p,q}(t,f,\theta) = \mathcal{R}\left\{\frac{Y_p(t,f)Y_q(t,f)^H}{|Y_p(t,f)||Y_q(t,f)^H|}e^{-j2\pi\frac{f}{N}f_s\tau_{p,q}(\theta)}\right\} \tag{8}$$

where $\mathcal{R}\{\cdot\}$ extracts the real part, $(\cdot)^H$ represents conjugate transpose, $|\cdot|$ computes the magnitude, j denotes the imaginary unit, N the number of discrete Fourier transform frequencies, and f_s 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 to microphone p and q, respectively.

We introduce WGCC to improve the performance of GCC-PHAT in the presence of reverberation and noise:

$$WGCC_{p,q}(t,f,\theta) = W_{p,q}(t,f)GCC_{p,q}(t,f,\theta) \tag{9}$$

$$W_{p,q}(t,f) = W_p(t,f)W_q(t,f) \tag{10}$$

Equations (9) and (10) bear similarities to the mask-weighted GCC-PHAT method introduced in [12]. However, there are two key differences. Firstly, in the mask-weighted method the GCC function is calculated from noisy observations as defined in (8), whereas the WGCC function in our study is calculated from the enhanced speech signals (see Fig. 1). Secondly, the mask-weighted method employs a ratio mask as the weight per T-F unit, whereas WGCC uses the mask value computed from the estimated target complex spectrogram.

Finally, frame-level DOA is estimated as

$$\widehat{\theta}_t = \arg\max_{\theta}\sum_{(p,q)\in\Omega}\sum_{f=1}^{N/2} WGCC_{p,q}(t,f,\theta) \tag{11}$$

III. Experimental Setup

We used the LibriSpeech corpus [23] for target speech signals, DEMAND dataset [24] for point-source noise simulations, and the TIMIT corpus [25] to create diffuse noises as described in [15]. 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. We use a 2-microphone linear array with an 8 cm inter-microphone distance and a 7-microphone circular array with a 4.25 cm radius. A microphone array is placed around the room center; more specifically, the array is randomly placed between [-0.5, 0.5] m along length, width, and height dimensions and rotated between [-60, 60]$^{\text{o}}$. 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 [26]. Speech and noise are convolved with the simulated multi-channel RIRs and SNR is computed with respect to reverberant speech and reverberant noise. 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 100k and 10k respectively.

For the test setup with simulated RIRs, room size is picked from three rooms: [6, 6, 2.4], [8, 8, 3], and [10, 10, 4] m^3. The 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]$^{\text{o}}$ and [0, 360]$^{\text{o}}$ in 5$^{\text{o}}$ intervals, resulting in a total of 37 and 72 possible DOA directions, for 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. For diffuse noise simulations, all possible speaker DOAs are considered for every T60, SNR, and source-to-microphone distance resulting in 37×5×3×3 = 1665 and 72×5×3×3 = 3240 configurations for linear and circular array in each room. For point-source noise simulations, room size and source-to-microphone distance are fixed to [6, 6, 2.4]

TABLE I

DOA RESULTS WITH SIMULATED RIRS IN DIFFUSE NOISE AND POINT-SOURCE NOISE CONDITIONS AT THREE SNR LEVELS USING A LINEAR ARRAY WITH TWO MICROPHONES. UTTERANCE-LEVEL DOA RESULTS ARE PRESENTED IN PARENTHESES. BOLD TYPE INDICATES THE BEST RESULT.

Method/Loss	Params	Diffuse noise						Point-source noise					
		-5 dB		0 dB		5 dB		-5 dB		0 dB		5 dB	
	(M)	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	14.0	49.8	18.9	45.2	24.6	39.8
Oracle MGCC	-	24.1	37.8	30.5	32.7	38.1	27.6	35.9	29.9	43.2	25.3	49.6	21.6
$\text{Mask}_{\text{BLSTM}}$ [12]	13.1	15.1 (55.0)	43.2 (11.2)	19.9 (62.5)	38.6 (7.8)	25.4 (68.2)	34.3 (6.0)	20.1 (70.6)	42.4 (11.2)	26.0 (83.3)	37.2 (13.8)	31.9 (88.9)	32.7 (2.8)
SADOA [13]	12.5	17.9 (58.9)	47.7 (11.5)	23.2 (69.2)	42.4 (8.1)	28.3 (75.2)	38.4 (6.3)	20.4 (76.2)	59.3 (10.6)	25.9 (86.3)	52.4 (5.2)	31.3 (91.1)	46.0 (3.6)
$\text{Phase}_{\text{CNN}}$ [7]	8.7	16.6	44.7	21.3	40.1	27.1	35.4	18.4	49.4	24.0	44.1	30.0	38.6
$\mathcal{L}_{\text{RI}}$	7.3	63.7	9.8	72.7	7.7	77.4	6.7	92.1	2.9	93.5	2.4	93.9	2.3
$\mathcal{L}_{\text{RI+Mag}}$	7.3	67.8	8.6	77.1	6.5	81.4	5.6	92.6	2.6	93.8	2.2	94.2	2.1
$\mathcal{L}_{\text{RI+IPD}}$	7.3	79.7	3.9	86.4	**2.9**	89.3	2.5	92.7	2.3	93.7	1.9	94.0	1.8
$\mathcal{L}_{\text{RI+Mag+IPD}}$	7.3	**80.9**	**3.9**	**87.5**	**2.9**	**90.4**	**2.4**	**94.0**	**1.8**	**94.9**	**1.5**	**95.2**	**1.5**

m^3 and 1 m respectively. All possible DOA pairs of speaker and noise (not from the same direction) are considered for every T60 and SNR resulting in 37×36×5×3 = 19980 configurations.

For the test setup with recorded RIRs, we pick the center two microphones from [27] to match our training geometry. Diffuse noises for real recordings are generated with the TIMIT speech convolved with RIRs from all possible directions, i.e. 26 including 1 m and 2 m distances. Note that real recorded diffuse noises are generated differently from our training simulations. 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. The hyperparameters include the mini-batch size of 8 and an initial learning rate of 0.001 which decays by 0.98 for every two 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-second segments for each experiment.

IV. EVALUATIONS AND COMPARISONS

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

$$\text{ACC } (\%) = \frac{\sum_t D_t C_t}{\sum_t D_t} \times 100 \tag{12}$$

$$\text{MAE } (^{\circ}) = \frac{\sum_t D_t |\theta_t - \widehat{\theta}_t|}{\sum_t D_t} \tag{13}$$

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 [28]. C_t = 1 if and only $|\theta_t - \widehat{\theta}_t| \leq \theta_{th}$, where $\theta_{th} = 5^{\circ}$ is commonly chosen as the tolerance threshold.

For speaker localization in reverberant and noisy environments, we compare with three DNN-based methods, GCC-PHAT and oracle mask-weighted GCC-PHAT (MGCC). 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 mask-weighted GCC-PHAT. $\text{Mask}_{\text{BLSTM}}$ [12] is a two-layer BLSTM (bidirectional long short-term memory) network that is trained on all individual microphones to estimate the PSM, and then performs mask-weighted GCC-PHAT for utterance-level localization. To generate frame-level results of $\text{Mask}_{\text{BLSTM}}$, we modify mask-weighted GCC-PHAT similar to (11). SADOA [13] employs cascaded DNNs for enhancement and localization. We modify the first DNN to BLSTM for a fair comparison. The model is trained end-to-end for 100 epochs at the utterance level with a cross entropy loss. The enhanced mask is used to weigh the features internally within the localization module referred to as feature masking [13]. Note that, even though the model is trained at the utterance level, it outputs DOA intrinsically at the frame level. For frame-level results, we pick the DOA as the angle with the highest value. $\text{Phase}_{\text{CNN}}$ [7] uses a CNN for DOA classification with frame-level mixture phase spectrograms as input.

Table I provides the frame-level DOA results of the proposed model with different loss functions and the comparison methods on the two-microphone linear array with simulated RIRs in the presence of diffuse and point-source noise. From Table I, we observe that the DOA performance of GCC-PHAT directly applied on microphone signals is very poor, reflecting reverberant and noisy test conditions. By incorporating the weighting mechanism in $\text{Mask}_{\text{BLSTM}}$, the utterance-level DOA performance is improved significantly but its frame-level DOA performance is poor. The SADOA results are slightly better than those of $\text{Mask}_{\text{BLSTM}}$. $\text{Phase}_{\text{CNN}}$ relies solely on phase spectrograms, and it performs better than GCC-PHAT and shows comparable performance with other baselines. Despite using the IRM, oracle MGCC's frame-level performance remains limited.

The proposed approach utilizes the clean phase estimated from MIMO based speech enhancement along with a weighting mechanism for speaker localization. As demonstrated by the results in Table I, the DOA performance of our systems surpasses oracle MGCC and is much better than the DNN baselines of $\text{Mask}_{\text{BLSTM}}$, SADOA and $\text{Phase}_{\text{CNN}}$. The frame-level results of our model exceed even the utterance-level results of $\text{Mask}_{\text{BLSTM}}$ and SADOA.

We now compare the DOA performance of different loss functions in our proposed MIMO approach for multi-channel speech enhancement. Although, $\mathcal{L}_{\text{RI}}$ and $\mathcal{L}_{\text{RI+Mag}}$ achieve high DOA performance, incorporating inter-channel phase differences into the loss functions, i.e., turning $\mathcal{L}_{\text{RI}}$ 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, especially in the diffuse noise case and in terms of mean absolute error. Among the loss functions, $\mathcal{L}_{\text{RI+Mag+IPD}}$ performs the best. Fig. 2 illustrates the DOA results of the different loss functions for the diffuse and point-source noise conditions. 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 tolerant range and fluctuate less compared to $\mathcal{L}_{\text{RI+Mag}}$. The same observation holds when comparing $\mathcal{L}_{\text{RI+IPD}}$ and $\mathcal{L}_{\text{RI}}$.

Table II provides the DOA results, averaged over all SNRs, T60 values, and source-to-microphone distances (see Sect. 3), with the recorded RIRs from [27] that match the trained two-channel array geometry. Our analysis for each speaker direction indicates that the endfire directions, i.e. $\{0, 180\}^{\circ}$, contribute to the majority of errors and are excluded from the evaluation as done in [12]. The results in Table II demonstrate that our approach generalizes well to real

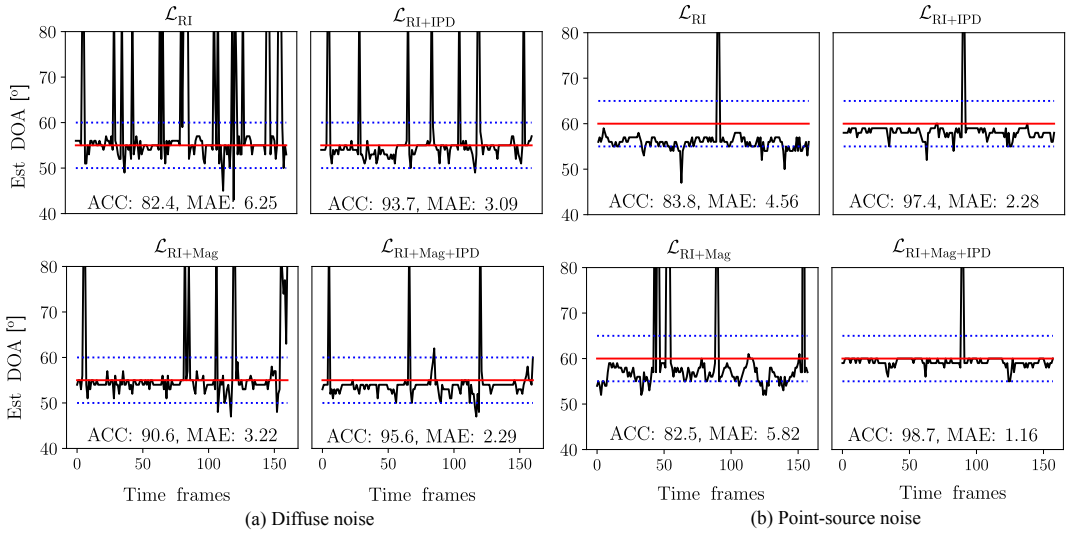

Fig. 2. Frame-level DOA estimation comparisons of different loss functions in a simulated environment with SNR = -5 dB and T60 = 1 s. The red solid and blue dotted lines represent the ground truth DOA and tolerance range respectively. The corresponding ACC and MAE results are shown in each case. Non-speech frames are excluded from the plot.

TABLE II
DOA RESULTS WITH RECORDED RIRS IN DIFFUSE NOISE AND POINT-SOURCE NOISE CONDITIONS.

Method/Loss	Diffuse noise		Point-source noise	
	ACC	MAE	ACC	MAE
GCC-PHAT	23.7	40.8	11.9	54.6
$\text{Mask}_{\text{BLSTM}}$ [12]	27.6 (74.2)	36.6 (8.3)	20.6 (54.0)	44.5 (21.5)
SADOA [13]	23.8 (63.6)	49.6 (19.3)	21.9 (58.0)	60.6 (20.7)
$\text{Phase}_{\text{CNN}}$ [7]	20.1	59.1	18.2	49.9
$\mathcal{L}_{\text{RI}}$	42.6	30.4	86.1	3.2
$\mathcal{L}_{\text{RI+Mag}}$	59.0	17.1	86.1	3.1
$\mathcal{L}_{\text{RI+IPD}}$	60.5	18.8	85.5	3.2
$\mathcal{L}_{\text{RI+Mag+IPD}}$	**64.9**	**16.6**	**86.2**	**3.0**

recordings and achieves much higher localization accuracy and lower MAE than the comparison baselines for both diffuse noise and point-source noise conditions. Like in Table I, the frame-level results of our model are better than the utterance-level results of the comparison baselines.

TABLE III
DOA RESULTS WITH SIMULATED RIRS IN DIFFUSE NOISE USING A CIRCULAR ARRAY.

Method/Loss	-5 dB		0 dB		5 dB	
	ACC	MAE	ACC	MAE	ACC	MAE
GCC-PHAT	14.6	71.2	21.5	61.8	30.4	51.3
Oracle MGCC	40.6	39.1	50.9	30.9	60.8	24.2
$\text{Mask}_{\text{BLSTM}}$ [12]	22.5	58.4	32.5	47.6	42.7	38.7
SADOA [13]	33.7	49.3	44.9	40.8	53.6	35.1
$\text{Phase}_{\text{CNN}}$ [7]	32	50.9	45.2	39.0	56.8	30.0
$\mathcal{L}_{\text{RI}}$	83.2	7.9	88.6	5.2	91.2	4.0
$\mathcal{L}_{\text{RI+Mag}}$	89.4	3.8	93.7	2.4	95.5	1.7
$\mathcal{L}_{\text{RI+IPD}}$	**96.1**	**1.7**	**97.4**	1.2	97.8	1.0
$\mathcal{L}_{\text{RI+Mag+IPD}}$	94.8	1.8	97.2	**1.1**	**98.0**	**0.8**

We further evaluate on the 7-microphone circular array our proposed method and the $\mathcal{L}_{\text{IPD}}$ loss for DOA and compare with other baselines in the simulated diffuse noise condition. The evaluation results are given in Table III. Consistent with the results on the 2-microphone linear array, the results in Table III demonstrate the superior performance of the proposed method over all the baselines. A comparison between the last two rows clearly shows that incorporating the $\mathcal{L}_{\text{IPD}}$ loss significantly enhances localization performance, demonstrating its utility across different array geometries.

As the localization results in Table 1 are based on multi-channel speech enhancement, their corresponding enhancement results are provided in Table IV where 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) [29], perceptual evaluation of speech quality (PESQ) [30], and scale-invariant signal-to-distortion ratio (SI-SDR) [31]. From Table IV, we notice that the inclusion of the IPD term in a loss function produces close speech enhancement results. However, the term clearly leads to better DOA results. This demonstrates that enhancement and localization are different tasks, justifying our two-stage approach. Overall, $\mathcal{L}_{\text{RI+Mag+IPD}}$ yields the best localization and speech enhancement performance.

TABLE IV
SPEECH ENHANCEMENT RESULTS IN DIFFUSE NOISE AND POINT-SOURCE NOISE CONDITIONS.

Loss	Diffuse noise			Point-source noise		
	STOI (%)	PESQ	SI-SDR	STOI (%)	PESQ	SI-SDR
Unprocessed	54.1	1.31	-9.1	58.4	1.39	-8.9
$\mathcal{L}_{\text{RI}}$	88.9	2.23	10.6	**97.3**	3.65	**16.2**
$\mathcal{L}_{\text{RI+Mag}}$	89.6	2.68	10.3	**97.3**	3.79	15.9
$\mathcal{L}_{\text{RI+IPD}}$	88.9	2.33	10.7	96.3	3.57	15.0
$\mathcal{L}_{\text{RI+Mag+IPD}}$	**89.8**	**2.70**	**10.7**	**97.3**	**3.80**	16.0

V. CONCLUSION

We have proposed novel MIMO complex spectral mapping models that are trained with the inclusion of inter-channel phase difference in loss functions. These models are then combined with a weighted GCC-PHAT method for frame-level speaker localization. The proposed approach yields accurate speaker localization in reverberant and noisy conditions and performs much better than related DNN based methods.

REFERENCES

[1] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, pp. 320–327, 1976.

[2] J. DiBiase, H. Silverman, and M. Brandstein, *Robust localization in reverberant rooms*. Berlin, Germany: Springer, 2001, pp. 157–180.

[3] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *J. Acoust. Soc. Amer.*, vol. 152, pp. 107–151, 2022.

[4] C. Pang, H. Liu, and X. Li, "Multitask learning of time-frequency CNN for sound source localization," *IEEE Access*, vol. 7, pp. 40 725–40 737, 2019.

[5] Shimada, K. Koyama, Y. Takahashi, S. N. Takahashi, and Y. Mitsufuji, "ACCDOA: Activity-coupled cartesian direction of arrival representation for sound event localization and detection," in *Proc. ICASSP*, 2021, pp. 915–919.

[6] S. Chakrabarty and E. A. P. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *Proc. WASPAA*, 2017, pp. 136–140.

[7] E. Vargas, J. R. Hopgood, K. Brown, and K. Subr, "On improved training of CNN for acoustic source localisation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 720–732, 2021.

[8] A. Bohlender, A. Spriet, W. Tirry, and N. Madhu, "Exploiting temporal context in CNN based multisource DOA estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1594–1608, 2021.

[9] B. Yang, X. Li, and H. Liu, "Supervised direct-path relative transfer function learning for binaural sound source localization," in *Proc. ICASSP*, 2021, pp. 825–829.

[10] J. Pak and J. W. Shin, "Sound localization based on phase difference enhancement using deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, pp. 1335–1345, 2019.

[11] P. Pertilä and E. Cakir, "Robust direction estimation with convolutional neural networks based steered response power," in *Proc. ICASSP*, 2017, pp. 6125–6129.

[12] 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, pp. 178–188, 2019.

[13] W. Mack, J. Wechsler, and E. Habets, "Signal-aware direction-of-arrival estimation using attention mechanisms," *Computer Speech & Language*, vol. 75, p. 101363, 2022.

[14] Z. Q. Wang and D. L. Wang, "Deep learning based target cancellation for speech dereverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 941–950, 2020.

[15] 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.

[16] T. Kabzinski and E. A. P. Habets, "A least squares narrowband DOA estimator with robustness against phase wrapping," in *Proc. EUSIPCO*, 2019, pp. 1–5.

[17] V. Tokala, E. Grinstein, M. Brookes, S. Doclo, J. Jensen, and P. A. Naylor, "Binaural speech enhancement using deep complex convolutional transformer networks," in *Proc. ICASSP*, 2024, pp. 681–685.

[18] C. H. Olivan, M. Delcroix, T. Ochiai, N. Tawara, T. Nakatani, and S. Araki, "Interaural time difference loss for binaural target sound extraction," in *Proc. IWAENC*, 2024.

[19] C. Quan and X. Li, "SpatialNet: Extensively learning spatial information for multichannel joint speech separation, denoising, and dereverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 1310–1323, 2024.

[20] 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, pp. 483–492, 2016.

[21] W. Fu, T. Y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *Proc. MLSP*, 2017, pp. 1–6.

[22] K. Tan and D. L. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 380–390, 2019.

[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[24] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *J. Acoust. Soc. Amer.*, vol. 133, pp. 3591–3591, 2013.

[25] S. Garofolo, "DARPA TIMIT acoustic-phonetic continuous speech corpus," *Nat. Inst. Standards Technol.*, 1993.

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

[27] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. IWAENC*, 2014, pp. 313–317.

[28] J. Wiseman, "Wiseman/py-webrtcvad," 2019.

[29] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, pp. 2125–2136, 2011.

[30] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, 2001, pp. 749–752.

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