applied sciences

MDPI

Article

Audio–Visual Sound Source Localization and Tracking Based on Mobile Robot for The Cocktail Party Problem

Zhanbo Shi, Lin Zhang * and Dongqing Wang

School of Software Engineering, Tongji University, Shanghai 201804, China; 2111291@tongji.edu.cn (Z.S.); wangdongqing@tongji.edu.cn (D.W.)
* Correspondence: cslinzhang@tongji.edu.cn

Abstract: Locating the sound source is one of the most important capabilities of robot audition. In recent years, single-source localization techniques have increasingly matured. However, localizing and tracking specific sound sources in multi-source scenarios, which is known as the cocktail party problem, is still unresolved. In order to address this challenge, in this paper, we propose a system for dynamically localizing and tracking sound sources based on audio–visual information that can be deployed on a mobile robot. Our system first locates specific targets using pre-registered voiceprint and face features. Subsequently, the robot moves to track the target while keeping away from other sound sources in the surroundings instructed by the motion module, which helps the robot gather clearer audio data of the target to perform downstream tasks better. Its effectiveness has been verified via extensive real-world experiments with a 20% improvement in the success rate of specific speaker localization and a 14% reduction in word error rate in speech recognition compared to its counterparts.

Keywords: robot audition; sound source localization; cocktail party problem; audio–visual; motion planning

check for updates

Citation: Shi, Z.; Zhang, L.; Wang, D. Audio–Visual Sound Source Localization and Tracking Based on Mobile Robot for The Cocktail Party Problem. *Appl. Sci.* **2023**, *13*, 6056. https://doi.org/10.3390/app13106056

Academic Editor: Marco Troncossi

Received: 21 April 2023
Revised: 11 May 2023
Accepted: 13 May 2023
Published: 15 May 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

To achieve natural and effective human–robot Interaction (HRI), robots need to utilize multiple senses to communicate with humans. Among these senses, robot audition has become increasingly important in recent years. The concept of robot audition was first proposed by Nakadai et al. [1] and is defined by three key abilities, (1) the ability to locate sound sources in unknown acoustic environments, (2) the ability to move actively to obtain more information about the sound and (3) the ability to continuously perform acoustic scene analysis in noisy environments. Current research on robot audition focuses on Sound Source Localization (SSL) [2], Sound Source Separation (SSS) [3] and Automatic Speech Recognition (ASR) [4]. While these techniques have become increasingly mature in single-speaker scenarios after more than two decades of development, locating and tracking a particular speaker in a noisy environment where multiple speakers are communicating simultaneously as illustrated in Figure 1 remains a challenge, and is known as the cocktail party problem [5]. This problem involves two challenges [6], how to locate the sound of interest and how to collect a clear audio signal of the target from a variety of background noises.

The first challenge can be solved via SSL. At present, traditional SSL methods have the ability to locate the sound sources relying only on microphone arrays [7,8]. However, when the sound sources are far from the robots compared to the aperture of the microphone array (called far-field effects), these SSL methods cannot estimate the distance between the sources and the robots [9]. To deal with this problem, more and more studies attempt to combine visual Simultaneous Localization and Mapping (SLAM) with SSL [9–12], which can provide the relative positions of the sources and the robots. Although these SLAM-based SSL approaches can accurately estimate the locations of sound sources, they cannot

determine which source a particular sound is emitted from. Specifically, suppose there are two people in a room, one is singing and the other is calling out. The above-mentioned SLAM-based SSL methods can estimate the relative position between these two active sound sources and the robot, but they do not know who is singing and who is making a phone call. Therefore, these techniques cannot be applied directly to resolve the cocktail party problem where the robot should locate the source of interest such as the one that is singing rather than all sound sources in the environment.

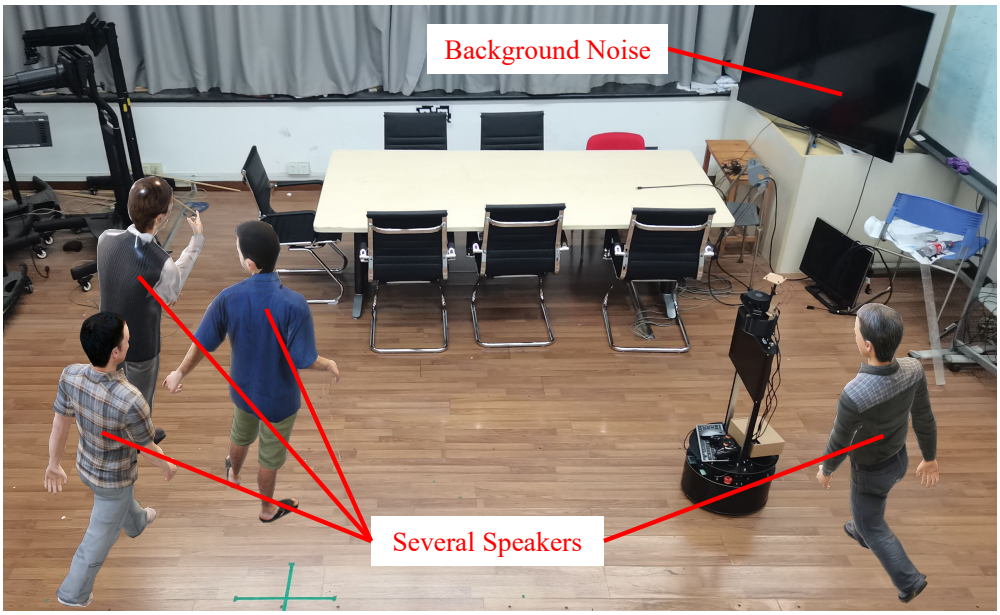


Figure 1. A common environment of the cocktail party problem. Multiple sound sources are active at the same time, e.g., several speakers are speaking simultaneously and the television is making background noise.

To meet the second challenge, there has been an amount of work investigating speech separation in cocktail party environments [3,13–15] and the most advanced techniques have been able to separate clear speech from mixed signals [3]. However, because of the front–back ambiguity of binaural arrays [16,17], these methods trained with binaural videos cannot be directly applied in a 3D cocktail party environment. Furthermore, most of them ignore the ability of robots to move actively to acquire more information about the target.

In this paper, we attempt to fill the aforementioned research gaps to some extent. We propose AV-SSLT (short for Audio–visual Sound Source Localization and Tracking), a framework that uses voiceprint and face features to identify the target speaker and locate her/his position and collects clear audio data of the target by moving towards it while staying away from other sound sources in the surroundings controlled via the motion module. The characteristics of AV-SSLT and our contributions can be summarized as follows:

- To identify and localize a specific sound source in a cocktail party environment, we propose an audio–visual approach since the vision can help the audition to distinguish the target speaker. Considering that voiceprint recognition and face recognition technologies are relatively mature, AV-SSLT combines these two types of techniques to find out the specific sound source and its position using microphones and RGB-D cameras. Compared with the existing work [18–21], the proposed approach is able to autonomously select the target of interest among several speakers. In addition, the method is well suited for situations where the users' identities are relatively fixed and the audio–visual features can be simply recorded, such as routine company meetings.
- Nowadays, most studies on the cocktail party problem ignore the ability of the robot to move actively to obtain more information about sound sources. To make full use of

this ability to acquire clearer audio of the target, a motion module is designed based on the Direction of Arrival (DoA) of the sounds. After locking on the target, the robot will move towards this speaker while keeping away from other interfering sound sources. To the best of our knowledge, this is the first work to improve the quality of the captured audio through robot motion.

- To verify the effectiveness of AV-SSLT, it was deployed on a mobile robot and real-world experiments were conducted with multi-speaker scenarios. Compared with its counterparts, AV-SSLT has a much higher success rate in localizing the particular target and significantly better accuracy concerning speech recognition at the end of tracking. To make our results reproducible, the source code is available at https://github.com/ZhanboShiSE/AV-SSLT (accessed on 10 May 2023).

2. Related Work

2.1. Audio–Visual Sound Source Localization (AV-SSL)

Audio–visual sound source localization techniques can be categorized into traditional methods and learning-based methods. The traditional ones mainly combine SSL with SLAM [9–12]. This type of method typically begins with estimating DoA using multi-channel audio data from microphone arrays and then calculates the relative distances between the sound sources and the robot referenced by SLAM. For instance, Emery et al. [10] used SEVD-MUSIC to estimate the DoA of the sound and RGB-D SLAM for mapping and localization. Subsequently, the extended Kalman filter was utilized to track the movement of the source. This method successfully located one static or moving speaker in the real world. Similarly, Chau et al. [11] made use of GSVD-MUSIC and EKF-SLAM to localize and track the sound sources. In addition, combined with OpenPose, a monocular camera-based human pose estimator, their system focused more on human–robot interaction. In contrast with the above two studies using just one robot, Michaud et al. [9] proposed a method using two robots referenced via RTAB-Map, a graph-based SLAM, to localize the sound source. In [9], each robot was equipped with a tower-shaped 16-mic 3D array. To locate and track the sound source, SRP-PHAT-HSDA and triangulation were used in their method. It is obvious that the aforementioned studies improve the accuracy of sound source localization by means of SLAM. In order to improve the robustness of SLAM to dynamic environments, an approach based on direct-path relative transfer functions was presented by Zhang et al. [12] to eliminate the moving objects.

With the development of machine learning, some researchers have been seeking learning-based solutions to locate sound sources. For instance, Masuyama et al. [22] introduced a self-supervised learning-based probabilistic spatial model for AV-SSL. To train this model, 360° images with multi-channel audio signals were fed into visual and auditory DNNs. Chen et al. [23] replaced the DNN with SVM and trained the models using 1-channel audio from a single microphone to simultaneously estimate the DoA and the distance between the sound source and the microphone. Regrettably, although the above systems can accurately locate multiple sound sources at the same time, they are unable to identify the sources and focus on a specific target, which is actually a highly desired capability of robot audition in cocktail party environments. To address this issue, SELDnet was proposed in [21] for Sound Event Localization and Detection (SELD) of overlapping sources to provide the robot with the ability to identify the temporal activities of each sound event and estimate their respective spatial location trajectories when active. In recent years, SELD has received increasing attention from researchers [24–26]. However, even the most advanced techniques of SELD only have a success rate of around 41.6% in sound event detection and an angular error of approximately 18.5° in sound event localization [20]. Due to the low performance, these techniques cannot be applied directly to resolve the cocktail party problem.

2.2. Motion Planning for Robot Audition

With the development of robot path planning techniques [27,28], the majority of current research attempts to solve the front–back ambiguity of linear arrays by designing the path of the robot based on the information theory criteria. For example, Vinecent et al. [16] proposed a dynamic programming algorithm that minimized the entropy of a discrete occupancy grid to guide the movement of the robot to resolve ambiguities. Similarly, the belief entropy at each future time step was used to quantify the uncertainty in estimating the source location by Nguyen et al. [17]. After that, the motion of the robot was controlled via a Monte Carlo tree search algorithm. Later, in [29], Nguyen et al. extended their former work [17] to locate moving, intermittent sounds using the extended mixture Kalman filter. Unlike the previous methods that capture the sound signals with the Kinect 4-microphone array, a Monte Carlo exploration approach for active binaural localization was presented by Schymura et al. [30]. Bustamante and Danes [31] also proposed a robot motion control method based on information entropy using binaural microphones. Furthermore, in [32], Bustamante et al. took the rotation of the robot head into account to reduce the number of steps to find the source. For situations with strong reverberation and echo such as lecture hall, Sewtz et al. [33] introduced the motion model enhanced MUSIC to improve the accuracy of SSL. Unfortunately, the studies mentioned are mainly designed for a single sound source. To the best of our knowledge, no work has attempted to design a motion planning algorithm for robots to perceive sound sources in a cocktail party environment.

3. Methodology

3.1. Overview

In order to enable the mobile robot to selectively locate and track the sound source of interest in the cocktail party environments, we propose an audio–visual sound source localization and tracking system called AV-SSLT. The system first collects the audio data from the environment with an omnidirectional microphone. Acoustic features such as Mel-scale Frequency Cepstral Coefficients (MFCC) are subsequently extracted from the 1-channel audio data for detecting keywords of interest. When awakened by a keyword, the system extracts the voiceprint features from the audio. By comparing the voiceprint features with the voiceprint–face database, the system obtains the face features of the target. By doing so, the robot can identify the target speaker among multiple speakers in RGB-D images. As for the control of the robot's movement, the system estimates the DoA using 4-channel audio data at first. After obtaining the DoA information, the robot rotates in these directions to capture the target speaker with the camera. Once the target is detected, the robot moves towards it while staying away from other interfering sound sources in the surroundings as much as possible under the control of the motion module. The overall framework of AV-SSLT is illustrated in Figure 2.

3.2. Specific Target Identification

To locate and track the sound source of interest in a cocktail party environment, the system first needs to identify the specific speaker uttering the keywords. To achieve this, we equip the robot with an omnidirectional microphone to constantly pick up sound signals in the environment. Meanwhile, the input signals are transformed into the complex domain using Short-Time Fourier transform (STFT) to extract the MFCC features. These MFCC features are then fed into a recurrent neural network transducer (RNN-T) [34–36] to detect the keywords. Once the system is awakened by a keyword, the part of the audio containing the keyword is fed into Tse-FV [37] for voiceprint feature extraction. The resulting features from Tse-FV consist of two parts, MFCC features and d-vector features. By comparing these features with the voiceprint–face matching database, the identity of the target speaker can be determined.

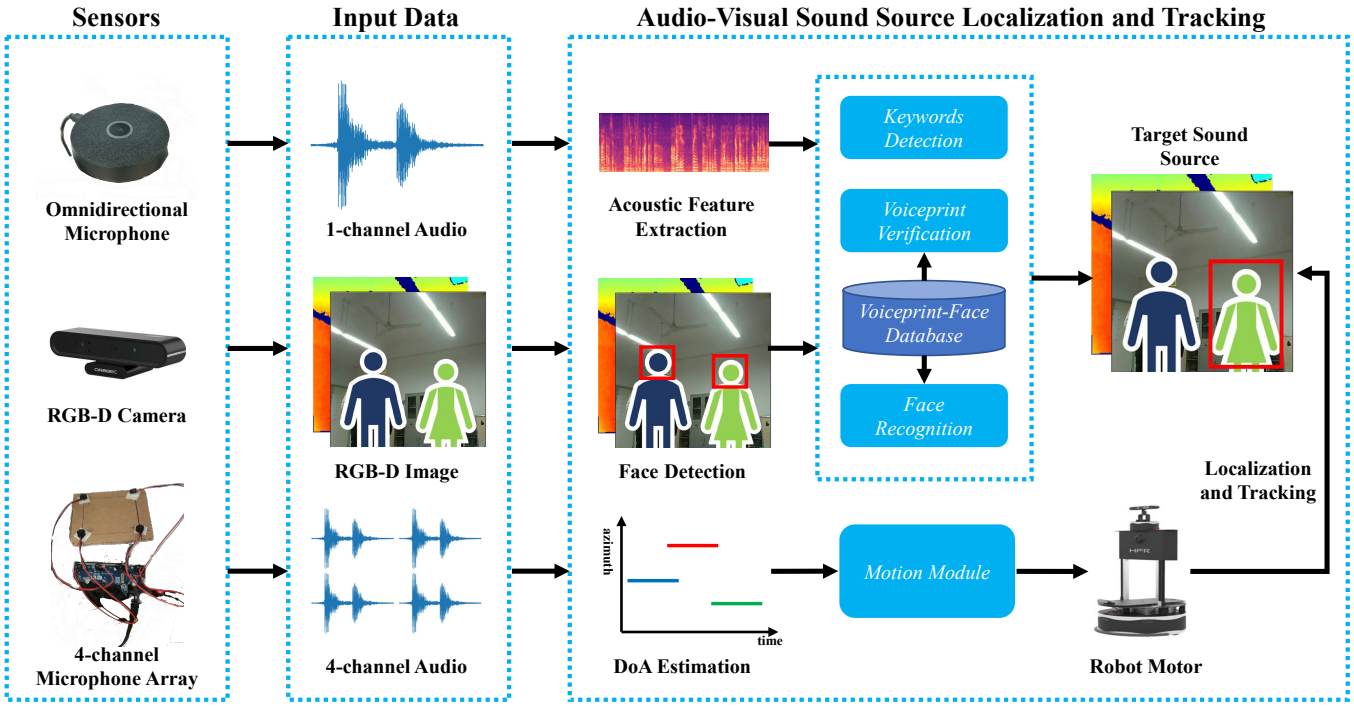


Figure 2. The overall framework of AV-SSLT. To identify a target sound source of interest, the robot continuously collects sound signals and waits for keyword detection. Once a keyword is detected, acoustic features extracted from the sound signals and face features obtained from RGB-D images are compared with the database to identify the target. To localize and track the target speaker, the motion module controls the robot's movement using the directions of arrival (DoA) estimated from the audio data.

Note that when the target's identity is retrieved from the database, her or his facial features are also retrieved. With such features, we further resort to ArcFace [38] to find the target face for the robot equipped with the RGB-D camera to locate and track the speaker of interest. It is easy for the above process if the target is within the camera's field of view (FOV). However, if the target is out of view, the robot must first rotate to face the target, which is guided by the DoA estimation discussed in the next subsection.

3.3. DoA Estimation

In a real-world environment, the sound source of interest is often out of the FOV of the camera. To locate and track such a target, the robot should first estimate its DoA and then rotate to find it. In our AV-SSLT, the DoA estimation is based on the SEVD-MUSIC algorithm [18] using another microphone array. MUSIC takes the transfer function from sound to each microphone as a prior. Consider a microphone array with N_m microphones (which is 4 in our system), the multichannel transfer function can be expressed as,

$$\boldsymbol{H}(\theta,\omega) = [h_1(\theta,\omega),\ldots,h_{N_m}(\theta,\omega)]^T, \tag{1}$$

where θ is the azimuth of the sound source relative to the microphone array in the 2D case, ω represents the frequency and $h_i, i \in (1, N_m)$ is the transfer function from the sound to the i-th microphone.

To perform sound source localization, we have to calculate the correlation matrix among the input signal channels. To achieve this, the signal vector in the frequency domain is firstly obtained via STFT of the input acoustic signal in N_m channels as,

$$\boldsymbol{X}(\omega,f) = [X_1(\omega,f),\ldots,X_{N_m}(\omega,f)]^T, \tag{2}$$

where f expresses the frame index. With $\boldsymbol{X}(\omega, f)$, the correlation matrix among its channels can be further defined as follows for every frame and every frequency,

$$\boldsymbol{R}(\omega, f) = \boldsymbol{X}(\omega, f)\boldsymbol{X}(\omega, f)^*, \tag{3}$$

where $(\cdot)^*$ denotes the conjugate transpose operator. In order to be robust against noise, in AV-SSLT, $\boldsymbol{R}(\omega, f)$ is time-averaged over T_r frames via,

$$\boldsymbol{R}'(\omega, f) = \frac{1}{T_r} \sum_{i=0}^{T_r-1} \boldsymbol{R}(\omega, f+i). \tag{4}$$

To separate the speech signal and noise signal from the mixed signal, standard eigenvalue decomposition of $\boldsymbol{R}'(\omega, f)$ is conducted to decompose the signal space into the speech and noise subspaces,

$$\boldsymbol{R}'(\omega, f) = \boldsymbol{E}(\omega, f)\boldsymbol{\Lambda}(\omega, f)\boldsymbol{E}^{-1}(\omega, f), \tag{5}$$

in which $\boldsymbol{\Lambda}(\omega, f) = \mathrm{diag}(\lambda_1(\omega, f), \ldots, \lambda_{N_m}(\omega, f))$ is a diagonal matrix composed of the eigenvalues of $\boldsymbol{R}'(\omega, f)$ in descending order and the i-th column of $\boldsymbol{E}(\omega, f) = [e_1(\omega, f), \ldots, e_{N_m}(\omega, f)]$ is the eigenvector of $\boldsymbol{R}'(\omega, f)$ associated with $\boldsymbol{R}'(\omega, f)$'s eigenvalue $\lambda_i(\omega, f)$.

Subsequently, the MUSIC spectrum ζ for SSL can be written as follows,

$$\zeta(\theta, \omega, f) = \frac{|\boldsymbol{H}^*(\theta, \omega)\boldsymbol{H}(\theta, \omega)|}{\sum_{i=N_s+1}^{N_m} |\boldsymbol{H}^*(\theta, \omega)e_i(\omega, f)|}, \tag{6}$$

where $\boldsymbol{H}(\theta, \omega)$ is the multichannel transfer function defined in Equation (1) and N_s is an empirical parameter considered as the number of sound sources in the SSL process to remove the noise from the time-averaged correlation matrix $\boldsymbol{R}'(\omega, f)$.

Note that the MUSIC spectrum defined by Equation (6) is for every frequency. To make it applicable to human speech within a certain frequency range between i_{low} and i_{high}, the response of the MUSIC spectrum can be calculated by summing out ω in Equation (6) as,

$$\overline{\zeta}(\theta, f) = \frac{1}{i_{high} - i_{low} + 1} \sum_{i=i_{low}}^{i_{high}} \zeta(\theta, \omega_i, f). \tag{7}$$

At every frame f, the peaks of $\overline{\zeta}(\theta, f)$ which have larger values than a threshold are detected as the direction of the active sound source.

In a cocktail party environment, there are typically several active sound sources present at the same time. To capture clearer speech of the speaker of interest in such an environment, it is natural to get close to that speaker while keeping away from other distracting sound sources, which can be achieved with the motion module in our AV-SSLT.

3.4. Motion Module

In order to be more robust against interfering sound sources in the cocktail party environment where multiple sound sources are active simultaneously, the motion module is designed to control the robot to move towards the target while maintaining a distance from other sound sources based on the estimated DoA. More specifically, the purpose of the motion module is to find an optimal direction, which is a weighted sum of the direction of the target sound source and the opposite directions of the other interfering sound sources. To calculate this direction, the directions of sound sources are firstly converted into unit vectors in the robot coordinate system by means of trigonometric functions as follows,

$$\begin{aligned} \boldsymbol{V} = [&(\cos\theta_t, \sin\theta_t)^T, \\ &(\cos\theta_0, \sin\theta_0)^T, \ldots, (\cos\theta_{N_o}, \sin\theta_{N_o})^T], \end{aligned} \tag{8}$$

where θ_t denotes the direction of the target sound source, N_o is the number of other sound sources in the environment and $\theta_i, i \in (1, N_o)$ represents other sound source directions.

After that, the response of the MUSIC spectrum of the target sound source and the other sound sources for the current time frame f can be written as,

$$\overline{\boldsymbol{\zeta}}' = [\overline{\zeta}(\theta_t, f), \overline{\zeta}(\theta_1, f), \ldots, \overline{\zeta}(\theta_{N_o}, f)], \tag{9}$$

which indicates the intensity of the effect of each sound source on the robot.

Thus, the direction of the robot's next movement can be straightforwardly expressed as,

$$(x, y)^T = \frac{\boldsymbol{V}(\boldsymbol{W}^T \odot \overline{\boldsymbol{\zeta}}'^T)}{||\boldsymbol{V}(\boldsymbol{W}^T \odot \overline{\boldsymbol{\zeta}}'^T)||_2}, \tag{10}$$

where $\boldsymbol{W} = [w_t, w_1, \ldots, w_{N_o}]$ indicates the weights of the directions of these sound sources, which affect the length of the path to the target, and $\odot$ denotes the element-wise product.

In order to instruct the robot to rotate, the direction obtained in Equation (10) is further converted into radians by the inverse trigonometric functions as,

$$\theta' = \arccos \frac{x}{\sqrt{x^2 + y^2}}. \tag{11}$$

4. Experiments and Results

To validate the performance of our proposed AV-SSLT in cocktail party environments, we performed experiments to test its ability to accurately identify the target speaker in a noisy environment with several speakers and its accuracy of speech recognition when the robot is at the end of tracking. The experimental setup and the experimental results are as follows.

4.1. Experimental Setup

4.1.1. Experimental Site

The room where the experiments were conducted was an office of size 10.5 × 10.5 m^2. There was an open space of the size about 3 × 5 m^2 in the middle surrounded by tables and chairs. The Room Impulse Response (RIR) was measured according to [39] as shown in Figure 3. During the measurements, a four-channel microphone array was positioned in the center of the room 1 m above the ground. At the same height, a loudspeaker was installed. The loudspeakers were placed at positions on a circle with a 1 m radius around the microphone array. An initial position of the loudspeaker was marked as 0°, from which the measurements were conducted in a counter-clockwise direction at 10° intervals. The Reverberation Time (RT60) of this room was about 1.52 s.

4.1.2. Hardware

The platform for experiments was a Handsfree stone robot equipped with an Orbbec Astro Pro RGB-D camera, a PlayStation Eye 4-channel microphone array and a conference-grade omnidirectional microphone. The configuration of the sensors is illustrated in Figure 4. Another high-resolution 32-channel spherical microphone array was used to generate the FOA format recordings. The processor hardware of the mobile robot is listed in Table 1.

Table 1. The processor hardware of the mobile robot.

Hardware	Model
CPU	Intel Core i7-8700 @ 3.20 GHz
GPU	NVIDIA GeForce GTX 1650 4 G
Memory	Micron 8G DDR4 2400 MHz ×2

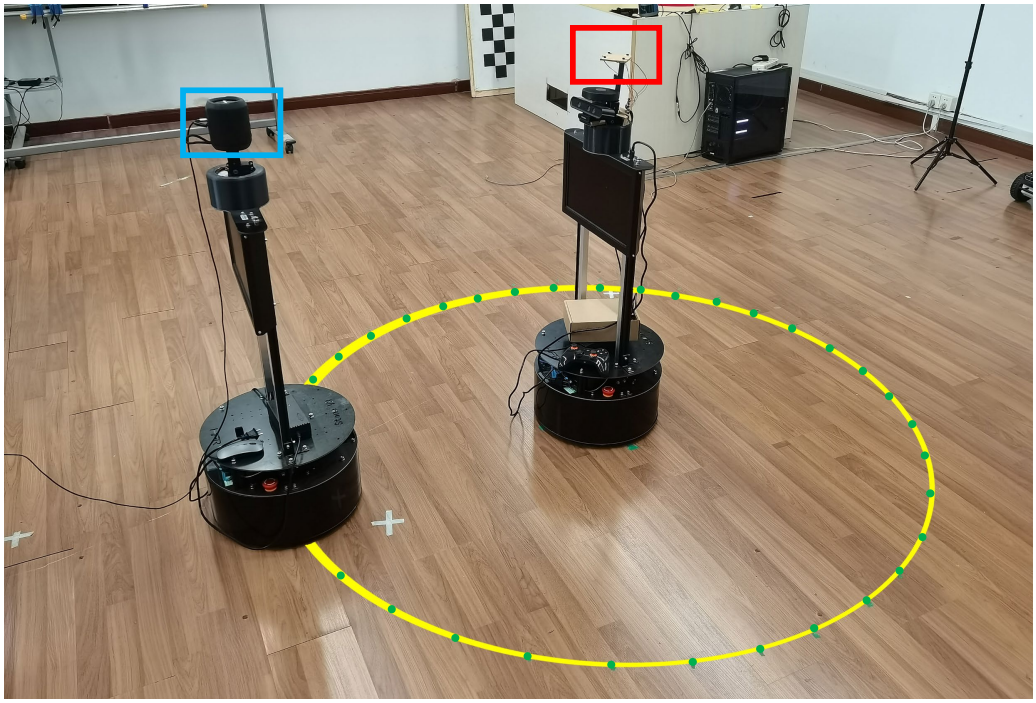


Figure 3. Room impulse response measurement. A 4-channel microphone array, indicated by the red rectangle, was positioned in the center of the room. A loudspeaker marked by the blue rectangle was placed at positions indicated by the yellow circle with a 1 m radius around the microphone array. The initial position of the loudspeaker was marked as 0° and measurements were taken in a counter-clockwise direction, at 10° intervals, from the positions marked by the green points.

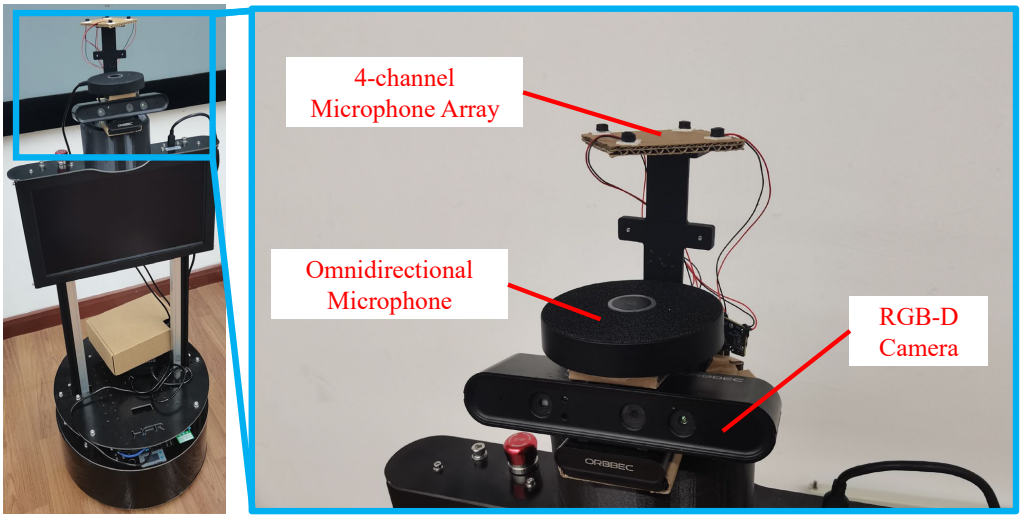


Figure 4. (**a**) The experimental platform. (**b**) The sensor configuration. The RGB-D camera is mounted on the head of the robot which is 1 m above the ground. The omnidirectional microphone is placed on top of the camera and the 4-channel microphone array is installed 0.1 m above the camera.

4.1.3. Episode Definition

In the experiments, three speakers stood casually in the room and had a conversation. For each episode, one of the speakers uttering the keywords was treated as the target speaker. The other two speakers were presented as interfering sound sources. The sound of a television program was considered as the background noise and its level was about 55 dB measured 1.5 m away from the television. All the sound sources mentioned above can be considered as far-field signals introduced in [40]. The starting position of the robot

was a random position in the room and the starting orientation of the robot was a random direction with the target speaker out of the FOV of the camera. An episode was considered successful if the robot was able to stop within 1.5 m of the speaker who spoke the keywords and keep the camera facing this speaker. As an example, a specific configuration of an episode in the experiments is presented in Figure 5.

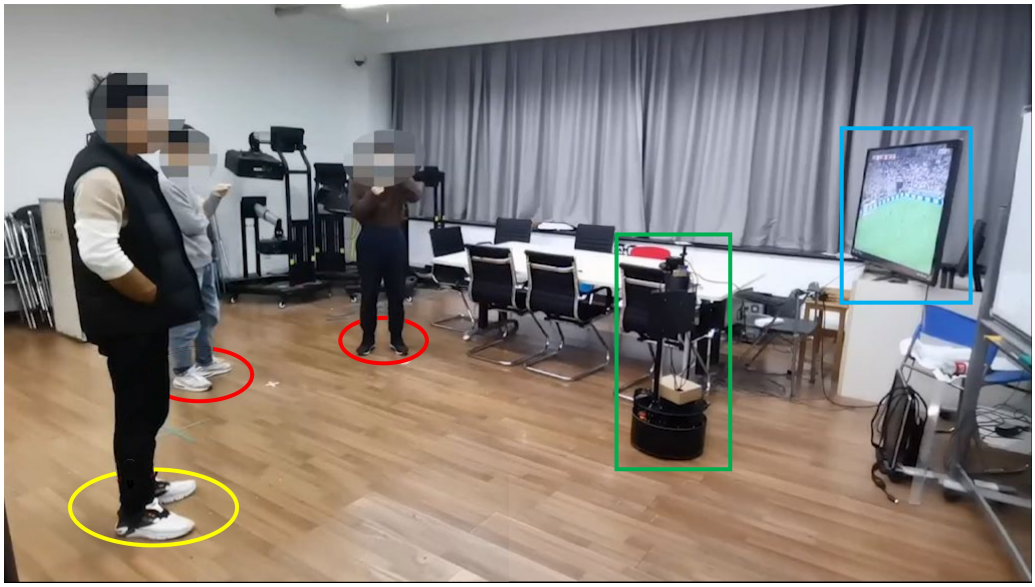


Figure 5. The configuration of an episode. The person marked by the yellow circle is the target speaker. The other people, identified by the red circles, are presented as the interfering sound sources. The robot highlighted by the green rectangle starts at a random position in the room. The sound from the television indicated by the blue rectangle served as the background noise.

4.1.4. Metrics

The average success rate of the robot in locating a speaker uttering the keywords, the average distance to the target at the end of tracking and the word error rate (WER) of speech recognition at the end of tracking serve as the evaluation metrics for the performance of the system. The first two correspond to the first challenge in the cocktail party problem, locating the sound of interest, while the latter corresponds to the second challenge, collecting a clear audio signal of the target from a multiple sound sources signal.

4.2. Results

4.2.1. The Performance of Locating the Sound of Interest

To verify the ability of our system in solving the first challenge in the cocktail party problem, we conducted experiments in the real world. We compared our AV-SSLT with three representative solutions in this field, the SEVD-MUSIC [18], the noise-robust GSVD-MUSIC [19] and the Multi-ACCDOA-based SELDnet [20,21]. The two MUSIC-based algorithms took the audio data recorded via the four-channel microphone array as input. Two strategies were applied to select the direction of the target sound source from all DoA estimations, a random direction or the direction with the highest MUSIC spectrum response. The SELDnet was trained on the STARSS22 dataset [41] and a dataset synthesized using the method proposed in [42] with a batch size of 128 and a dropout rate of 0.05. The model was then fine-tuned on the real-world FOA format recordings from the experimental room generated via the 32-channel spherical microphone array. The target sound source of SELDnet was selected as the sound source whose sound event category was human speech and which was active at the same time as the keywords were detected. More detailed parameters of the MUSIC-based methods and SELDnet are presented in Tables 2 and 3, respectively.

Table 2. Parameter list of the MUSIC-based algorithm.

Parameter	Value
Sampling Rate	16,000 Hz
Frame Length	512
Window Size	64
Dimension of Mel-spectrograms	13
Lower Bound Frequency	300 Hz
Higher Bound Frequency	2800 Hz
Max Number of Sounds	3
Interval of Azimuth	10°

Table 3. Parameter list of the SELDnet algorithm.

Parameter	Value
Sampling Rate	24,000 Hz
Frame Length	480
Label Frame Length	2400
Dimension of Mel-spectrograms	64
Unify Threshold of Multi-ACCDOA	15
Feature Sequence Length	250
Label Sequence Length	50

The results are summarized in Table 4. It can be seen that our AV-SSLT has a much higher success rate in locating the specific target and has the ability to get closer to the target than its counterparts relying solely on the microphone array. That is to say, by combining voiceprint and face features, our approach performs well in locating a specific speaker among multiple speakers, which is the first challenge in the cocktail party problem. Furthermore, in order to demonstrate the capability of our AV-SSLT system to work in real time, the average response time of each module was measured during the experiments and is presented in Table 5. It can be seen that the total response time of the entire system is about 0.5 s, which is sufficient for real-time interaction with humans.

Table 4. Performance on locating the sound of interest.

Method	Success Rate ↑	Distance to Target ↓
SEVD-MUSIC [18] (random)	24%	2.34 m
SEVD-MUSIC [18] (high response)	40%	2.02 m
GSVD-MUSIC [19] (random)	28%	2.28 m
GSVD-MUSIC [19] (high response)	42%	1.96 m
SELDnet [20,21]	62%	1.68 m
AV-SSLT (ours)	82%	1.28 m

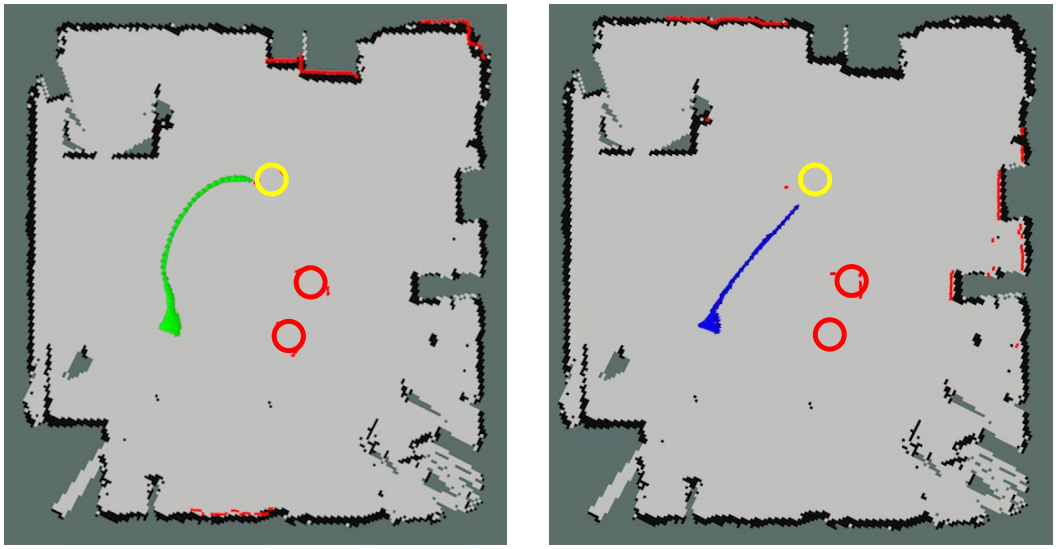
Table 5. The average response time of each module.

Module	Response Time
Keywords Detection	101 ms
Voiceprint Recognition	216 ms
Face Recognition	91 ms
DoA Estimation	160 ms

4.2.2. The Accuracy of Speech Recognition at the End of Tracking

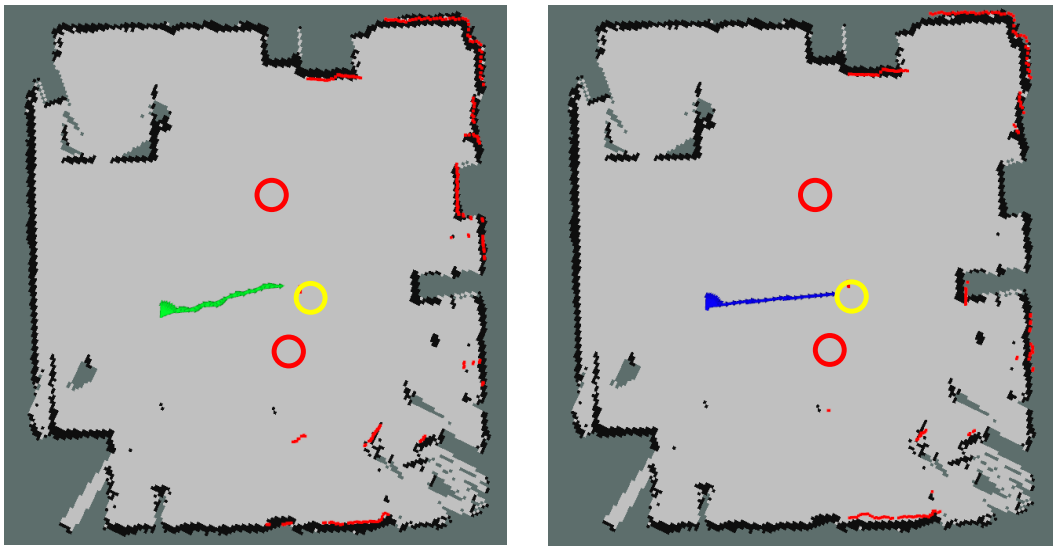
For the second challenge in the cocktail party problem, the clarity of the target audio can be reflected by the accuracy of speech recognition. To verify the effectiveness of our motion module, the WER was calculated at the end of tracking. Two strategies were applied to control the robot to track the target, the robot went directly to the target or the

robot moved relying on the motion module. Figure 6 shows the trajectories of the robot in two different episode settings, one with the target speaker outside of the interference sources and the other with the target speaker in the middle of the interference sources. It can be seen that the motion module can control the robot to move it toward the target while keeping away from other interfering sound sources in the surroundings. Table 6 reports the average WERs for these two strategies. Obviously, the WER is lower when the robot was guided by our motion module, implying that AV-SSLT can improve the quality of captured speech audio under the control of the motion module.



(i) (ii)

(**a**)



(i) (ii)

(**b**)

Figure 6. The trajectories of the robot in two different episodes. (**i**) The trajectory of the robot moving under the control of the motion module and (**ii**) the trajectory of the robot moving directly towards the sound source. The yellow circle indicates the target speaker and the red circles represent the interfering sound sources. (**a**) The trajectories of the robot in the situation where the target speaker was on the outside of the interfering sound sources. (**b**) The trajectories of the robot in the situation where the target speaker was in the middle of the interfering sound sources.

Table 6. Word Error Rates (WERs) of speech recognition at the end of tracking.

Method	WER
Going directly to the target	32%
Moving with motion module	18%

5. Discussion

Although the success rate of our system in locating the specific sound source is significantly higher compared to the methods without using voiceprint and face features, the overall success rate still has room for improvement. By reviewing the process of the experiments, we identified two main reasons for the failure as follows.

The voiceprint features extracted directly from the mixed signal sometimes cannot be matched with the features in the database. This happened particularly when the interfering signal in the environment was dominant. In other words, the ambient noise was much louder than the desired speech. At this point, it was difficult to extract the exact voiceprint of the target speaker and to obtain the correct face features of the target from the database, leading to the failure in localization and tracking of the target. We hold the view that differential microphone arrays can address this difficulty to some extent. In detail, by designing a specific beam pattern of the arrays, it is expected that the response from the non-target direction will be significantly reduced.

In order to ensure the accuracy of face recognition, a high threshold was set for the face verification process. This resulted in the robot being unable to properly detect and recognize the target face when the face was in an extreme pose, thereby failing to locate and track the target speaker. To deal with this problem, an additional robot motion controller needs to be designed to equip the robot with the ability to actively move towards the front of a person to catch the target's face when there is no frontal human face in the FOVs of the camera, for example, when the target speakers have their backs to the robot.

6. Conclusions

In this paper, we propose AV-SSLT to localize and track the specific sound source in a cocktail party environment. In order to focus on the target of interest, we present a way by combining voiceprint recognition and face recognition to locate the sound source. To obtain high-quality audio of the target for downstream tasks such as speech recognition, we designed a motion module to control the robot's movement towards the target while keeping away from other sound sources in the environment. The effectiveness of AV-SSLT has been fully corroborated via extensive practical experiments, which is demonstrated by the higher success rate in locating a specific target, the closer distance to the target at the end of tracking and the lower WER of speech recognition at the end of tracking than its counterparts. However, we have to admit that our system fails in some cases, such as when the target speaker has their back to the camera. Meanwhile, there is still room for improvement in the accuracy of speech recognition. Our future work will focus on improving the accuracy of face recognition when the face is in extreme poses to increase the success rate of specific target localization and on designing the geometry of the microphone array to reduce the interference of noises from other directions.

Author Contributions: Supervision, L.Z.; Writing—original draft, Z.S.; Writing—review & editing, L.Z. and D.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant 62272343 and Grant 61973235; in part by the Shanghai Science and Technology Innovation Plan under Grant 20510760400; in part by the Shuguang Program of Shanghai Education Development Foundation and Shanghai Municipal Education Commission under Grant 21SG23; and in part by the Fundamental Research Funds for the Central Universities.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nakadai, K.; Lourens, T.; Okuno, H.G.; Kitano, H. Active audition for humanoid. In Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on on Innovative Applications of Artificial Intelligence, Austin, TX, USA, 30 July–3 August 2000; pp. 832–839.
2. Grumiaux, P.A.; Kitić, S.; Girin, L.; Guérin, A. A survey of sound source localization with deep learning methods. *J. Acoust. Soc. Am.* **2022**, *152*, 107–151. [CrossRef] [PubMed]
3. Rahimi, A.; Afouras, T.; Zisserman, A. Reading to listen at the cocktail party: Multi-modal speech separation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 10493–10502.
4. Zhu, Q.S.; Zhang, J.; Zhang, Z.Q.; Wu, M.H.; Fang, X.; Dai, L.R. A noise-robust self-supervised pre-training model based speech representation learning for automatic speech recognition. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 3174–3178.
5. Cherry, E.C. Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.* **1953**, *25*, 975–979. [CrossRef]
6. Qian, Y.M.; Weng, C.; Chang, X.K.; Wang, S.; Yu, D. Past review, current progress and challenges ahead on the cocktail party problem. *Front. Inf. Technol. Electron. Eng.* **2018**, *19*, 40–63. [CrossRef]
7. Chiariotti, P.; Martarelli, M.; Castellini, P. Acoustic beamforming for noise source localization–Reviews, methodology and applications. *Mech. Syst. Signal Process.* **2019**, *120*, 422–448. [CrossRef]
8. Grondin, F.; Michaud, F. Lightweight and optimized sound source localization and tracking methods for open and closed microphone array configurations. *Robot. Auton. Syst.* **2019**, *113*, 63–80. [CrossRef]
9. Michaud, S.; Faucher, S.; Grondin, F.; Lauzon, J.S.; Labbé, M.; Létourneau, D.; Ferland, F.; Michaud, F. 3D localization of a sound source using mobile microphone arrays referenced by SLAM. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 24 October 2020–24 January 2021; pp. 10402–10407.
10. Emery, B.M.; Jadidi, M.G.; Nakamura, K.; Miro, J.V. An audio–visual solution to sound source localization and tracking with applications to HRI. In Proceedings of the 8th Asian Conference on Refrigeration and Air-Conditioning, Taipei, Taiwan, 15–17 May 2016; pp. 268–277.
11. Chau, A.; Sekiguchi, K.; Nugraha, A.A.; Yoshii, K.; Funakoshi, K. Audio–visual SLAM towards human tracking and human–robot interaction in indoor environments. In Proceedings of the 2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), New Delhi, India, 14–18 October 2019; pp. 1–8.
12. Zhang, T.; Zhang, H.; Li, X.; Chen, J.; lam, T.L.; Vijayakumar, S. AcousticFusion: Fusing sound source localization to visual SLAM in dynamic environments. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September–1 October 2021; pp. 6868–6875.
13. Ephrat, A.; Mosseri, I.; Lang, O.; Dekel, T.; Wilson, K.; Hassidim, A.; Freeman, W.; Rubinstein, M. Looking to listen at the cocktail party: A speaker-independent audio–visual model for speech separation. *arXiv* **2018**, arXiv:1804.03619.
14. Gu, R.; Zhang, S.X.; Xu, Y.; Chen, L.; Zou, Y.; Yu, D. Multi-modal multi-channel target speech separation. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 530–541. [CrossRef]
15. Gao, R.; Grauman, K. Visualvoice: Audio–visual speech separation with cross-modal consistency. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 15490–15500.
16. Vincent, E.; Sini, A.; Charpillet, F. Audio source localization by optimal control of a mobile robot. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015; pp. 5630–5634.
17. Nguyen, Q.V.; Colas, F.; Vincent, E.; Charpillet, F. Long-term robot motion planning for active sound source localization with Monte Carlo tree search. In Proceedings of the 2017 Hands-free Speech Communications and Microphone Arrays (HSCMA), San Francisco, CA, USA, 1–3 March 2017; pp. 61–65.
18. Schmidt, R. Multiple emitter location and signal parameter estimation. *IEEE Trans. Antennas Propag.* **1986**, *34*, 276–280. [CrossRef]
19. Nakamura, K.; Nakadai, K.; Ince, G. Real-time super-resolution sound source localization for robots. In Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, Vilamoura-Algarve, Portugal, 7–12 October 2012; pp. 694–699.
20. Shimada, K.; Koyama, Y.; Takahashi, S.; Takahashi, N.; Tsunoo, E.; Mitsufuji, Y. Multi-ACCDOA: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 316–320.

21. Adavanne, S.; Politis, A.; Nikunen, J.; Virtanen, T. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE J. Sel. Top. Signal Process.* **2018**, *13*, 34–48. [CrossRef]
22. Masuyama, Y.; Bando, Y.; Yatabe, K.; Sasaki, Y.; Onishi, M.; Oikawa, Y. Self-supervised neural audio–visual sound source localization via probabilistic spatial modeling. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 24 October 2020–24 January 2021; pp. 4848–4854.
23. Chen, J.; Takashima, R.; Gou, X.; Zhang, Z.; Xu, X.; Takiguchi, T.; Hancock, E.R. Multimodal fusion for indoor sound source localization. *Pattern Recognit.* **2021**, *115*, 107906. [CrossRef]
24. Politis, A.; Mesaros, A.; Adavanne, S.; Heittola, T.; Virtanen, T. Overview and evaluation of sound event localization and detection in DCASE 2019. *IEEE Trans. Audio Speech, Lang. Process.* **2020**, *29*, 684–698. [CrossRef]
25. Guizzo, E.; Gramaccioni, R.F.; Jamili, S.; Marinoni, C.; Massaro, E.; Medaglia, C.; Nachira, G.; Nucciarelli, L.; Paglialunga, L.; Pennese, M.; et al. L3DAS21 challenge: Machine learning for 3D audio signal processing. In Proceedings of the 2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP), Gold Coast, Australia, 25–28 October 2021; pp. 1–6.
26. Guizzo, E.; Marinoni, C.; Pennese, M.; Ren, X.; Zheng, X.; Zhang, C.; Masiero, B.; Uncini, A.; Comminiello, D. L3DAS22 challenge: Learning 3D audio sources in a real office environment. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 9186–9190.
27. Sergiyenko, O.Y.; Ivanov, M.; Tyrsa, V.V.; Kartashov, V.M.; Rivas-López, M.; Hernández-Balbuena, D.; Flores-Fuentes, W.; Rodríguez-Quiñonez, J.C.; Nieto-Hipólito, J.I.; Hernandez, W.; et al. Data transferring model determination in robotic group. *Robot. Auton. Syst.* **2016**, *83*, 251–260. [CrossRef]
28. Sergiyenko, O.Y.; Tyrsa, V.V. 3D optical machine vision sensors with intelligent data management for robotic swarm navigation improvement. *IEEE Sens. J.* **2020**, *21*, 11262–11274. [CrossRef]
29. Nguyen, Q.V.; Colas, F.; Vincent, E.; Charpillet, F. Motion planning for robot audition. *Auton. Robot.* **2019**, *43*, 2293–2317. [CrossRef]
30. Schymura, C.; Grajales, J.D.R.; Kolossa, D. Monte Carlo exploration for active binaural localization. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 491–495.
31. Bustamante, G.; Danès, P. Multi-step-ahead information-based feedback control for active binaural localization. In Proceedings of the International Conference on Intelligent Robots and Systems, Vancouver, BC, Canada, 24–28 September 2017; pp. 6729–6734.
32. Bustamante, G.; Danes, P.; Forgue, T.; Podlubne, A.; Manhès, J. An information based feedback control for audio-motor binaural localization. *Auton. Robot.* **2018**, *42*, 477–490. [CrossRef]
33. Sewtz, M.; Bodenmüller, T.; Triebel, R. Robust MUSIC-based sound source localization in reverberant and echoic environments. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 24 October 2020–24 January 2021; pp. 2474–2480.
34. Graves, A. Sequence transduction with recurrent neural networks. *arXiv* **2012**, arXiv:1211.3711.
35. Punjabi, S.; Arsikere, H.; Raeesy, Z.; Chandak, C.; Bhave, N.; Bansal, A.; Müller, M.; Murillo, S.; Rastrow, A.; Stolcke, A.; et al. Joint ASR and language identification using RNN-T: An efficient approach to dynamic language switching. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 7218–7222.
36. Saon, G.; Tüske, Z.; Bolanos, D.; Kingsbury, B. Advancing RNN transducer technology for speech recognition. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 5654–5658.
37. Cheng, S.; Shen, Y.; Wang, D. Target speaker extraction by fusing voiceprint features. *Appl. Sci.* **2022**, *12*, 8152. [CrossRef]
38. Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4690–4699.
39. Suzuki, Y.; Asano, F.; Kim, H.Y.; Sone, T. An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses. *J. Acoust. Soc. Am.* **1995**, *97*, 1119–1123. [CrossRef]
40. Benesty, J.; Chen, G.; Huang, Y. *Microphone Array Signal Processing*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2008.
41. Politis, A.; Shimada, K.; Sudarsanam, P.; Adavanne, S.; Krause, D.; Koyama, Y.; Takahashi, N.; Takahashi, S.; Mitsufuji, Y.; Virtanen, T. STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. *arXiv* **2022**, arXiv:2206.01948.
42. Politis, A.; Adavanne, S.; Krause, D.; Deleforge, A.; Srivastava, P.; Virtanen, T. A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection. *arXiv* **2021**, arXiv:2106.06999.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.