IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 April 2024, accepted 20 May 2024, date of publication 24 May 2024, date of current version 4 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3405174

RESEARCH ARTICLE

Direction-of-Arrival Estimation for Mobile Agents Utilizing the Relationship Between Agent's Trajectory and Binaural Audio

TOMOYA SATO, YUSUKE SUGANO, (Member, IEEE), AND YOICHI SATO, (Senior Member, IEEE)
Institute of Industrial Science, The University of Tokyo, Tokyo 113-8654, Japan
Corresponding author: Tomoya Sato (tomosato@iis.u-tokyo.ac.jp)

This work was supported by JST SPRING under Grant JPMJSP2108.

ABSTRACT With the development of robotics and wearable devices, there is a need for information processing under the assumption that an agent itself is mobile. Especially, understanding an acoustic environment around an agent is an important issue. In this paper, we solve a task in which a moving agent estimates the Direction of Arrival (DoA) of the surrounding sound sources. To this end, we propose a novel training method, Trajectory-based Direction Selection (TDS). In TDS, a mixture of binaural audio recorded by two agents and their trajectories are given as input to a network. Then, the network is trained to estimate the DoA of surrounding sounds that correspond to each agent's trajectory separately. By corresponding the agent's trajectory to the binaural audio with TDS, we can estimate DoAs of multiple sounds even with binaural audio as audio input, which has not been realized by sound-only methods. In simulated environments covering both single and multiple sources, our method outperforms existing DoA estimation methods.

INDEX TERMS Audio processing, direction of arrival estimation, embodied agents, multi-modal learning.

I. INTRODUCTION

Growing developments in robotics and wearable cameras have increased the demand for analyzing acoustic environments in scenarios where the devices are in motion. To address such situations, various sound-tracking methods have been proposed [1], [2], [3]. These sound-tracking methods estimate the Direction of Arrival (DoA) at each time in a situation where the positional relationship between the device and the sound source is continuously changing. However, these methods have the limitation that they assume the localization of a smaller number of sound sources than the number of microphones. This assumption necessitates the use of specialized or costly equipment, such as microphone arrays, to effectively analyze complex environments in which multiple sound sources are involved.

On the other hand, focusing on the fact that most devices are equipped with multiple sensors such as cameras and microphones, various models utilizing multi-modal input have been proposed [4], [5], [6], [7], [8], [9], [10], [11], [12]. These methods include those that leverage the movement of the devices themselves to analyze scenes with multiple sound sources, even with devices like binaural microphones that have insufficient spatial information. For example, there are navigation models guiding the agent to multiple target sound sources [7], as well as models searching for the ''sweet spot'' where each sound is easily distinguishable [8], [9].

However, these multi-modal models for autonomous agents are applied for limited scenarios. Specifically, in these multi-modal models, agent movements are specialized in localizing sound sources, and it is difficult to realize applications in which an agent solves multiple tasks simultaneously. For example, multi-modal navigation models employ reinforcement learning techniques to guide agents incrementally toward positions where they can identify the desired sound source. These methods are specialized in localizing sound sources, and as a result, an agent may struggle to estimate the DoA of surrounding sound sources while moving toward a specific destination. To overcome this limitation, we need a multi-modal model capable of

The associate editor coordinating the review of this manuscript and approving it for publication was Alessandro Floris.

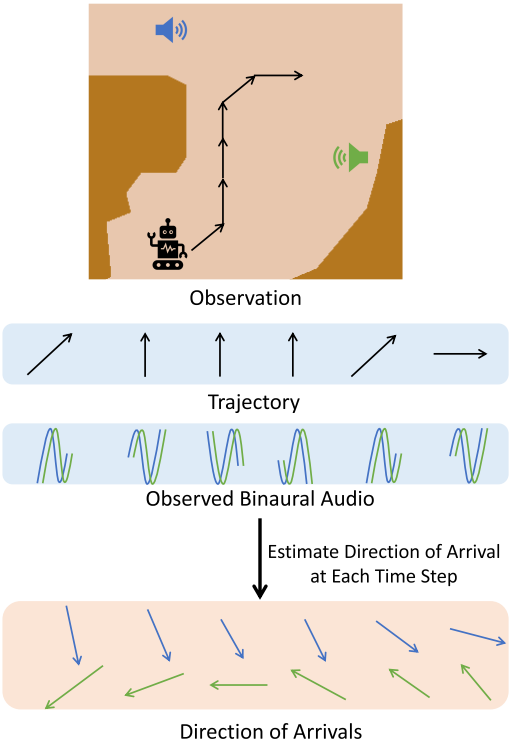

FIGURE 1. Overview of our task. We address the situation in which an embodied agent understands the status of sound sources while moving.

accommodating agents that move freely, without specializing in DoA estimation. In this work, we assume that the agent not only moves freely but also continuously observes the surrounding acoustic environment, akin to sound-tracking methods.

Our goal is to develop a model that enables a continuously moving agent to understand the surrounding acoustic environment. Specifically, we address the DoA estimation depicted in Figure 1. In this task, we have observations of agents moving through acoustic environments toward a certain destination. We input the agent's trajectory and the binaural audio obtained from the observation, and estimate the DoA of sound sources at each time step. Note that we assume that the agent has only a camera and a binaural microphone as sensors, but even in this case, we suppose that the agent's trajectory can be obtained using techniques such as Visual SLAM [13], [14]. The proposed model utilizes the agent's trajectory to achieve DoA estimation for multiple sound sources with binaural microphones, which has been challenging with methods that use only sound as input. Furthermore, the problem setting in the proposed method assumes that agents move freely through the acoustic environment, rather than moving to solve a specific task, which means that we relax the limitations of existing multi-modal methods.

Our key idea is twofold: i) applying multi-modal input of trajectory and binaural audio to the DoA estimation model, and ii) applying a combination of multi-modal tasks to the training of the model with the trajectory and binaural audio. First, an agent's trajectory is a critical factor in DoA estimation for moving agents. This is because the DoA of fixed sound sources observed by a moving agent changes according to the agent's movement, i.e., trajectory. For example, in Figure 1, as the agent moves diagonally forward, the DoA of the blue sound changes to the right, and the DoA of the green sound changes to the front. To this end, i) we leverage the multi-modal input of agent trajectory and binaural audio to improve the accuracy of DoA estimation.

Furthermore, we ii) effectively learn the cross-modal correspondence between the agent's trajectory and binaural audio based on a combination of multi-modal tasks in audio-visual learning. Specifically, we focus on the fact that the combination of audio-visual localization and audio-visual separation improves the performance of each task [15], [16]. Here, audio-visual localization is a task that highlights an image region corresponding to a sound input, while audio-visual separation is a task that distinguishes between audio related to the image input and irrelevant audio. This suggests that utilizing a task that distinguishes audio corresponding to the other input from irrelevant audio, in addition to a task that simply correlates audio to the other input, will emphasize the audio consistent with the other input. Inspired by these works, we propose a novel training method for DoA estimation model shown in Figure 2.

Figure 2 shows an overview of our proposed method for training a DoA estimation model with a moving agent. We train the DoA estimation model not only by (a) applying multi-modal inputs from the agent's trajectory and binaural audio to the model, but also by (b) a newly proposed Trajectory-based Direction Selection (TDS). TDS is inspired by the Mix-and-Separate approach [17], a common method for audio-visual separation. In TDS, we have trajectories and binaural audio from videos of two different moving agents. First, (1) we mix the two binaural audio. Then, (2) the mixed audio and one of the two trajectories are fed into the DoA estimation model, which estimates only the DoA consistent with the given trajectory. The addition of TDS improves the performance of DoA estimation by effectively correlating the agent's trajectory with binaural audio, as in the case of the combination of audio-visual localization and separation.

The rest of this paper is organized as follows. In Section II, we review the related work and show the position of this work in multi-modal learning and DoA estimation. In Section III, we propose TDS that improves the performance of DoA estimation for continuous moving agents. In Section IV, we show the experiments of DoA estimation. Experimental results with simulator-generated data demonstrate that the proposed model outperforms prior models in both the single-source and multiple-source cases. Finally, in Section V, we conclude this paper.

II. RELATED WORK

A. AUDIO-VISUAL LEARNING

Audio-visual learning is a popular multi-modal learning framework in which we train models based on the

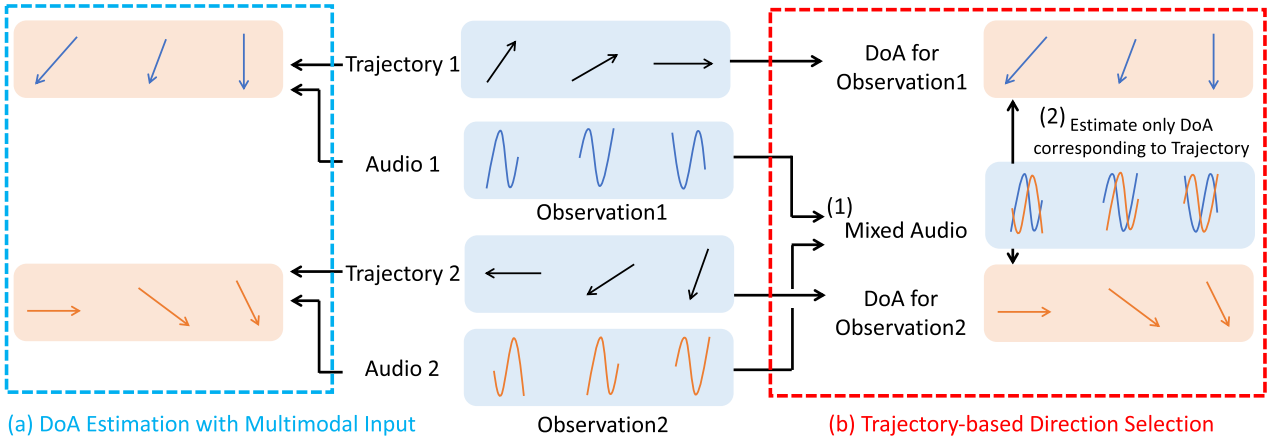

FIGURE 2. Overview of the proposed method. To train the DoA estimation model, we perform both (a) DoA estimation with multi-modal input of agent trajectories and binaural audio and (b) Trajectory-based Direction Selection (TDS).

correspondence between video and sound. Audio-visual learning leads to improving the performance of tasks conventionally solved with only a single modality such as action recognition [18], [19]. Also, audio-visual learning realizes novel tasks such as audio-visual localization and separation [16], [17], [20], [21], [22], [23]. For representing relationships between video and sound, various pretraining methods based on semantic correspondence [18], [20] and temporal synchronization [19], [24] have been proposed.

In recent years, several methods have been proposed to represent the audio-visual spatial relationship by adopting multi-channel sound input instead of monaural sound [25], [26], [27], [28], [29]. We note that most methods use stereo sound as multi-channel sound because of its low collection cost. By corresponding video and stereo sound, various audio-visual applications including vehicle detection [25], face detection in ASMR videos [26], and stereo sound generation [27], [28], [29] have been proposed.

While these audio-visual methods can utilize the relationship between video and multi-channel sound, they assume that devices such as cameras and microphones are placed in fixed positions. We address this limitation by constructing a DoA estimation model for continuously moving agents.

B. DoA ESTIMATION USING ONLY SOUND INPUT

DoA estimation models have traditionally been solved using only sound input. Beamforming with microphone arrays [30] and the MUSIC [31] are typical methods. With the development of deep learning, various models using neural networks have been proposed [32], [33]. Among them, convolutional neural networks are often adopted. Recently, sound localization methods based on the transformer [34] have been proposed [35], [36]. Moreover, models for moving sound sources have also been proposed [1], [2], [3]. For example, Adavanne et al. combined CNN and GRU to improve the DoA estimation performance of moving sound sources by considering time series changes [1]. 3D CNN and temporal convolutional networks have also been proposed as models for representing time series changes [2], [3].

While these methods effectively estimate DoA from sound input, they do not explicitly assume that an agent itself moves. We propose an effective sound source localization method for the continuous moving agents by correlating the agent's trajectories with the sounds.

In addition, previous DoA estimation methods have limitations on the number of sound sources to be localized. For example, MUSIC can only localize fewer sound sources than the number of microphones. Furthermore, most deep learning-based methods use multi-channel sounds to localize a single sound source [2], [32], [33], [36]. Some works have attempted to localize up to three sound sources, but they use four-channel inputs, called ambisonics [1], [3], [35]. Thus, existing DoA estimation addresses a smaller number of sound sources than the number of channels of multi-channel sound input. On the other hand, our method proposes to localize two and three sources from binaural audio utilizing the trajectory of the agent. Namely, the proposed method can localize more sound sources than the number of input channels.

C. MULTI-MODAL LEARNING FOR EMBODIED AGENTS

Multi-modal learning has enabled a variety of applications for embodied agents, i.e., autonomously moving agents. For example, there are audio-visual navigation, depth estimation, and camera pose estimation [4], [5], [6], [7], [8], [9], [10], [11], [12]. In particular, several methods have been proposed to process multiple sound sources by taking advantage of the agent's movement, even though the sound input is binaural [7], [8], [9]. For example, Kondoh and Kanezaki proposed audio-visual navigation that guides the agent to each sound source sequentially in an environment with multiple sound sources [7]. Although these methods can handle multiple sound sources by correlating the agent's movement to associated sound changes, the agent's movements need to

be specialized only to solve the task, and these multi-modal methods restrict the agent's behavior.

There are also methods for embodied agents that analyze a given agent's observations rather than manipulating the agent's movement [10], [11], [12], [37]. Chen et al. propose a method that improves the performance of DoA estimation by corresponding the agent's rotation to changes in DoA, although still working with a single sound source [37]. While these methods have been successful in improving task performance by taking advantage of agent observations, they are based only on observations at specific points and do not target freely moving agents.

We tackle a novel problem setting of estimating the DoA of multiple sound sources for binaural microphones, devices with insufficient spatial information, by corresponding the trajectory of a continuously moving agent to binaural audio.

III. PROPOSED METHOD

A. OVERVIEW

Our goal is to propose a DoA estimation method for a continuously moving agent utilizing multi-modal input of the agent trajectory and the binaural audio. In this paper, we assume that an agent moves around the acoustic environment where sound sources are fixed. To this end, we propose a network shown in Figure 3. Our idea is to emphasize the correspondence between the trajectory and the audio by combining (a) DoA estimation with trajectory and audio and (b) TDS, distinguishing between audio consistent with the trajectory and inconsistent audio. This is inspired by the combination of audio-visual localization and audio-visual separation in audio-visual learning. Specifically, (a) we estimate the DoA at each time step from each multi-modal pair. Then, (b) we incorporate a mixture of multiple binaural audio and each agent's trajectory as additional inputs. By having the model estimate only each DoA corresponding to each trajectory from the sound mixture, the model explicitly learns the correspondence between the agent's trajectory and the DoA. As a result, the model improves the accuracy of the DoA estimation when inputting pairs of the agent's trajectory and the binaural audio even in difficult cases where there are two or three multiple sources.

Let $\boldsymbol{D} = \{(\boldsymbol{p}_i, \boldsymbol{a}_i, \boldsymbol{d}_i), 1 \leq i \leq N\}$ be the dataset for training the proposed model, where $\boldsymbol{p}_i$, $\boldsymbol{a}_i$, and $\boldsymbol{d}_i$ are the i-th trajectory path, binaural audio, and DoA labels in the dataset, respectively. Here, N is the number of data in the dataset. $\boldsymbol{a}_i$ is a T-second binaural audio, which is divided into S segments of equal length for input into the model. In other words, the length of one segment is the T/S second, and $\boldsymbol{a}_i = \{\boldsymbol{a}_i^s, 1 \leq s \leq S\}$. Also, the agent's trajectory $\boldsymbol{p}_i = \{\boldsymbol{p}_i^s, 1 \leq s \leq S\}$ consists of S segments, where $\boldsymbol{p}_i^s = [x_i^s, y_i^s, \theta_i^s]$ contains the coordinates (x_i^s, y_i^s) and orientation θ_i^s of the agent at the $T \cdot s/S$ second. Note that $\boldsymbol{p}_i^1 = [0, 0, 0]$, i.e., $\boldsymbol{p}_i$ is calculated using the relative values of the agent's initial position and orientation as $(0, 0)$ and 0, respectively. When calculating d_i^s in $\boldsymbol{d}_i = \{d_i^s, 1 \leq s \leq S\}$, we assume K bins and M sound sources. We use $r_{s,i}^m$, representing the DoA of the m-th $(1 \leq m \leq M)$ sound source in radians at the $T \cdot s/S$ second. (i)When $M = 1$, d_i^s is a single class label as $d_i^s = \lfloor r_{s,i}^1 \cdot K/2\pi \rfloor$. (ii)When $M > 1$, d_i^s is a vector of class labels. Specifically, $d_i^s \in \{0, 1\}^K$. Here, the k-th element of d_i^s is 1 if $k \in \{\lfloor r_{s,i}^m \cdot K/2\pi \rfloor, 1 \leq m \leq M\}$ and 0 otherwise. The continuously moving agent dataset represented in this way is used to train the DoA estimation model described below.

B. TRAJECTORY-BASED DIRECTION SELECTION

In this paper, we propose Trajectory-based Direction Selection (TDS). First, we randomly select different observations $(\boldsymbol{p}_i', \boldsymbol{a}_i', \boldsymbol{d}_i')$ for the i-th observation $(\boldsymbol{p}_i, \boldsymbol{a}_i, \boldsymbol{d}_i)$ from the dataset. In this experiment, $(\boldsymbol{p}_i', \boldsymbol{a}_i', \boldsymbol{d}_i')$ is selected from observations that contain only a single sound source. For an additional loss in TDS, we introduce the sound mixture $\boldsymbol{m}_i = \boldsymbol{a}_i + \boldsymbol{a}_i'$ in the two observations and calculate

$$L_{mix} = \sum_{i=1}^{N} \sum_{s=1}^{S} (C(o^s(\boldsymbol{p}_i, \boldsymbol{m}_i), d_i^s) + C(o^s(\boldsymbol{p}_i', \boldsymbol{m}_i), d'^s_i)), \quad (1)$$

where C is the criterion of the loss function, and $o^s(\boldsymbol{p}, \boldsymbol{a})$ is the s-th DoA estimation model output of the sequence length S when trajectory $\boldsymbol{p}$ and binaural audio $\boldsymbol{a}$ are input. We similarly calculate the loss for the original agent trajectory and binaural audio pairs as

$$L_{original} = \sum_{i=1}^{N} \sum_{s=1}^{S} (C(o^s(\boldsymbol{p}_i, \boldsymbol{a}_i), d_i^s) + C(o^s(\boldsymbol{p'}_i, \boldsymbol{a}_i'), d'^s_i)). \quad (2)$$

Finally, the loss for TDS is calculated as

$$L = L_{mix} + L_{original}. \quad (3)$$

We show the pseudo code of this network training procedure in Figure 4.

C. NETWORK ARCHITECTURE

Our network shown in Figure 3 consists of an audio CNN and a transformer-based encoder. The audio CNN is based on ResNet18 [38]. Because the dimension of the sound input is four as described in Section III-D, the dimension of the first convolutional layer is changed from three to four. The sound feature after the fourth residual block is extracted and passed through a FC layer. Note that, as in the previous work [21], the stride in the fourth residual block is changed from two to one. The transformer encoder is constructed with the transformer encoder layer which has the same structure as the implementation in the original paper [34]. We construct the transformer encoder with 6 encoder layers and 16 heads. The features from these encoder layers are passed through FC layers to obtain the final output. Additionally, we apply the same positional encoding as in the original paper to the input of the transformer encoder.

Next, we describe how to obtain the network output. First, we detail the feature extraction of the agent trajectory.

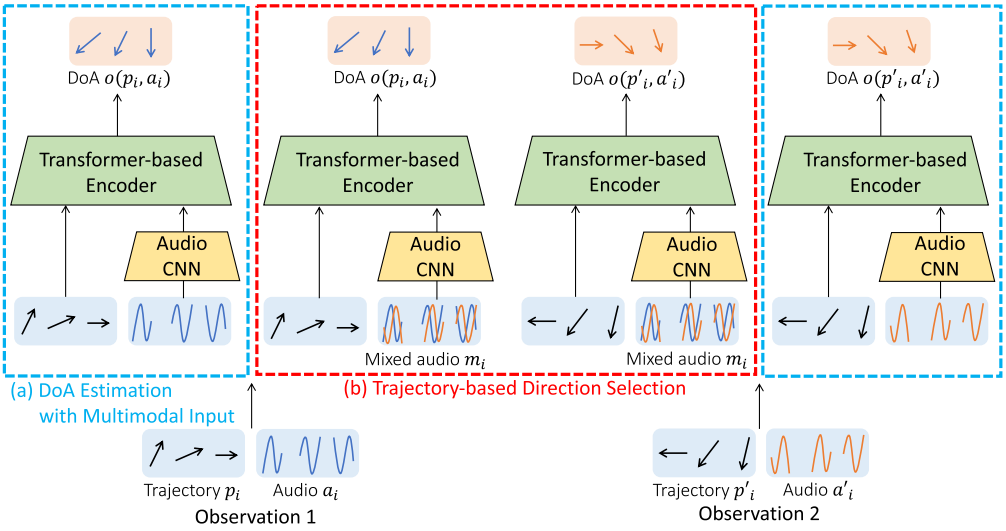

FIGURE 3. Proposed network for DoA estimation. The network consists of an audio CNN and a transformer-based encoder. We note that the weights of the audio CNN and the transformer encoder are shared.

for (p_i, a_i, d_i) **in** D**:**
$(p'_i, a'_i, d'_i) \xleftarrow{randomly} D$
$m_i = a_i + a'_i$ `#generate mixed audio`

$$
\begin{aligned}
L_{mix} &= \text{loss}(d_i, \text{model}(p_i, m_i)) \\
&\quad + \text{loss}(d'_i, \text{model}(p'_i, m_i)) \\
L_{original} &= \text{loss}(d_i, \text{model}(p_i, a_i)) \\
&\quad + \text{loss}(d'_i, \text{model}(p'_i, a'_i)) \\
L &= L_{mix} + L_{original}
\end{aligned}
$$

update model by backpropagating L

def model(p, a):
estimate DoA d from (p, a)
return d

FIGURE 4. Pseudo code for training the proposed network.

As described in section III-A, we represent the trajectory of the agent as coordinates and orientation at each time step. Following the previous work [37], we transform these values into high-dimensional vectors by positional encoding instead of directly inputting them. We transform each element of the agent trajectory $p_i^s = [x_i^s, y_i^s, \theta_i^s]$ into D' dimensional feature vector $\boldsymbol{f}_{i,s}^p = [PE(2\pi \cdot x_i^s, D'/4), PE(2\pi \cdot y_i^s, D'/4), PE(\theta_i^s, D'/2)]$, where the function $PE(x, D)$ is the positional encoding proposed in [34] that transforms the value x into a D-dimensional vector. Finally, the agent trajectory $\boldsymbol{p}_i = \{\boldsymbol{p}_i^s, 1 \leq s \leq S\}$ is transformed into $\boldsymbol{f}_i^p = \{\boldsymbol{f}_{i,s}^p, 1 \leq s \leq S\}$. We next describe feature extraction from the binaural audio. Each element a_i^s of the input of binaural audio $\boldsymbol{a}_i = \{a_i^s, 1 \leq s \leq S\}$ is transformed into a spectrogram and fed into the audio CNN to extract the sound feature $\boldsymbol{f}_{i,s}^a \in \mathbb{R}^{D'}$. Thus, we obtain the binaural audio feature $\boldsymbol{f}_i^a = \{\boldsymbol{f}_{i,s}^a, 1 \leq s \leq S\}$. Finally, we combine the agent trajectory feature $\boldsymbol{f}_i^p$ and the binaural audio feature $\boldsymbol{f}_i^a$ to obtain $\boldsymbol{f}_i \in \mathbb{R}^{S \times 2D'}$. We input $\boldsymbol{f}_i$ into the transformer-based encoder and pass it through a FC layer. Then, we have the network output $o(\boldsymbol{p}_i, \boldsymbol{a}_i) \in \mathbb{R}^{S \times K}$.

D. IMPLEMENTATION DETAILS

First, we describe the calculation of spectrograms from binaural audio. Following previous works [27], [28], [29], we calculate a spectrogram that preserves both the amplitude and phase information of the binaural audio. Specifically, we first apply a Short-Time Fourier Transform (STFT) to the left and right channels of the binaural audio. Here, instead of calculating the amplitude spectrogram from the norm of the real and imaginary parts, the real and imaginary parts are calculated as distinct spectrograms. As a result, each channel is converted into two spectrograms, resulting in four channels in total. The binaural audio is sampled as 16 kHz and the STFT is applied by a 25-ms Hann window with a 10-ms hop and an FFT size of 512. Therefore, the binaural audio a_i^s is converted into a spectrogram of size $4 \times 256 \times 100T/S$.

Finally, we show the training parameters. We set the input data length $T = 3.6$ and the sequence length $S = 24$, such that each feature element corresponds to an agent observation of $T/S = 0.15$ seconds. The number of bins $K = 64$ and the dimensionality of each feature $D' = 512$. For the loss function, we employ Cross Entropy Loss when the number of sound sources is one, and Binary Cross Entropy Loss when multiple sound sources are assumed. The network is trained using Adam optimizer with a batch size of 16. The learning rate and weight decay are set to 0.0001 and 0.001, respectively.

IV. EXPERIMENTS

In this section, we present the experimental results to evaluate the proposed method for DoA estimation in continuously moving agents.

TABLE 1. Number of parameters for the proposed method and baselines.

	# of parameters	
	audio backbone	DoA estimator
CNN	15.9M	32.8K
CRNN	15.9M	7.94M
Tran	15.9M	25.2M
SLfM	13.3M	32.8K
CNN-Traj	15.9M	65.6K
CRNN-Traj	15.9M	31.6M
SLfM-Traj	13.3M	65.6K
Proposed	15.9M	50.4M

A. DATASET

To obtain observations in which an agent moves continuously through an acoustic environment, we use SoundSpaces [4] and SoundSpaces2.0 [5]. SoundSpaces generates transfer functions from pairs of sound source and agent positions in an indoor scene. We use 3D models of indoor scenes in the Matterport3D dataset [39] and LibriSpeech [40] as a sound source dataset. We note that SoundSpaces can reproduce reberveration based on reverberations based on the wall material and floor plan of an indoor scene from Matterport3D.

1) DATASET SPLITTING

First, we describe the splitting of each dataset. We split the indoor scenes contained in Matterport3D as in the previous work [11], assigning 59/10/8 scenes for training/validation/test subset, respectively. The splitting of the LibriSpeech dataset is also based on the previous work [8]. Although this split allows for overlapping speakers, we remove speaker overlap from this splitting to assess the robustness of our model to unknown sound sources. Specifically, we split the 100 speakers in the split from [8] into 80/10/10 speakers for the train/validation/test subset. Consequently, we obtain 205/24/25 voices for the train/validation/test subset.

2) DETERMINATION OF AGENT'S PATH AND SOUND LOCATION

We then describe how we determine the agent's path and the sound source location. We utilize the graph provided by SoundSpaces [4] for each scene in Matterport3D. This graph contains nodes for every 1.0 m where the agent can move. From these nodes, we first determine the start and end points of the agent's path. Specifically, among all combinations of node pairs, we select those whose shortest paths are less than four in the Manhattan distance and make them candidates for the start and end points. Next, to each start and end point, we assign candidates for the sound source location. In this experiment, all nodes whose shortest paths from the start and end points are within three in the Manhattan distance are considered candidate source locations. Finally, 50 candidate pairs of start and end points are selected from each scene, and three sound source locations are assigned to each start and end point. An example of the start and end points and sound source locations is shown on the left side of Figure 5.

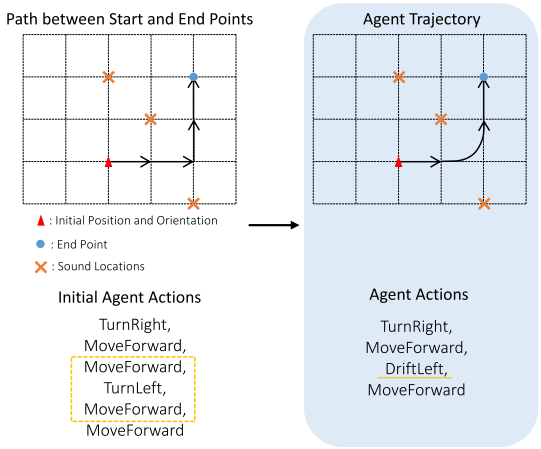

FIGURE 5. An example of an agent's actions. Based on the graph paths provided by SoundSpaces, we determine the sequence of the agent's actions by combining *MoveForward*, *TurnLeft/Right*, and *DriftLeft/Right*.

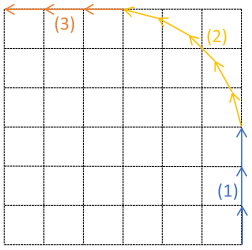

FIGURE 6. Illustration of *DriftLeft*. In this figure, each scale is 1/6 m.

3) DETAILS OF AGENT ACTIONS

We show how we determine agent behavior based on the paths in the graph. We define five types of agent action: *MoveForward*, *TurnLeft/Right*, and *DriftLeft/Right*. *MoveForward* is an action in which the agent moves 1.0 m straightforward, and *TurnLeft/Right* are actions in which the agent rotates 90° to the left or right. First, we transform the graph paths with these three actions. An example of the transformation is shown on the left side of Figure 5. Furthermore, to represent the agent's actions more naturally, we define *DriftLeft/Right* as a new action for diagonal forward movement and replace (*MoveForward*, *TurnLeft/Right*, *MoveForward*) shown in the right side of Figure 5. When an agent follows the path as depicted on the left side of Figure 5, it moves through without hindrance because this path is based on the graph provided by SoundSpaces. However, with the introduction of *DriftLeft/Right*, the agent may deviate from the path and make unexpected movements, such as colliding with obstacles on the new path. In this experiment, such cases are removed and we finally generate 2,479/422/400 agent observations for train/validation/test, respectively.

4) DETAILS IN BINAURAL AUDIO RENDERING

Finally, we describe how we render the binaural audio. In *MoveForward* and *TurnLeft/Right* actions, the transfer functions are calculated by SoundSpaces at each of the

TABLE 2. Results of trajectory-based direction selection. We adopt accuracy and angular error for evaluation metrics. *n* is the number of sound sources in the agent observations.

	accuracy	anguler error
n=1	0.773	11.2
n=2	0.638	19.5
n=3	0.692	17.8

six subdivided points for each action and binaural audio is rendered. Note that the agent moves forward by 1/6 m or rotates by 15° in 0.15 seconds. For interpolation between points, we use linear crossfading as in previous work [5]. However, *DriftLeft/Right* is slightly more complex. Take *DriftLeft* as an example in Figure 6. The agent (1) moves forward by 1/2 m, (2) repeats moving forward with 15° rotation five times, and then (3) moves forward by 1/2 m after 15° rotation. Here, the agent's velocity is consistent with other actions except during moving forward in (2), where the agent progresses by $\frac{1}{2}/\sum_{i=1}^{5}\sin(15\cdot i)$ m in 0.15 seconds.

B. EVALUATION DETAILS

1) BASELINE METHODS AND ABLATIONS

Throughout the experiments, *Proposed* indicates the proposed method and is compared with the following baseline methods and ablations.

a: DoA ESTIMATION USING ONLY AUDIO INPUT

To evaluate the effectiveness of multi-modal input of agent trajectories and binaural audio, we first compare *Proposed* with DoA estimation methods using only audio input. These methods train the DoA estimation model with binaural audio input as a class classification problem in the same way as $L_{original}$. Note that L_{mix} is not used because there is no agent trajectory input. Specifically, we construct the following four methods.

CNN: In *CNN*, the DoA of sound at each time is estimated directly from $\boldsymbol{a}_i$, a binaural audio input. Specifically, this method uses the same idea as previous works that estimate the DoA of fixed sound sources [32], [33].

CRNN: CRNN adopts an architecture that combines CNN and GRUs, similar to the previous sound source tracking method [1]. In this method, the transformer in the proposed network is replaced with the GRU as in the previous work and we only input the binaural audio $\boldsymbol{a}_i$ into the network.

Tran: We further construct *Tran* as a method that employs a transformer to capture time-series information. *Tran* uses the same architecture as *Proposed* except for inputting the agent's trajectory.

SLfM: Chen et al. proposed a method for self-supervised learning of the audio CNN by utilizing the correspondence between DoA changes and agent rotation [37]. To compare our method with this method, we use the pre-trained audio CNN provided by Chen et al. Specifically, we re-train the pre-trained audio CNN with our dataset and denote it as *SLfM*.

TABLE 3. Results of DoA estimation. The evaluation metrics and *n* is as in Table 2. *acc* and *err* are accuracy and angular error, respectively.

	$n=1$		$n=2$		$n=3$	
	acc	*err*	*acc*	*err*	*acc*	*err*
CNN	0.223	20.1	0.209	35.5	0.163	45.6
CRNN	0.738	4.18	0.315	24.0	0.280	29.7
Tran	0.566	8.34	0.285	25.4	0.227	32.7
SLfM	0.262	22.0	0.224	36.9	0.211	41.3
CNN-Traj	0.280	20.9	0.239	35.0	0.191	42.7
CRNN-Traj	0.905	1.54	0.640	11.8	0.676	14.9
Tran-Traj	0.906	1.83	0.684	11.9	0.699	14.6
SLfM-Traj	0.296	20.3	0.251	36.3	0.241	38.8
Mixed	0.862	2.95	0.615	16.7	0.648	20.2
Proposed	**0.936**	**1.23**	**0.722**	**10.2**	**0.735**	**12.9**

b: DoA ESTIMATION USING AUDIO AND TRAJECTORY INPUT

We also construct DoA estimation methods that use both audio and agent trajectory inputs. These methods are ablations of our method that evaluate that the proposed network and loss function effectively utilize the multi-modal input of agent trajectory and binaural audio. Specifically, we modify the input of the networks for *CNN*, *CRNN*, *Tran*, and *SLfM* to incorporate a feature that combines the agent's trajectory and the binaural audio. We call these methods *CNN-Traj*, *CRNN-Traj*, *Tran-Traj*, and *SLfM-Traj*, respectively. We train these methods only with $L_{original}$ as well as methods that input only binaural audio.

Moreover, we add *Mixed*. In *Mixed*, the model is trained with L_{mix} only. *Mixed* is an ablation to evaluate the loss function in *Proposed*, which combines $L_{original}$ and L_{mix}.

Table 1 shows the number of parameters of *Proposed* and baselines. Here, the column audio backbone denotes the number of parematers for encoder of binaural audio input, and the column DoA estimator denotes the number of parameters for the model that estimates the DoA at each time step from the input features. In particular, in the network of *Proposed* in Figure 3, the audio backbone and DoA estimator refer to Audio CNN and Transformer-based Encoder, respectively. Note that *Tran-Traj* and *Mixed* are not included in Table 1 because they use the same architecture as *Proposed*.

2) EVALUATION METRICS

We adopt two evaluation metrics for evaluating DoA estimation.

a: ACCURACY

The first metric is the accuracy of the DoA estimation. In this paper, the DoA estimation is solved as a classification problem, and the network output $o(\boldsymbol{p},\boldsymbol{a}) = \{o^s(\boldsymbol{p},\boldsymbol{a}), 1 \le s \le S\}$ represents the degree of presence of sound sources in each bin, where $o^s(\boldsymbol{p},\boldsymbol{a})$ is the s-th output for $\boldsymbol{p}$ and $\boldsymbol{a}$. When the number of sound sources is only one, the ground truth d_i^s is a single class label, and the accuracy *acc* is calculated as

$$acc = \frac{1}{N\cdot S}\sum_{i=1}^{N}\sum_{s=1}^{S}\delta_{d_i^s,\mathrm{argmax}(o^s(\boldsymbol{p}_i,\boldsymbol{a}_i))}, \tag{4}$$

where $\delta_{i,j}$ is Kronecker delta and $\mathrm{argmax}(\boldsymbol{x})$ is the index of the maximum value of $\boldsymbol{x}$. In summary, *acc* is the average of the exact correspondence between $\boldsymbol{d}_i$ and $o(\boldsymbol{p},\boldsymbol{a})$.

For multiple sources, we extend *acc* to multi-label classification manner. We first extract indices corresponding to the top n_i^s values from $o^s(\boldsymbol{p}_i, \boldsymbol{a}_i)$, where n_i^s is the number of sound sources corresponding to $o^s(\boldsymbol{p}_i, \boldsymbol{a}_i)$. We denote the values by $\boldsymbol{o}_{i,s} = \{o_{i,s}^n, 1 \leq n \leq n_i^s\}$. From d_i^s, a vector of class labels, we also extract the indices whose corresponding value is 1 as $\boldsymbol{d}_{i,s} = \{d_{i,s}^n, 1 \leq n \leq n_i^s\}$. Then, we calculate accuracy *acc* as

$$acc = \frac{1}{N \cdot S} \sum_{i=1}^{N} \sum_{s=1}^{S} \frac{1}{n_i^s} \max_{x,y} \left(\sum_{n=1}^{n_i^s} \delta_{o_{i,s}^{x(n)}, d_{i,s}^{y(n)}} \right), \tag{5}$$

where x and y are arbitrarily ordered sequences of integers with values from 1 to n_i^s.

b: ANGULAR ERROR

The second metric is the estimated angular error. This error indicates the gap between the estimated DoAs and the ground truth. When there is only a single sound source, the angular error *err* is calculated as

$$err = \frac{1}{N \cdot S} \sum_{i=1}^{N} \sum_{s=1}^{S} \min(a_{i,s}, K - a_{i,s}) \cdot \frac{360}{K},$$
$$a_{i,s} = |\mathrm{argmax}(o^s(\boldsymbol{p}_i, \boldsymbol{a}_i)) - d_i^s|, \tag{6}$$

where $a_{i,s}$ is the distance between indices of estimated DoA and ground truth. In the case we have multiple sounds, we calculate the angular error as

$$err = \frac{1}{N \cdot S} \sum_{i=1}^{N} \sum_{s=1}^{S} \frac{1}{n_i^s} \sum_{n=1}^{n_i^s} \min(a_{i,s}, K - a_{i,s}) \cdot \frac{360}{K}, \tag{7}$$

where $a_{i,s} = |o_{i,s}^{x(n)} - d_{i,s}^{y(n)}|$, and x and y are the same as Eqn. 5.

C. PERFORMANCE OF TRAJECTORY-BASED DIRECTION SELECTION

We first evaluate whether TDS is a valid task. Table 2 shows the accuracy and angular error of TDS for each number of sources. Considering that DoA estimation is treated as a 64-class classification problem in this experiment, these results are significantly above the chance rate. Therefore, we see that the proposed network solves the TDS in both single-source and multiple-source cases, i.e., the network can choose only the DoA corresponding to the agent's trajectory from multiple DoAs.

D. PERFORMANCE OF DoA ESTIMATION FOR SINGLE SOUND SOURCE

We perform DoA estimation for single sources, a common problem setting in previous works [2], [32], [33], [36]. The $n = 1$ results in Table 3 correspond to the DoA estimation results for a single source. We first compare *Proposed* with those that use only binaural audio input. *Proposed* outperforms *CNN*, and *CRNN*. Thus, we can see that not only binaural audio features which are essential for DoA estimation but also agent trajectory features are important for the analysis of mobile agents. Moreover, *Proposed* outperforms *Tran*. This result indicates that simply introducing transformer architecture does not improve the DoA estimation performance, and the multi-modal input and the proposed loss function are effective for the analysis of moving agents. We also compare *Proposed* to *SLfM*. *Proposed* shows superior performance to *SLfM*. This indicates that the proposed method also outperforms existing approaches as DoA estimation utilizing mobile agents.

We also compare our method with those that use both binaural audio and trajectory inputs. *Proposed* performs better than *CNN-Traj*. This indicates that *Proposed* can effectively analyze the acoustic environment of a continuously moving agent with the network that captures time-series information. In addition, the estimation performance of *Proposed* is higher than those of *CRNN-Traj* and *SLfM-Traj*. This shows that the proposed model is more suitable for moving agents than simply adding agent trajectory features to the existing sound-tracking method or the existing DoA estimation method utilizing moving agents. Furthermore, *Proposed* is better than *Tran-Traj*, which is the same method except for the introduction of L_{mix}. This result indicates that the proposed loss function, L_{mix}, captures the correspondence between the agent's trajectory and binaural audio, and consequently improves the performance of DoA estimation. Also, *Proposed* performs better than *Mixed*. This shows the importance of $L_{original}$ that trains the network with similar data that is used during inference. Another observation is that for $n = 1$, *CRNN-Traj* and *Tran-Traj* have comparable results. This indicates that in the simple problem setting of estimating the DoA of a single sound source, merely introducing the transformer architecture may not directly lead to improved performance.

We visualize the results of the DoA estimation by *Proposed* in Figure 7. Figure 7 shows the results of DoA estimation at time steps $S = 1$, 12, and 24. As the column of $n = 1$ shows, *Proposed* successfully localizes the sound source at each time step.

E. PERFORMANCE OF DoA ESTIMATION FOR MULTIPLE SOUND SOURCES

Next, we address the DoA estimation for sound sources above the number of microphones in the binaural microphones. This corresponds to the results for $n = 2, 3$ in Table 3. *Proposed* shows higher DoA estimation performance than *CNN*, *CRNN*, and *Tran*. Furthermore, as in the case of a single source, *Proposed* performs better than *SLfM*. These results indicate that the combination of binaural audio features and agent trajectory features is also effective in multiple-source cases.

We also compare *Proposed* with methods that use both the agent's trajectory and binaural audio as input. *Proposed* performs better than *CNN-Traj*, *CRNN-Traj*, *Tran-Traj*, and *SLfM-Traj*. These results show that the proposed architecture and loss function are effective even in difficult problem settings where more sound sources than the number of microphones are simultaneously emitting sound. Also, *Proposed*

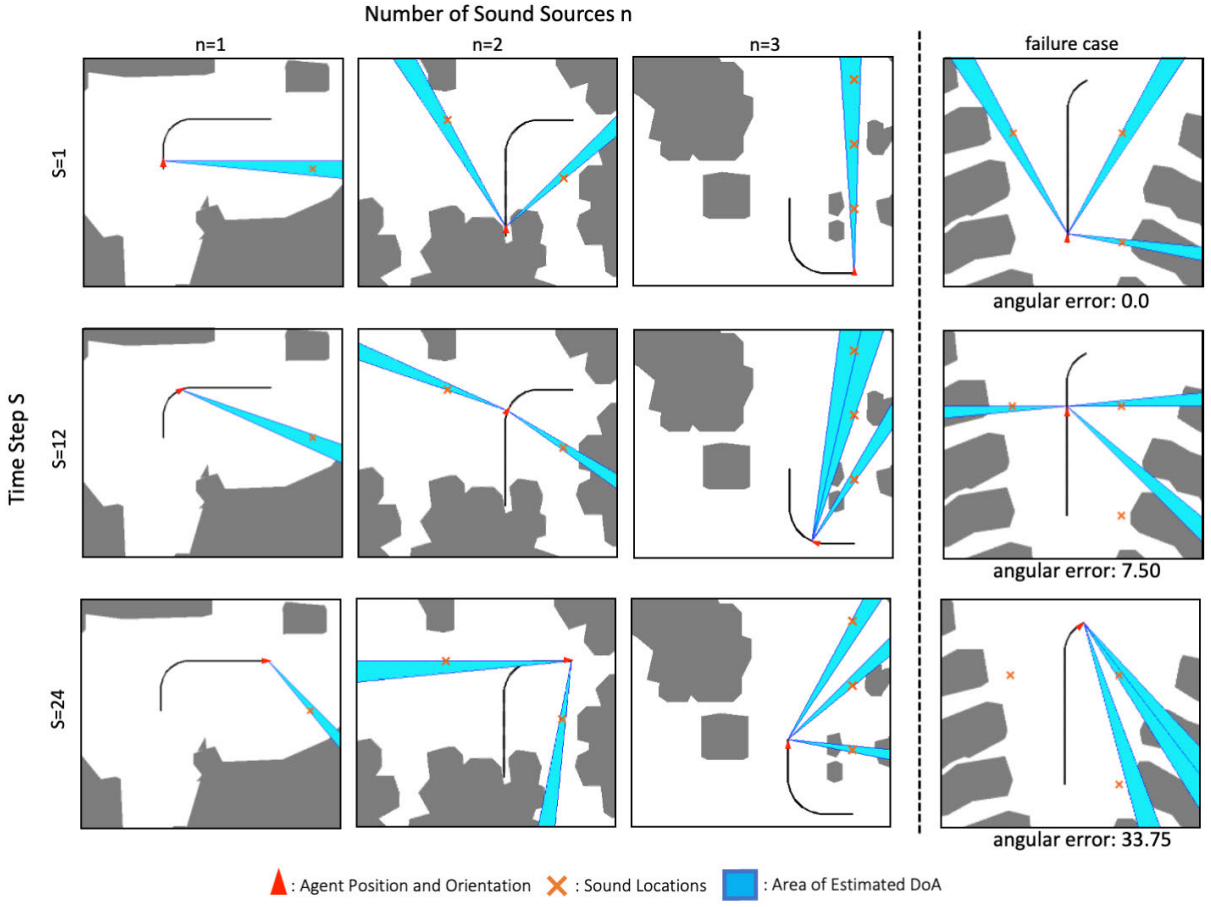

FIGURE 7. Visualization of DoA estimation using the proposed method. The three columns from the left are for 1, 2, and 3 sources, respectively. The rightmost column is the failure case.

shows better performance than *Mixed*. This means that the combination of $L_{origina}$ and L_{mix} is still effective in improving the performance in the DoA estimation for sound sources above the number of microphones. In addition, the accuracy of *Tran-Traj* outperforms the accuracy of *CRNN-Traj* in these settings. Therefore, the introduction of the transformer architecture is effective in difficult problem settings, where estimating the DoA of sound sources with more than the number of microphones is required.

In Figure 7, we also show visualizations of the DoA estimation using *Proposed* for multiple sound sources as well as that for a single sound source. The columns of $n = 2$ and 3 show that *Proposed* can localize each source at each time step. We also show a failure case in the rightmost column of Figure 7. As this example shows, our method may incorrectly estimate some DoAs of multiple sources. However, we note that in many cases, the proposed method improves the DoA estimation performance for multiple sources.

F. PERFORMANCE UNDER NOISY ENVIRONMENT

To demonstrate the robustness of *Proposed*, we further evaluate our method in a noisy environment following previous works [41], [42], [43]. Here, we consider reverberation, directional noise, and spatially white Gaussian noise as noise. In Section IV-D and IV-E, we have already shown that *Proposed* is robust to reverberation and directional noise. First, *Proposed* shows robustness to reverberation reproduced by SoundSpaces in these experiments. Second, *Proposed* achieves DoA estimation of multiple sound sources in Section IV-E. In multiple DoA estimation, *Proposed* also shows robustness to directional noise because when estimating the DoA for each sound source, other sound

TABLE 4. Results of DoA estimation under noisy environment. We set the SNR as 5 dB. We show these results in the same manner as Table 3.

	$n=1$		$n=2$		$n=3$	
	acc	*err*	*acc*	*err*	*acc*	*err*
CNN	0.012	91.1	0.027	81.9	0.047	68.9
CRNN	0.022	88.2	0.029	82.1	0.047	68.5
Tran	0.032	86.1	0.037	73.8	0.056	68.3
SLfM	0.012	93.5	0.032	78.0	0.065	65.6
CNN-Traj	0.016	89.0	0.055	72.8	0.052	67.8
CRNN-Traj	0.087	71.4	0.263	48.8	0.377	46.1
Tran-Traj	0.224	48.7	0.412	**33.9**	0.458	33.5
SLfM-Traj	0.017	91.4	0.046	72.2	0.090	63.6
Mixed	0.348	40.0	0.394	40.1	0.553	28.3
Proposed	**0.418**	**33.6**	**0.437**	34.8	**0.559**	**26.7**

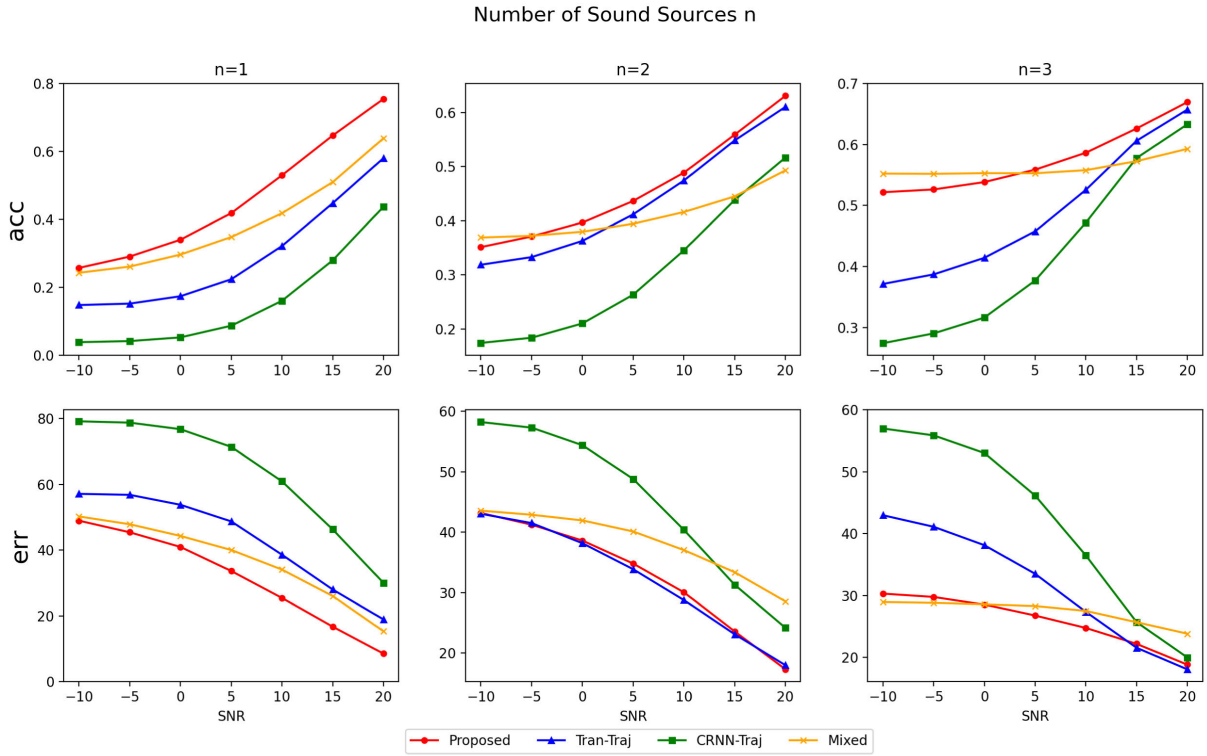

FIGURE 8. Performance of DoA estimation in multiple noise environments. We add spatially white Gaussian noise from -10 dB to 20 dB with a step size of 5 dB. *n* is the number of sound sources.

sources are to be directional noise. In this section, we further show that *Proposed* is robust to spatially white noise. Specifically, we add spatially white Gaussian noise to each binaural audio of the test data and observe the performance of DoA estimation. Note that we have not retrained the model to evaluate the performance in noisy environments.

Table 4 shows the DoA estimation performance of *Proposed* and baseline methods on a noisy dataset with a signal-to-noise ratio (SNR) of 5 dB. *Proposed* shows high DoA estimation performance in noisy environments, although its performance is lower than the original setting. This indicates that the network architecture of *Proposed* and the network training by TDS provide robust DoA estimation. We further compare *Proposed* with baseline methods. First, *Proposed* performs significantly better than *CNN*, *CRNN*, *Tran*, and *SLfM* which are methods with only binaural audio input. Also, *Proposed* shows much better performance than *CNN-Traj* and *SLfM-Traj*, methods with architecture that do not consider time series. These results show that agent trajectories and time series information contribute to the performance of DoA estimation for autonomous agents.

We further compare *Proposed* with *CRNN-Traj*, *Tran-Traj*, and *Mixed*, methods that have multi-modal input of agent trajectories and binaural audio and an architecture considering time series. *Proposed* significantly improves DoA estimation performance for $n = 1$ compared to *CRNN-Traj* and *Tran-Traj*. This is because *Proposed* is trained in an environment with directional noise by TDS, while *CRNN-Traj* and *Tran-Traj* are trained in an environment with no noise other than reverberation in the case of n = 1. However, *Proposed* shows the best performance except for the angular error at $n = 2$, and performs as well as *Tran-Traj* for the angular error at $n = 2$. These results indicate that training the network based on the correspondence between the agent's trajectory and binaural audio by TDS is effective for the difficult task of DoA estimation of multiple sources in noisy environments. *Proposed* also performs better in DoA estimation than *Mixed*. This indicates that the combination of $L_{original}$ and L_{mix}, as in the original environment, contributes to the improved performance.

We also show in Figure 8 the accuracy and angular error for test data with SNR changing from -10 dB to 20 dB with a step size of 5 dB. We choose *Proposed*, *Tran-Traj*, *CRNN-Traj*, and *Mixed* because they are the ones that show good performance in Table 4 because they have both binaural audio and agent trajectory inputs, and that use an architecture considering time series information. As Figure 8 shows, *Proposed* performs well in all SNR. This result shows that *Proposed* is effective for various noise intensities. Interestingly, in noisy settings such as -5 and -10 dB at $n = 2$ and 3, *Mixed* shows comparable performance to *Proposed*. This suggests that L_{mix} contributes significantly to robustness in the difficult situation of DoA estimation for multiple sound sources in a noisy environment.

V. CONCLUSION

We have proposed a novel task, Trajectory-based Direction Selection, to estimate the DoA without disturbing the movement of the embodied agent. Our idea is to effectively acquire cross-modal correspondence between the agent trajectory and the binaural audio by learning to decompose a sound mixture conditioned on the trajectory of each agent. Unlike sound-only methods, the proposed method can localize more sound sources than the number of microphones using low-cost devices, including binaural microphones. Furthermore, unlike existing multi-modal methods for embodied agents, the proposed method can localize sound sources while the agent moves freely. Experimental results show that the performance of our proposed model with agent trajectories and binaural audio as input outperforms existing DoA estimation models.

REFERENCES

[1] S. Adavanne, A. Politis, and T. Virtanen, "Localization, detection and tracking of multiple moving sound sources with a convolutional recurrent neural network," in *Proc. Detection Classification Acoustic Scenes Events Workshop (DCASE)*, 2019, pp. 20–24.

[2] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "Robust sound source tracking using SRP-PHAT and 3D convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Lang., Process.*, vol. 29, pp. 300–311, 2021.

[3] K. Guirguis, C. Schorn, A. Guntoro, S. Abdulatif, and B. Yang, "SELD-TCN: Sound event localization & detection via temporal convolutional networks," in *Proc. 28th Eur. Signal Process. Conf. (EUSIPCO)*, Jan. 2021, pp. 16–20.

[4] C. Chen, U. Jain, C. Schissler, S. V. A. Gari, Z. Al-Halah, V. K. Ithapu, P. Robinson, and K. Grauman, "SoundSpaces: Audio-visual navigation in 3D environments," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 17–36.

[5] C. Chen, C. Schissler, S. Garg, P. Kobernik, A. Clegg, P. Calamia, D. Batra, P. W. Robinson, and K. Grauman, "SoundSpaces 2.0: A simulation platform for visual-acoustic learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 8896–8911.

[6] C. Chen, Z. Al-Halah, and K. Grauman, "Semantic audio-visual navigation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15511–15520.

[7] H. Kondoh and A. Kanezaki, "Multi-goal audio-visual navigation using sound direction map," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2023, pp. 5219–5226.

[8] S. Majumder, Z. Al-Halah, and K. Grauman, "Move2Hear: Active audio-visual source separation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 275–285.

[9] S. Majumder and K. Grauman, "Active audio-visual separation of dynamic sound sources," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 551–569.

[10] R. Gao, C. Chen, Z. Al-Halab, C. Schissler, and K. Grauman, "VisualEchoes: Spatial image representation learning through echolocation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 658–676.

[11] K. K. Parida, S. Srivastava, and G. Sharma, "Beyond image to depth: Improving depth prediction using echoes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8264–8273.

[12] K. Yang, M. Firman, E. Brachmann, and C. Godard, "Camera pose estimation and & localization with & active audio sensing," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 271–291.

[13] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras," *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1255–1262, Oct. 2017.

[14] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap SLAM," *IEEE Trans. Robot.*, vol. 37, no. 6, pp. 1874–1890, Dec. 2021.

[15] Y. Tian, D. Hu, and C. Xu, "Cyclic co-learning of sounding object visual grounding and sound separation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2744–2753.

[16] X. Hu, Z. Chen, and A. Owens, "Mix and localize: Localizing sound sources in mixtures," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10473–10482.

[17] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 570–586.

[18] R. Arandjelović and A. Zisserman, "Objects that sound," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 435–451.

[19] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 435–451.

[20] A. Senocak, T. Oh, J. Kim, M. Yang, and I. S. Kweon, "Learning to localize sound source in visual scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 4358–4366.

[21] H. Chen, W. Xie, T. Afouras, A. Nagrani, A. Vedaldi, and A. Zisserman, "Localizing visual sounds the hard way," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16862–16871.

[22] S. Mo and P. Morgado, "Localizing visual sounds the easy way," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 218–234.

[23] H. Zhao, C. Gan, W.-C. Ma, and A. Torralba, "The sound of motions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1735–1744.

[24] B. Korbar, D. Tran, and L. Torresani, "Cooperative learning of audio and video models from self-supervised synchronization," in *Proc. Adv. Neural Inf. Process. Syst.*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. 2018, pp. 7763–7774.

[25] C. Gan, H. Zhao, P. Chen, D. Cox, and A. Torralba, "Self-supervised moving vehicle tracking with stereo sound," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7052–7061.

[26] K. Yang, B. Russell, and J. Salamon, "Telling left from right: Learning spatial correspondence of sight and sound," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 9932–9941.

[27] R. Gao and K. Grauman, "2.5D visual sound," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 324–333.

[28] H. Zhou, X. Xu, D. Lin, X. Wang, and Z. Liu, "Sep-stereo: Visually guided stereophonic audio generation by associating source separation," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 52–69.

[29] X. Xu, H. Zhou, Z. Liu, B. Dai, X. Wang, and D. Lin, "Visually informed binaural audio generation without binaural audios," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15480–15489.

[30] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*. New York, NY, USA: Simon & Schuster, 1992.

[31] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. AP-34, no. 3, pp. 276–280, Mar. 1986.

[32] D. Suvorov, G. Dong, and R. Zhukov, "Deep residual network for sound source localization in the time domain," 2018, *arXiv:1808.06429*.

[33] P. Vecchiotti, N. Ma, S. Squartini, and G. J. Brown, "End-to-end binaural sound localisation from the raw waveform," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2019, pp. 451–455.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. U. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2017, pp. 1–11.

[35] 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," in *Proc. Interspeech*, 2021, pp. 2117–2121.

[36] S. Kuang, K. van der Heijden, and S. Mehrkanoon, "BAST: Binaural audio spectrogram transformer for binaural sound localization," 2022, *arXiv:2207.03927*.

[37] Z. Chen, S. Qian, and A. Owens, "Sound localization from motion: Jointly learning sound direction and camera rotation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 7897–7908.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[39] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niebner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3D: Learning from RGB-D data in indoor environments," in *Proc. Int. Conf. 3D Vis. (3DV)*, Oct. 2017, pp. 667–676.

[40] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.
[41] W. Xue, S. Liang, and W. Liu, "DOA estimation of speech source in noisy environments with weighted spatial bispectrum correlation matrix," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 2282–2286.
[42] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 2814–2818.
[43] O. Schwartz, Y. Dorfan, M. Taseska, E. A. P. Habets, and S. Gannot, "DOA estimation in noisy environment with unknown noise power using the EM algorithm," in *Proc. Hands-Free Speech Commun. Microphone Arrays (HSCMA)*, Mar. 2017, pp. 86–90.

TOMOYA SATO received the B.S. degree from the Department of Information and Communication Engineering, The University of Tokyo, in 2019, and the M.S. and Ph.D. degrees from the Graduate School of Information Science and Technology, The University of Tokyo, in 2021 and 2024, respectively.

YUSUKE SUGANO (Member, IEEE) received the Ph.D. degree in information science and technology from The University of Tokyo, in 2010. He was an Associate Professor with the Graduate School of Information Science and Technology, Osaka University, a Postdoctoral Researcher with the Max Planck Institute for Informatics, and a Project Research Associate with the Institute of Industrial Science, The University of Tokyo. He is currently an Associate Professor with the Institute of Industrial Science, The University of Tokyo. His research interests include computer vision and human–computer interaction.

YOICHI SATO (Senior Member, IEEE) received the B.S. degree from The University of Tokyo, in 1990, and the M.S. and Ph.D. degrees in robotics from the School of Computer Science, Carnegie Mellon University, in 1993 and 1997, respectively. He is currently a Professor with the Institute of Industrial Science, The University of Tokyo. His research interests include first-person vision, gaze sensing and analysis, and illumination and reflectance analysis. He served/is serving in several conference organization and journal editorial roles, including IEEE TRANSACTIONS ON PATTERN ANALYSIS and MACHINE INTELLIGENCE, *International Journal of Computer Vision*, *Computer Vision and Image Understanding*, the CVPR 2023 General Co-Chair, the ICCV 2021 Program Co-Chair, the ECCV 2012 Program Co-Chair, the ACCV 2018 General Co-Chair, and the ACCV 2016 Program Co-Chair.