IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

Microphone Pair Training for Robust Sound Source Localization with Diverse Array Configurations

Inkyu An[1], Guoyuan An[2], Taeyoung Kim[3], and Sung-eui Yoon[2]

***Abstract*—We present a novel sound source localization method that leverages microphone pair training, designed to deliver robust performance in various real-world environments. Existing deep learning (DL)-based approaches face scalability issues when dealing with various types of microphone arrays. To address these issues, our approach has been structured into two training steps: the first step focuses on microphone pair training, while the second step is designed for array geometry-aware training. The first training step enables our model to learn from multiple datasets covering various real-world situations, allowing it to robustly estimate the time difference of arrival (TDoA). Our robust-TDoA model incorporates a Mel scale learnable filter bank (MLFB) and a hierarchical frequency-to-time attention network (HiFTA-net). This allows it to effectively learn from various situations in multiple datasets, including those involving simultaneous sources and various sound events. The second training step enables our approach to estimate the direction of arrival (DoA) of sound based on TDoA information computed by our robust-TDoA model, which begins with parameters acquired during the first training step. During this process, our approach can be trained to accommodate geometry information of the target microphone array, which can span diverse array types. As a result, our method demonstrates robust performance across two DoA estimation tasks using three different types of arrays.**

I. INTRODUCTION

Sound source localization (SSL), also known as direction of arrival (DoA) estimation, is a fundamental problem in the field of robot audition [1]. Numerous studies have been dedicated to solving this SSL problem by leveraging various signal processing techniques, including MUSIC methods [2], [3] and beamforming methods [4].

In real-world scenarios, robots encounter complex situations, including noise and simultaneous sound events, which pose challenges for traditional signal processing methods. To address these challenges, several deep learning (DL)-based methods have been introduced. He *et al.* [5] introduced a method for the localization of simultaneous speech sources using a multi-layer perceptron. Additionally, Wang *et al.* [6] improved speech localization performance in noisy conditions by creating a trainable mask to reduce noise impact.

Many studies have also focused on the task of sound event localization and detection (SELD). Adavanne *et al.* [7] proposed a convolutional recurrent neural network (CRNN) to perform the SELD task with simultaneous sources. Schymura *et al.* [8] introduced a self-attention-based network to enhance the performance of the SELD task.

However, existing deep learning-based methods are typically tailored for specific types of microphone arrays. This design choice can lead to issues when different types of microphone arrays are used. Datasets used for training and testing must be acquired using a consistent type of microphone array. Therefore, these methods often face challenges in obtaining a sufficient amount of data for a specific array type. Moreover, models trained with a specific type of microphone array may not effectively adapt to other array types. We define these challenges as scalability issues in existing DL-based methods. Given the diversity of microphone arrays employed in robots, there is an increasing demand for solutions to overcome these scalability issues.

Several audio datasets have been released for SSL, including TUT-CA [7], DCASE2021 [9], and SSLR [5]. However, their combined usage is limited by scalability issues related to the various microphone array types used during recording. These datasets contain a wide range of situational audio data. For instance, the TUT-CA dataset includes various sound events recorded in an anechoic chamber, DCASE2021 captures dynamic sources in real, reverberating indoor environments, and SSLR collects human conversations in real environments amidst noise interference. Therefore, the potential for improving SSL performance through the combined utilization of these datasets is significant.

Main contribution. To address scalability issues associated with different types of microphone arrays, we propose a two-step training methodology. In the first step, we conduct microphone pair training (Sec. II-A), allowing our model to access multiple datasets collected by a range of microphone array types for the time difference of arrival (TDoA) estimation. Therefore, our model gains exposure to various situations present in these datasets. Additionally, we develop a robust TDoA model, which includes a Mel scale learnable filter bank (MLFB) (Sec. II-A1) and a hierarchical frequency-to-time attention network (HiFTA-net) (Sec. II-A2). This model is explicitly designed to learn from a variety of situations presented in multiple datasets.

Following the first step, our method performs the second step, array geometry-aware training (Sec. II-B), for DoA estimation (SSL). At this stage, our method can be trained to consider the geometry information of the target microphone array, accommodating diverse array types.

Thanks to our two-step training, our approach demonstrates

This work was supported by IITP grant funded by the Korea government(MSIT) (RS-2023-00237965). *(Corresponding author: Sung-eui Yoon)*

[1]Inkyu An is with the Integrated Intelligence Research Section, Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea inkyu.an@etri.re.kr

[2]Guoyuan An and Sung-eui Yoon are with School of Computing, Korea Advanced Institute of Science and Technology, Daejeon, South Korea anguoyuan@kaist.ac.kr, sungeui@kaist.edu

[3]Taeyoung Kim is with the Center for Artificial Intelligence, Korea Institute of Science and Technology, Seoul, South Korea ty.kim@kist.re.kr

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

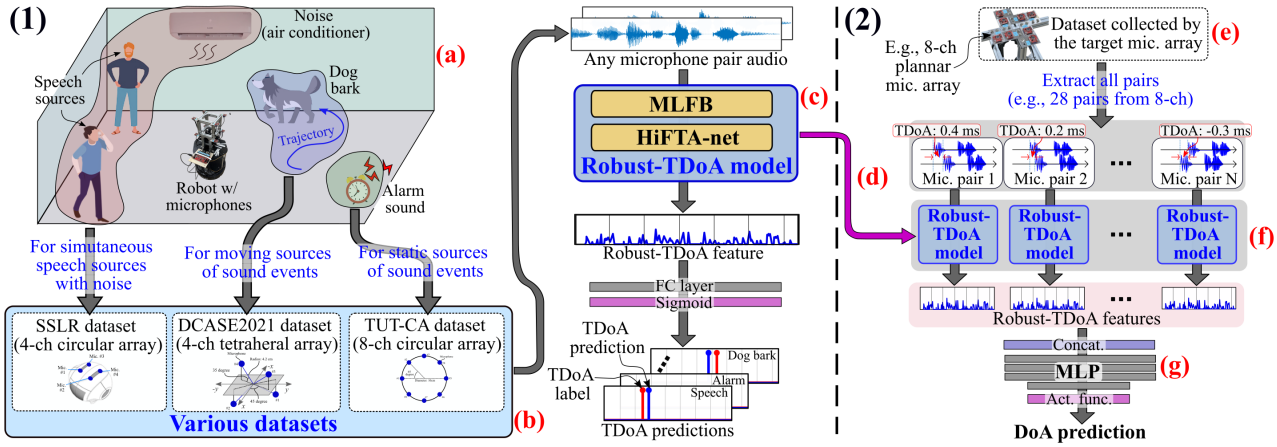

Fig. 1. The overview of our two-step training pipeline, which includes microphone pair training (1) and array geometry-aware training (2). In first microphone pair training (1), we employ our robust-TDoA model (c), which comprises a Mel scale learnable filter bank (MLFB) and a hierarchical frequency-to-time attention network (HiFTA-net). This model is trained to predict time differences of arrival (TDoAs) for sound events using audio data from microphone pairs. Microphone pair audios are extracted from various datasets (b), covering different scenarios (a), including simultaneous speech sources with noise, simultaneous static sound events, and simultaneous moving sound events. After microphone pair training, our approach performs second array geometry-aware training (2) to predict a direction of arrival (DoA). Our approach starts by extracting audio data from all microphone pairs within the target microphone array (e). It then utilizes the robust-TDoA model (f) to compute TDoA information encoded in robust-TDoA features. It is worth noting that the parameters of the robust-TDoA model (d) are initialized with those acquired during the first microphone pair training, ensuring robust TDoA computation for real-world scenarios (a). Subsequently, a multi-layer perceptron (MLP) (g) is trained to predict DoAs using the robust-TDoA features, incoporating geometry information of the target microphone array.

robust performance across various DoA estimation tasks, such as the localization of simultaneous speech sources (Sec. III-A, III-B, and III-D) and sound event localization and detection (SELD, Sec. III-E). Furthermore, through our array geometry-aware training, our approach is applicable to diverse types of microphone arrays, including the 4-ch circular array in the Pepper robot (Sec. III-A and III-B), the 8-ch planar array (Sec. III-D), and the Respeaker Mic Array v2 from Seeed (Sec. III-E). We also verified that our approach can work in real-time (Sec. III-F).

II. Robust DoA Estimation with Microphone Pair Training

Our training pipeline for robust direction of arrival (DoA) estimation is depicted in Fig. 1. During first microphone pair training, our method focuses on estimating the time difference of arrival (TDoA), which contains valuable information for DoA estimation [5]–[7], of sound events given two microphone signals. In second array geometry-aware training, our method leverages the estimated TDoA information to predict DoAs by learning the geometry of the target array.

This two-step training offers several benefits. It eliminates scalability issues related to microphone array types during the first step. TDoA estimation involves calculating the time difference between two coherent signals; therefore, microphone array types become irrelevant when our method estimates TDoA during the first step. Microphone pair training can utilize diverse datasets, detailed in Sec. I, collected with various microphone array types, thereby providing exposure to a wide range of real-world scenarios. After microphone pair training, our method can be applied to various microphone array types, leveraging array geometry-aware training, and it consistently demonstrates robust performance.

A. Microphone Pair Training for TDoA estimation

Some TDoA features [6], [10], [11] exist based on generalized cross correlation-phase transform (GCC-PHAT) [12]. However, existing TDoA features require improvement to function effectively in real environments with simultaneous sources and diverse sound events (see Sec. III-A, III-D, and III-E). We propose a robust-TDoA model designed to address challenges by learning from multiple datasets recorded using diverse array types. The robust-TDoA model takes microphone pair signals from multiple datasets and predicts TDoAs of sound events. The robust-TDoA model comprises two components: a Mel scale learnable filter bank (MLFB) for generating advantageous audio features and a hierarchical frequency-to-time attention network (HiFTA-net) for estimating TDoAs from these MLFB-generated features.

*1) **Mel scale learnable filter bank (MLFB)**:* In real environments, there exist diverse types of sound like speech, alarm, and dog bark, i.e., different sound events. Moreover, these sound events can occur simultaneously. The robust-TDoA model needs to be capable of distinguishing sound events while maintaining phase information for estimating TDoAs, even in cases of simultaneous sources. We first generate a useful audio feature for managing sound events in this section, then in Sec. II-A2, we estimate TDoAs while distinguishing between these events.

The Mel-spectrogram [13] serves as a useful hand-crafted feature for capturing different characteristics of various sound events. By converting the frequency scale to the Mel frequency scale, the Mel-spectrogram can show greater discriminative ability for low frequency and be advantageous in differentiating sound events, as many such events like speech, alarms, and dog barks dominate the low-frequency range. Nonetheless, the Mel-spectrogram is not suitable for computing TDoAs because it neglects a phase spectrum of short time Fourier transform

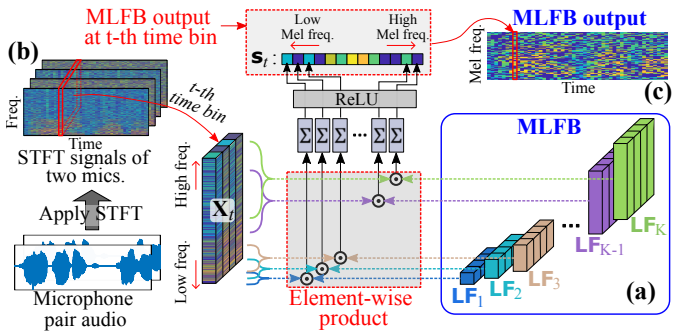

Fig. 2. An example of applying a Mel scale learnable filter bank (MLFB), consisting of K learnable filters (LFs), as depicted in (a), to STFT signals of two microphones, as illustrated in (b). Each LF consists of 4-ch learnable parameters and has a unique frequency bandwidth, notably, the frequency bandwidth of each LF becomes more narrow as the frequency decreases. Each LF is utilized on the selectively cropped frequency signal present at the t-th time bin within the STFT signals. As a result, the processed output from the k-th LF subsequently becomes the k-th value of the MLFB output (c) at the respective t-th time bin.

(STFT) signals. A phase difference of two microphone signals plays an key role in the TDoA estimation [11], [12], but the Mel-spectrogram is computed only from a magnitude spectrum of STFT signals.

We propose a Mel scale learnable filter bank (MLFB) that generates effective audio features capable of both computing TDoAs and distinguishing between sound events. Unlike the Mel-spectrogram, our MLFB is designed to take into account the phase difference of two microphone signals, thereby accepting real and imaginary parts of the STFT signals as input. The MLFB is composed of K learnable filters (LFs), which facilitate the conversion of the frequency scale into the Mel frequency scale.

The process of applying MLFB to a pair of microphone audio inputs is illustrated in Fig. 2. Our method begins by applying the short-time Fourier transform (STFT) to the microphone pair audio, resulting in STFT signals which include both the real and imaginary parts from the two microphones. The K LFs are then applied to a frequency signal at each time bin of the STFT signals. The process of filtering through MLFB for each time bin of the STFT signals can be expressed as:

$$s_t[k] = \text{ReLU}(\Sigma(\mathbf{X}_t \odot \mathbf{LF}_k)), \tag{1}$$

where $s_t[k]$ represents the k-th value of the MLFB output at the t-th time bin, ReLU is the rectified linear unit, Σ denotes the function summing all elements, $\odot$ is the element-wise product, $\mathbf{X}_t$ is the cropped frequency signal at the t-th time bin of the STFT signals, and $\mathbf{LF}_k$ is the k-th LF. Our method applies (1) across all time bins to produce the comprehensive MLFB output.

To convert the frequency scale to the Mel frequency scale, each LF is designed to possess different frequency bandwidths [14]. LFs for low frequencies have narrower bandwidths compared to those for high frequencies. As a result, when each LF is applied to the cropped frequency signal $\mathbf{X}_t$, which has the same frequency bandwidth, the MLFB output can be more discriminative for low frequencies.

Furthermore, each LF is designed with learnable parameters having 4-channel. Each channel corresponds to real and imaginary components of the STFT signals of two microphones. The learnable parameters of each LF can be trained to consider the phase difference between the two microphones during the scale microphone pair training process. A ReLU function in (1) is employed to facilitate efficient LF training.

Through the implementation of multiple MLFBs, our approach can generate N MLFB outputs, thereby increasing the model's learnable parameters and subsequently its ability to handle diverse situations found in real-world environments.

*2) **Hierarchical frequency-to-time attention network**:* In this section, we introduce a hierarchical frequency-to-time attention network (HiFTA-net). This model is designed to estimate time differences of arrival (TDoAs) between two microphones while simultaneously distinguishing different sound events from N MLFB outputs.

Different sound events inherently carry unique frequency and temporal characteristics. In an intuitive sense, each sound event consists of sequential elements unfolding over time (i.e., temporal characteristics), and these components exhibit distinct frequency attributes. Consider, for instance, human speech, which comprises a sequence of phonemes, or the periodic pattern of footsteps caused by the alternating contact of two feet with the ground. The temporal aspects, such as the sequential arrangement of phonemes in speech or the rhythm of footfalls, differ greatly. In addition, both phonemes and the sounds generated by individual footsteps exhibit distinct frequency attributes.

In an attempt to estimate TDoAs from a microphone pair while distinguishing between sound events, our approach strives to comprehend and leverage these temporal and frequency characteristics inherent to sound events. This is particularly relevant in scenarios where multiple sound sources exist simultaneously. Our hypothesis is that by considering the unique frequency and temporal properties of each sound source, our method can effectively estimate the multiple TDoAs that originate from these simultaneous sound sources.

HiFTA-net operations are demonstrated in Fig. 3; we are inspired by the image-based neural architecture [15] when designing our HiFTA-net. Our method initially divides the N MLFB outputs into T uniform time frames and seeks to estimate the TDoA of each frame. The HiFTA-net hierarchically learns both frequency and temporal characteristics of sound events. The frequency-attention network (FA-net) initially learns the frequency characteristics of each time frame, followed by the temporal-attention network (TA-net) learning temporal characteristics across the T time frames.

To learn the frequency characteristics of each time frame, each frame of the N MLFB outputs is divided into Q frequency patches. These patches are then transformed into hidden features using a fully connected layer FC: $\text{FC}([\mathbf{p}_{(t,1)}, \cdots, \mathbf{p}_{(t,Q)}]) \rightarrow [\mathbf{h}_{(t,1)}, \cdots, \mathbf{h}_{(t,Q)}]$, where $\mathbf{p}_{(t,q)}$ is the flattened q-th frequency patch on the t-th time frame and $\mathbf{h}_{(t,q)}$ is the q-th hidden feature on the t-th time frame. These hidden features are input into the FA-net, which learns the relationships between hidden features of frequency patches,

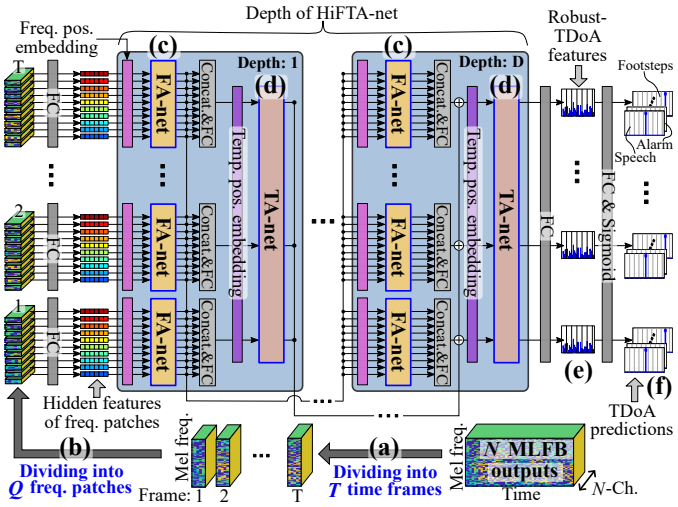

Fig. 3. An illustration of performing our HiFTA-net. Our method begins by dividing the N-channel MLFB output into T time frames, as shown in (a). Each time frame is further divided into Q frequency patches, as shown in (b). The HiFTA-net is designed to hierarchically comprehend both the frequency and temporal aspects inherent in the input, the MLFB output, derived from the divided frequency patches. A frequency-attention network (FA-net), presented in (c), initially learns the frequency characteristics within each time frame, followed by a temporal-attention network (TA-net), presented in (d), grasping the temporal properties spanning across all T time frames. Finally, from the output of the TA-net, our method generates robust-TDoA features (e) and calculates TDoA predictions (f) for sound events.

representing frequency characteristics of various sound events:

$$[\mathbf{h}'_{(t,1)}, \cdots, \mathbf{h}'_{(t,Q)}] = \text{FA-net}([\mathbf{h}_{(t,1)}, \cdots, \mathbf{h}_{(t,Q)}] + \mathbf{e}_f), \quad (2)$$

where $\mathbf{h}'_{(t,q)}$ is the q-th FA-net output on the t-th time frame and $\mathbf{e}_f$ is the frequency positional embedding (e.g., the sinusoidal signal [16]). FA-net is designed based on the self attention mechanism [15].

The outputs of the FA-net include frequency characteristics of each time frame. The temporal-attention network (TA-net) uses these outputs to learn the temporal characteristics across all time frames. The outputs of the FA-net at the t-th time frames are converted to another hidden feature: $\mathbf{f}_t = \text{FC}(\text{Concat}([\mathbf{h}'_{(t,1)}, \cdots, \mathbf{h}'_{(t,Q)}]))$, where FC is the fully connected layer and Concat is the function concatenating all elements. By repeating for all T time frames, our approach obtains the T hidden features, $[\mathbf{f}_1, \cdots, \mathbf{f}_T]$, which is used as the TA-net input:

$$[\mathbf{f}'_1, \cdots, \mathbf{f}'_T] = \text{TA-net}([\mathbf{f}_1, \cdots, \mathbf{f}_T] + \mathbf{e}_t), \quad (3)$$

where $\mathbf{f}'_t$ is the output of TA-net for the t-th time frame and $\mathbf{e}_t$ is the temporal positional embedding similar to $\mathbf{e}_f$ in (2). The TA-net is also based on the self attention mechanism, similar to the FA-net.

In order to accommodate more diverse scenarios in real environments, such as an increase in the number of distinct sound events to be discerned, we expand the trainable parameters of our HiFTA-net. This is achieved by stacking multiple HiFTA-nets, providing the depth D to the network. In this configuration, the output of the FA-net at the t-th time frame ($[\mathbf{h}'_{(t,1)}, \cdots, \mathbf{h}'_{(t,Q)}]$, in (2)) serves as the input for the subsequent FA-net at the same time frame. Furthermore, the outputs of the TA-net ($[\mathbf{f}'_1, \cdots, \mathbf{f}'_T]$, in (3)) are added to the input of the subsequent TA-net, which is the hidden features ($[\mathbf{f}_1, \cdots, \mathbf{f}_T]$).

The final step involves estimating TDoAs from the outputs of the last TA-net, which has the depth D and ideally should encapsulate the temporal and frequency characteristics of the input microphone pair audio:

$$\textbf{r-tdoa}_t = \text{FC}(\mathbf{f}'_t), \quad (4)$$

$$[\textbf{tdoa}^1_t, \cdots, \textbf{tdoa}^S_t] = \sigma(\text{FC}(\textbf{r-tdoa}_t)), \quad (5)$$

where $\textbf{r-tdoa}_t$ is our robust-TDoA feature at the t-th time frame, $\textbf{tdoa}^s_t$ is the TDoA prediction of the s-th sound event at the t-th time frame, and σ is the sigmoid activation function. We compute the BCE loss with a TDoA label for training our robust-TDoA model, which consists of the MLFB and HiFTA-net. The TDoA label is computed from the direction of arrival (DoA) labels provided by datasets, considering the positions of microphone pairs.

B. Array geometry-aware training

In this section, our objective is to localize sound sources by leveraging the estimated time difference of arrival (TDoA) information from a target microphone array. This target microphone array can encompass various types, as long as it satisfies the requirement that each direction of arrival (DoA) corresponds to a specific combination of TDoAs [17]. We can calculate DoAs based on the TDoAs from all microphone pairs within the target microphone array, while considering the array's geometry. In our case, we acquire TDoA information through the robust-TDoA model, as detailed in Sec.II-A. Consequently, in the second array geometry-aware training, as illustrated in Fig.1, our approach is trained to incorporate the geometry information of the target microphone array to estimate DoAs from TDoA information.

Initially, we extract audio from each microphone pair within the target microphone array. Subsequently, we utilize the robust-TDoA model to estimate TDoA information. The robust-TDoA model is initialized with parameters acquired during microphone pair training (Fig. II-A), and these parameters are shared across all microphone pairs. The TDoA information is then encoded in the robust-TDoA features, denoted as $[\textbf{r-tdoa}^1_t, \cdots, \textbf{r-tdoa}^P_t]$ at the t-th time frame given P microphone pairs. These robust-TDOA features are inputted into a multi-layer perceptron (MLP) aiming to predict the DoA, DoA_t:

$$\textbf{doa}_t = \Phi(\text{MLP}(\text{Concat}([\textbf{r-tdoa}^1_t, \cdots, \textbf{r-tdoa}^P_t]))), \quad (6)$$

where Φ is the activation function, MLP consists of four fully connected layers, and Concat is a function concatenating all elements. MLP is trained to account for microphone positions, i.e., array geometry information, during the array geometry-aware training.

By modifying MLP and the activation function Φ, we can reformat $\textbf{doa}_t$ to match the desired DoA representation. Examples include 3-D DoA vectors (x, y, and z) for various sound events in Sec.III-E and 2-D DoA angles (360 cells corresponding to 360 degrees while ignoring the elevation)

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

for speech in Sec.III-A and III-D. For these examples, we respectively use mean squared error and binary cross-entropy as the loss functions.

III. Result and Discussion

In this section, we evaluate the effectiveness of our approach and highlight its advantages. Firstly, we performed microphone pair training, as elaborated in Sec.II-A, to train our robust-time difference of arrival model (robust-TDoA model). This model was trained using three datasets: SSLR, TUT-CA, and DCASE2021, as introduced in Sec. I. These datasets encompass challenging situations such as instances with simultaneous sources of various sound events. Consequently, our approach, which operates based on this robust-TDoA model, delivers substantial performance in various direction of arrival (DoA) estimation tasks.

Additionally, we implemented an exisitng deep learning (DL)-based TDoA estimation method, i.e., DeepGCC [10] which was trained using the same three datasets. Our aim in this comparison is to show the effectiveness of the structural componenets of our robust-TDoA model in learning diverse situations found within the aforementioned datasets.

Secondly, following microphone pair training, our approach adapts to various types of microphone arrays for DoA estimation tasks through array geometry-aware training. These tasks include the multiple speaker localization and detection (MSLD), the localization of simultaneous speech sources, in Sec. III-A, III-B, and III-D as well as sound event localization and detection (SELD) in Sec.III-E. We compare our approach with previous methods used in MSLD [5], [6] and SELD [7]. We then compare our approach with the existing DL-based TDoA estimation method, DeepGCC [10], which also underwent array geometry-aware training following microphone pair training by replacing our robust-TDoA model to the DeepGCC model in Fig. 1.

Given the diversity of microphone arrays in different robots, adapting to various types of arrays is crucial. We have demonstrated our approach by applying three distinct target microphone arrays: the 4-ch circular array in the Pepper robot (Sec. III-A and III-B), the 8-ch planar array (Sec. III-D), and the Respeaker Mic Array v2 from Seeed (Sec. III-E).

Lastly, we validate the real-time functionality of our approach in Sec. III-F. We have included a demonstration video of our approach in the multimedia materials.

The input microphone pair audio has 1 s length, comprising 10 time frames, each lasting 0.1 s, with 24 kHz sample rate. The robot records audio for 1 s using the microphone array and then utilizes our method to estimate DoAs. If the audio length exceeds 1 s, the robot's response becomes slow. We believe that 1 s is not only appropriate for expecting a rapid response from robots but also sufficient to capture adequate time and frequency information. Other hyperparameters employed in our robust-TDoA model were selected by referencing prior DoA estimation methods [5], [9] and the image-based neural architecture [15]. For further details, please refer to the GitHub project (github.com/InkyuAn/MicPairTrain).

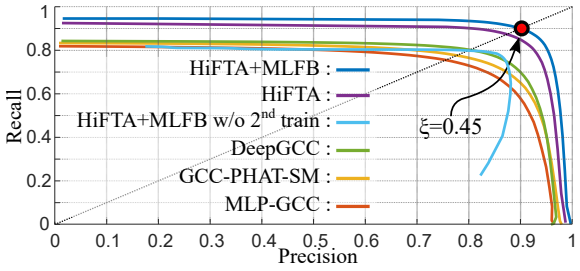

Fig. 4. The graph of *precision* vs. *recall* curves of different methods by varying the prediction threshold ξ, in the case of an unknown number of sources.

TABLE I
THE ACCURACY OF MSLD WITH EXISTING DATASET.

	Overall		*Source=1*		*Source=2*	
	MAE (↓)	ACC (↑)	MAE (↓)	ACC (↑)	MAE (↓)	ACC (↑)
MLP-GCC	9.65°	81.82%	9.46°	86.35%	10.16°	69.74%
GCC-PHAT -SM	7.41°	83.02%	6.88°	88.08%	8.84°	69.59%
DeepGCC	7.16°	84.63%	6.75°	89.12%	8.25°	72.66%
HiFTA (ours)	3.20°	93.95%	3.11°	95.19%	3.42°	90.64%
HiFTA+MLFB (ours)	**2.84°**	**94.94%**	**2.81°**	**95.91%**	**2.92°**	**92.34%**
HiFTA+MLFB w/o 2nd train (ours)	12.01°	81.63%	7.44°	88.95%	24.19°	62.13%

A. MSLD with existing dataset

Suppose the target microphone array for the multiple speaker localization and detection (MSLD) task is the 4-ch circular microphone array within the Pepper robot. Our approach then performs array geometry-aware training over 10 epochs using the SSLR dataset [5], collected using the same 4-ch circular microphone array. Other datasets like DCASE2021 and TUT-CA cannot be used in conjunction because they were recorded using different array types.

We compare our approach to previous MSLD methods, including *MLP-GCC* [5] and *GCC-PHAT-SM* [6], as well as the existing DL-based TDoA method, *DeepGCC* [10]. Additionally, we conduct an ablation study to highlight the advantages of the components in our method. Specifically, *HiFTA+MLFB* incorporates all components, while *HiFTA* utilizes only the HiFTA-net without MLFB. Moreover, we test our method in a different version, *HiFTA+MLFB w/o 2nd train*, to show the necessity of second array geometry-aware training. In this variant, *HiFTA+MLFB w/o 2nd train* predicts DoAs solely using TDoA prediction computed by our robust-TDoA model, which was trained during first microphone pair training without second array geometry-aware training. We implement this version by referring to the beamforming-based method [18].

We apply evaluation metrics previously suggested in the work [5] for the MSLD task, covering scenarios with known and unknown numbers of sources. The results of the first condition, i.e., the results with the known number of sources, are shown in Table. I. The mean absolute error (MAE) is the average azimuth estimation error, and the accuracy (ACC) denotes the percentage of correct azimuth estimates. We measure the results of both metrics under three conditions based on the

TABLE II
THE ACCURACY BY INCREASING THE SIZE OF THE TRAINING DATASET.

Training datasets for our robust -TDoA model	*Overall*		$\xi = 0.45$	
	MAE (↓)	ACC (↑)	Precision (↑)	Recall (↑)
SSLR	3.53°	91.27%	85.51%	86.45%
SSLR, DCASE2021	2.97°	93.63%	87.84%	88.94%
SSLR, DCASE2021 TUT-CA	**2.84°**	**94.94%**	**90.38%**	**89.95%**

number of overlapping sources: *Source=1* (a single source), *Source=2* (two simultaneous sources), and *Overall* (averaging *Source=1* and *Source=2*)

Our approach, *HiFTA+MLFB*, outperforms previous methods and our different versions across all situations and both metrics. Comparing *HiFTA+MLFB* with *MLP-GCC* and *GCC-PHAT-SM*, our two-step training strategy successfully estimates DoAs by learning from multiple datasets during the first training step. *DeepGCC*, performing the proposed two-step training, also shows better performance than *MLP-GCC* and *GCC-PHAT-SM*, thus we are convinced of the benefit of our two-step training strategy.

Comparing *HiFTA+MLFB* to *DeepGCC*, we show that our robust-TDoA model, comprising MLFB and HiFTA-net, excels at learning diverse situations from multiple datasets in the first step. Conversely, *DeepGCC*, which employs a distinct deep learning (DL) architecture based on a CNN-based encoder-decoder, underperforms in this aspect. Furthermore, both componenets, MLFB and HiFTA-net, in our robust-TDoA model perform well for estimating TDoAs, as evidenced by the superior performance of *HiFTA+MLFB* compared to *HiFTA*.

Comparing *HiFTA+MLFB* to *HiFTA+MLFB w/o 2nd train*, we confirm the necessity of second array geometry-aware training for more accurate DoA estimation. The performance of *HiFTA+MLFB w/o 2nd train* deteriorates, especially in *Source=2*, due to the low resolution of the TDoA predictions at low sampling rates (e.g., 24 kHz), which often place the TDoAs of simultaneous sources in the same delay bin. However, *HiFTA+MLFB* incorporates the robust-TDoA features, which are free from the resolution of the TDoA prediction, during the array geometry-aware training. This allows it to overcome this issue.

The results of the second type of metrics, i.e., the results with an unknown number of sources, are shown in Fig. 4. The *precision* vs. *recall* curve is proposed by varying the prediction threshold ξ [5] to verify the ability of detection as well as localization; the curve showing better results is closer to 1 value in both axes corresponding to *precision* and *recall*. Our approach, *HiFTA+MLFB*, shows the best performance among the reported results.

B. The ablation study with varying sizes of datasets

Our approach is evaluated by expanding the training datasets during the microphone pair training process. This experiment aims to ascertain the impact of learning from an extensive training dataset. We assess our method using various dataset sizes, incrementally adding the DCASE2021 and TUT-CA datasets to the SSLR dataset. This results in three versions: SSLR, (SSLR + DCASE2021), and (SSLR + DCASE2021 + TUT-CA).

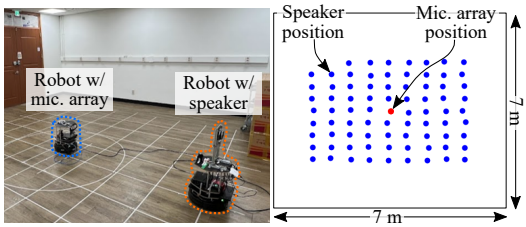

Fig. 5. The experiment for recording real RIRs in the left figure and the positions of robots equipped with the speaker (blue dots) and the microphone array (a red dot), respectively, in the right figure. The positions of robots are obtained using a SLAM technique [19] with 2D LiDAR and IMU sensors.

We utilize the MAE and ACC metrics of the *Overall* case, as well as the *precision* and *recall* values discussed in Sec. III-A. The *precision* and *recall* values have a trade-off relation according to the prediction threshold ξ. For our approach to work on the fly in a real-time, ξ needs to be chosen with a fixed value. We choose the proper value of ξ, which makes both precision and recall have large values. In Fig. 4, we assume that both *precision* and *recall* can be large at the intersection point between the *precision* vs. *recall* curve and the symmetric line, i.e., the dotted black line; thus, we utilize $\xi = 0.45$, i.e., the red point.

The results are shown in Table. II. The performance of our method improves across all metrics when the size of the training dataset is increased. These findings suggest that a large dataset is advantageous in our microphone pair training, implying that our robust-TDoA model, comprised of MLFB and the HiFTA-net, has sufficient capacity to learn from multiple datasets. Moreover, even in the case of just utilizing the SSLR dataset during the microphone pair training stage, the DoA estimation performance is bettern than the prior works, *MLP-GCC* and *GCC-PHAT-SM*, in Table. I.

C. Synthetic datasets for training and evaluating processes.

Our method is designed to operate with diverse types of microphone arrays mounted on various robots, made possible through our array geometry-aware training. Notably, when the dataset of the target microphone array for array geometry-aware training is not available, our method can effectively leverage a synthetic dataset generated by sound simulators. This involves initially generating room impulse responses (RIRs) using a sound simulator, then creating a synthetic dataset by convoluting RIRs with dry signals. We exclusively utilize dry speech signals for the MSLD task and all dry signals for the SELD task.

To generate synthetic datasets with two target microphone arrays, i.e., the ReSpeaker Mic Array v2 and 8-ch planar microphone array, we employ the Habitat 2.0 sound simulator [20] and a NIGENS general sound events database [21], which contains dry signals of fourteen sound events. The sound simulator can operate in a 3-D reconstructed environment using an iPhone equipped with a camera and LiDAR sensors, replicating a real environment of 7m width, 7m depth,

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

TABLE III
THE ACCURACY OF MSLD WITH SYNTHETIC DATASET OF THE 8-CH PLANAR ARRAY USING THE SOUND SIMULATOR.

	Overall		*Source=1*		*Source=2*	
	MAE (↓)	ACC (↑)	MAE (↓)	ACC (↑)	MAE (↓)	ACC (↑)
MUSIC	11.68°	78.02%	6.74°	79.61%	13.99°	77.27%
MLP-GCC	18.49°	63.08%	17.87°	64.95%	18.80°	62.14%
GCC-PHAT-SM	15.19°	62.72%	14.57°	65.70%	15.49°	61.22%
DeepGCC	11.02°	68.42%	9.33°	71.68%	11.86°	77.27%
Ours	**6.19°**	**88.67%**	**4.94°**	**91.5%**	**6.82°**	**87.26%**

TABLE IV
THE ACCURACY OF SELD WITH SYNTHETIC DATASET OF RESPEAKER V2 USING THE SOUND SIMULATOR.

	SED metrics		*DoA metrics*		*Overall*
	ER (↓)	F-score (↑)	DOA error (↓)	Frame recall (↑)	SELD score (↓)
SELDnet	0.67	47.92	28.77	66.21	0.42
DeeGCC	0.89	18.88	18.57	19.18	0.65
Ours	**0.45**	**69.79**	**11.90**	**73.43**	**0.27**

and 3m height. The synthetic dataset includes 5 hours of multi-channel audio with up to two simultaneous sources.

In addition, we record real RIRs with robots in an actual environment to verify our approach following the array geometry-aware training process using synthetic datasets from the sound simulator. We equip two mobile robots with each microphone array and a speaker, then record RIRs at 78 locations using sine sweeps [22], as illustrated in Fig. 5. We generate a 2-hour evaluation dataset using real RIRs, thereby making the evaluation dataset more realistic than the synthetic dataset from the sound simulator.

To mimic the noisy real-world conditions, we incorporate additional white noise; both our synthetic dataset, created using a sound simulator, and our evaluation dataset, utilizing real RIRs, have average signal-to-noise ratios (SNR) of 10 dB and 18 dB, respectively. Furthermore, the reverberation time (RT60s), which signifies the reverberation factor in our experimental environments, is 0.45 seconds for the synthetic dataset and 0.4 seconds for the evaluation dataset.

D. MSLD with synthetic dataset

When targeting the use of an 8-ch planar microphone array for the Multiple Speaker Localization and Detection (MSLD) task, we face a challenge: there are no published datasets recorded using this array. To address this, we employ the synthetic and evaluation datasets, both generated using the 8-ch planar microphone array (Sec. III-C) for the array geometry-aware training process. The DL-based TDoA method, *DeepGCC*, also undergoes array geometry-aware training using the same synthetic and evaluation datasets. The prior DL-based methods, *MLP-GCC* [5] and *GCC-PHAT-SM* [6], are modified to access the 8-ch audio of the synthetic dataset.

Additionally, we compare our approach with MUSIC [3], a signal processing-based technique. As MUSIC can function without a training process, it can easily be used when a training dataset is unavailable. To apply MUSIC, we must be aware of the number of active sources at each frame; we run MUSIC with the actual number of active sources, gleaned from the DoA labels. Therefore, the results of MUSIC in this paper can be seen as the optimal performance achievable of MUSIC. We use evaluation metrics, namely MAE and ACC, for the three situations: *Source=1*, *Source=2*, and *Overall*, as described in Sec. III-A.

The evaluation results are depicted in Table. III. Even though the sound simulator can generate a realistic synthetic dataset, it does not fully replicate real-world recorded data. Due to scalability issues related to microphone array types, prior DL-based methods, *MLP-GCC* and *GCC-PHAT-SM*, can only use the synthetic dataset for model training, resulting in suboptimal performances. The performance of these methods are inferior to *DeepGCC*, which employs the proposed two-step training, as well as *MUSIC*, the signal processing-based method. In contrast, our method exhibits superior performance by learning from multiple real datasets during the first microphone pair training step. Even when compared to *DeepGCC*, which was also trained from multiple real datasets during the first training step, our approach consistently outperforms it. Thus, our method with the robust-TDoA model demonstrates its effectiveness in learning realistic scenarios from diverse datasets.

E. SELD with synthetic dataset

When the target microphone array is the 4-ch Respeaker Mic Array v2, for which no published dataset exists, we use synthetic and evaluation datasets (Sec.III-C) for array geometry-aware training and performance evaluation. We also extend our approach to the Sound Event Localization and Detection (SELD) task, which requires differentiating sound events and is a different DoA estimation task than the experiments in Sec.III-A, III-B, and III-D.

This task demands the differentiation of sound events, making traditional signal processing-based methods like MUSIC [3] unsuitable. We compare our approach with a prior SELD method [7], which employs the GCC-PHAT feature and the Mel spectrogram within the CRNN model. This method is retrained using the synthetic dataset from the 4-ch Respeaker Mic Array v2 due to the discrepancy in the microphone array types used in its study [7] and our target microphone array. Scalability issues associated with the microphone array types prevent this method from utilizing other datasets, such as SSLR, DCASE2021, and TUT-CA, simultaneously.

For evaluation, we use metrics proposed in previous SELD work [7], consisting of five metrics: ER, F-score, DoA error, Frame recall, and SELD score. The ER and F-score are used to evaluate sound event detection (SED), whereas DoA error and Frame recall measure the performance of DoA estimation. The SELD score represents the overall performance by averaging the other four metrics. Detailed information about these metrics is available in [7].

As shown in Table. IV, our approach outperforms the prior works in all metrics, demonstrating its applicability to different SSL tasks with high performance, thanks to microphone pair training from multiple datasets. Conversely, *DeepGCC* was

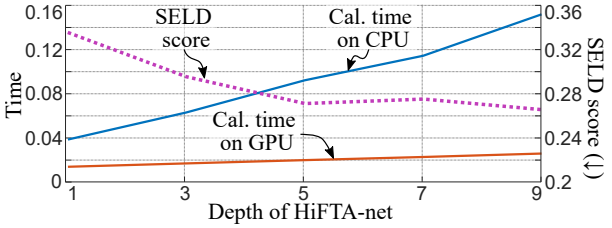

Fig. 6. The calculation times on CPU and GPU and the SELD scores of the SELD task by increasing the depth of our HiFTA-net.

not designed to differentiate between different sound events. While *DeepGCC* also utilizes microphone pair training, its performance is unsatisfactory due to its inability to distinguish between different sound events.

F. Real-time computation verification

So far, the results and analysis confirm the applicability of our approach to various microphone array types and diverse tasks. In this section, we aim to validate the feasibility of our approach for real-time application with actual robots. The key factor influencing computation time is the depth D of our HiFTA-net (Sec. II-A2). We tested our work in the SELD task with Respeaker Mic Array v2 in Sec. III-E by incrementally increasing the depth D from 1 to 9 in Fig. 6.

Our approach is computed in a laptop computer with Intel i7-11800H CPU and NVIDIA GeForce RTX 3070 Laptop GPU. However, even at maximum depth of 9, the computation time is approximately 0.15 s, validating that our approach can operate in real-time. Furthermore, we observe that the SELD score reaches saturation beyond a depth of 5. Therefore, in all our experiments, we set our model to have a depth of 5. We demonstrate the real-world applicability of our method with real robots by implementing it in an SSL application, for example, directing the robot to face the speech sources. Demonstration videos of these SSL applications can be found in our multimedia attachment.

IV. Conclusion

We have demonstrated that our approach is capable of robust DoA estimation across two SSL tasks using various microphone array types, thanks to our microphone pair training. We were unable to conduct detailed experiments analyzing the accuracy of DoA estimation for varying distances among simultaneous sources with similar frequencies. This is an area we plan to explore in future studies. In this paper, we focused on evaluating the performance of the robust-TDoA model for DoA estimation. Nevertheless, a more comprehensive analysis is needed, including an examination of the relationship between TDoA and DoA estimation performance. We anticipate that our method has the potential to handle indirect sounds, such as reflections and diffractions, by integrating it with ray tracing-based SSL methods [23]. To make our proposed method suitable for complex environments with diverse objects, we plan to extend it to accommodate a larger number of sound sources, exceeding three. Additionally, we believe that our method can be adapted to address open-world problems, including handling unseen sounds, through techniques such as model generalization, domain adaptation, or online learning.

References

[1] H. G. Okuno and K. Nakadai, "Robot audition: Its rise and perspectives," in *IEEE ICASSP*, 2015.
[2] K. Nakamura, K. Nakadai, F. Asano, Y. Hasegawa, and H. Tsujino, "Intelligent sound source localization for dynamic environments," in *IEEE/RSJ IROS*, 2009.
[3] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.
[4] C. Zhang, D. Florêncio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 538–548, 2008.
[5] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *IEEE ICRA*, 2018.
[6] J. Wang, X. Qian, Z. Pan, M. Zhang, and H. Li, "Gcc-phat with speech-oriented attention for robotic sound source localization," in *IEEE ICRA*, 2021.
[7] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2018.
[8] C. Schymura, B. Bönninghoff, T. Ochiai, M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, and D. Kolossa, "Pilot: Introducing transformers for probabilistic sound event localization," *arXiv preprint arXiv:2106.03903*, 2021.
[9] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection," *arXiv preprint arXiv:2106.06999*, 2021.
[10] J. M. Vera-Diaz, D. Pizarro, and J. Macias-Guarasa, "Towards domain independence in cnn-based acoustic localization using deep cross correlations," in *IEEE EUSIPCO*, 2021.
[11] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source tdoa estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, no. 8, pp. 1950–1960, 2012.
[12] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, 1976.
[13] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The journal of the acoustical society of america*, vol. 8, no. 3, pp. 185–190, 1937.
[14] X. Huang, A. Acero, H.-W. Hon, and R. Reddy, *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR, 2001.
[15] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," *arXiv preprint arXiv:2103.00112*, 2021.
[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, 2017.
[17] A. K. Tellakula, "Acoustic source localization using time delay estimation," *Degree Thesis. Bangalore, India: Supercomputer Education and Research Centre Indian Institute of Science*, 2007.
[18] J.-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robot. Auton. Syst.*, vol. 55, no. 3, 2007.
[19] W. Hess, D. Kohler, H. Rapp, and D. Andor, "Real-time loop closure in 2d lidar slam," in *IEEE ICRA*, 2016.
[20] A. Szot *et al.*, "Habitat 2.0: Training home assistants to rearrange their habitat," in *NeurIPS*, 2021.
[21] I. Trowitzsch, J. Taghia, Y. Kashef, and K. Obermayer, "The nigens general sound events database," *arXiv preprint arXiv:1902.08314*, 2019.
[22] A. Farina, "Advancements in impulse response measurements by sine sweeps," in *Audio engineering society convention 122*. Audio Engineering Society, 2007.
[23] I. An, Y. Kwon, and S.-e. Yoon, "Diffraction-and reflection-aware multiple sound source localization," *IEEE Transactions on Robotics*, 2021.