drones

MDPI

Article

Placement Planning for Sound Source Tracking in Active Drone Audition

Taiki Yamada [1,*], Katsutoshi Itoyama [1,2], Kenji Nishida [1] and Kazuhiro Nakadai [1]

[1] Department of Systems and Control Engineering, School of Engineering, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan; itoyama@ra.sc.e.titech.ac.jp (K.I.); nishida@ra.sc.e.titech.ac.jp (K.N.); nakadai@ra.sc.e.titech.ac.jp (K.N.)
[2] Honda Research Institute Japan Co., Ltd., 8-1 Honcho, Wako, Saitama 351-0188, Japan
* Correspondence: yamada@ra.sc.e.titech.ac.jp

Abstract: This paper addresses a placement planning method for drones to improve the performance of source tracking by multiple drones equipped with microphone arrays. By equipping the drone with a microphone array, the drone will be able to locate the person in need of rescue, and by deploying multiple drones, the 3D location of the sound source can be estimated. However, effective drone placement for sound source tracking has not been well explored. Therefore, this paper proposes a new drone placement planning method to improve the performance of sound source tracking. By placing multiple drones close to the sound source with multiple angles, it is expected that tracking will be performed with small variance. The placement planning algorithm is also extended to be applicable to multiple sound sources. Through numerical simulations, it is confirmed that the proposed method reduces the sound source tracking error. In conclusion, the contribution of this research is to extend the field of drone audition to active drone audition that allows drones to move by themselves to achieve better tracking results.

Keywords: drone audition; sound source tracking; action planning

Citation: Yamada, T.; Itoyama, K; Nishida, K; Nakadai, K; Placement Planning for Sound Source Tracking in Active Drone Audition. *Drones* **2023**, *7*, 405. https://doi.org/10.3390/drones7070405

Academic Editor: Diego González-Aguilera

Received: 16 May 2023
Revised: 14 June 2023
Accepted: 15 June 2023
Published: 21 June 2023

1. Introduction

Audio signal processing has been playing an important role which enables robots and drones to recognize the external world. The field known as computational auditory scene analysis (CASA) [1,2] has been extensively studied for the purpose of understanding acoustic scenes and has a wide range of applications. This paper deals with computational auditory scene analysis (CASA) in drone applications, which is referred to as drone audition, and especially discusses sound source localization, which is one of the basic concepts in CASA, using drones. For example, recognition of spatial information of sound sources will help drones search for people calling for help in the disaster sites when they are covered by rubble and are not visible [3–6]. In the field of robot/drone audition, microphone arrays are generally used for audio sensing. A microphone array is an aggregation of microphones and it is useful for estimating the direction of sound sources by utilizing the phase difference between each input and the microphones [7–9]. When we equip a microphone array for a drone, we can locate sound sources from the sky, which could be used in people searching tasks in disaster sites. Existing work has shown both sound source localization for stationary sound sources [3,10] and sound source tracking for moving sound sources [11,12]. While existing work has shown significant tracking results, the drones are normally assumed to be either staying still or flying in a fixed route. Yamada et al. [13] have shown that tracking results for any sound source tracking method can be affected by how the drones and the microphone arrays attached are placed. Therefore, realizing optimal microphone array placement to achieve improved sound source tracking performance is a remaining challenge. With the mobility of drones, we believe we can move the microphone arrays to realize such formation. This paper pursues the concept of "active drone audition"

in which the drones act in a manner that they can efficiently perform sound source tracking (Figure 1). As a stepping stone toward realizing active drone audition, we considered what would constitute an optimal microphone array arrangement for effective tracking. We proposed a method based on the assumption that an ideal placement is one in which each drone is able to approach the sound source while simultaneously maintaining a multi-angle observation. Furthermore, to extend our proposal to the tracking of multiple sound sources, we quantified the contribution of each drone to the tracking of each sound source, striving to improve the efficiency of sound source tracking. The proposed placement optimization method is evaluated through various tracking methods to understand whether it improves the tracking performance with generality among the tracking methods.

The remainder of this paper is organized as follows: Section 2 discusses related work, Section 3 presents our method, Section 4 provides an evaluation of our approach, and finally, Section 5 concludes the paper and suggests areas for future research.

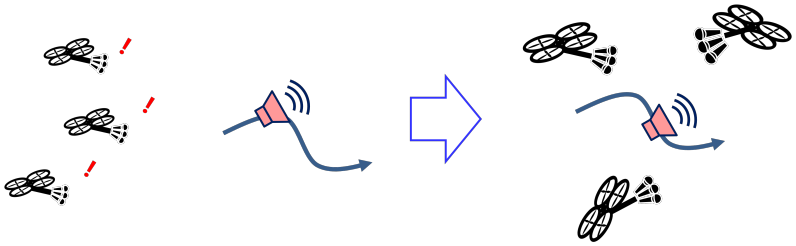

Figure 1. Concept image of active drone audition. Since microphone arrays estimate the sound source direction, it is difficult to localize the source location by nearby drones. Hence, drones should try to form a shape to suppress the tracking error by making a small localization variance in a specific direction.

2. Related Work

In this section, we present an overview of the related studies that have contributed to sound source tracking using drones and state the position of this research.

2.1. Definition of Active Drone Audition

In this section, the terms "sound source tracking" and "active drone audition" are explained. In the field of array processing, "sound source localization" sometimes means "estimation of the sound source direction", and in this perspective, "sound source tracking" indicates "tracking the sound source direction". However, this paper is focusing on estimation of sound source 'location'. Therefore, when "sound source tracking" is used in this paper, it indicates "tracking the 3D sound source location". Furthermore, we "track" the sound source location by repeating the location estimation through time. This paper also considers tracking as an estimation using not only the current observation but also the previous sound source state.

In this paper, a novel concept, "active drone audition", is introduced. While the existing work of "drone audition" only concentrates on signal processing, "active drone audition" also utilizes the operation of drones for auditory scene analysis. In addition, this paper also uses the term "active sound source tracking" or "active tracking" when the topic is specific to sound source tracking.

2.2. Sound Source Tracking Using Microphone Arrays

This section introduces sound source location and tracking methods using microphone arrays. Array processing aims to capture the phase difference or time difference between microphone arrays, and from these differences, we can also estimate the direction of sound sources from the phase/time difference. As mentioned in Section 2.1, some studies aim to estimate the sound source direction, while others aim to estimate the sound source location. Since most location estimation methods are based on direction estimation, this section first reviews direction estimation methods and then reviews location estimation

methods. Existing work has proposed several approaches to process multiple channels of audio signals. Generalized cross-correlation phase transform method (GCC-PHAT) [8], also known as cross-power spectrum phase (CSP) analysis, calculates the correlation between two channel recordings and obtain the time difference of arrival (TDOA) between channels, where the source direction can be estimated from the TDOA. The beamforming method [7], or steered-response power phase transform (SRP-PHAT) method, in other words, is a linear spatial filtering method that calculates the steered-response power (SRP), which peaks in the sound source direction. GCC-PHAT is a direction estimation method using two microphones, while SRP-PHAT can use more microphones, and mathematically, SRP-PHAT can be expressed as the sum of GCC-PHAT for all pairs of arrays of microphones. Multiple signal classification (MUSIC) [9] is known as a subspace method that analyzes the correlation matrix derived in the time–frequency domain, calculates the orthogonality between the noise subspace and the signal assuming that the noise subspace is orthogonal to the target signal space. Similar to beamforming, MUSIC method calculates a spatial spectrum called MUSIC spectrum which lies on the azimuth-elevation plane, and the direction in which the MUSIC spectrum has a peak value is the direction of the sound source. In the field of drone audition, the MUSIC method is frequently used, and several variations have been proposed to tackle severe drone noise.

Expanding on the "direction" estimation method described in the previous section, this section describes a method for estimating the "location" of the sound source. Since microphone arrays are generally used for direction estimation, additional measures should be taken to obtain the sound source location. One approach is to move around the sound source and observe the sound source from multiple points of view [14–16]. By projecting directional spectrum to a 2D/3D field, we can obtain the locational spectrum and estimate the source location. However, this approach is time consuming and cannot be applied to moving sound sources. Therefore, using multiple microphone arrays is a common approach for capturing the sound source location with a single direction estimation. Triangulation using multiple arrays is one popular way to accomplish this and is widely used for many purposes. For indoor examples, localizing speakers with multiple microphone arrays have been done [17–19], and for outdoor examples, triangulation methods have also been applied to ecological surveys such as detecting and locating birdsong [20–22]. In terms of drone audition, rather than performing triangulation with multiple arrays, the sound source (the person calling for help) is located by estimating the intersection point of the estimated direction and a virtual terrain surface [6,10,23,24].

When the sound source is moving, estimating its location might be difficult especially for single-microphone arrays as they cannot estimate the distance to the sound source. Although there are related studies which have performed sound source tracking with a single array and showed its effectiveness in indoor environments [25–27], the variance in the distance direction can be seen. Therefore, observing the sound source by distributed microphone arrays is a popular method to track the sound source location. Several tracking strategies have been conducted for sound source localization. As mentioned in the previous section, calculating the triangulation points from estimated directions of multiple arrays is a popular way to obtain the sound source, and by applying the average of triangulation points to Kalman filtering methods, we can obtain the sound source trajectory [28–30]. Since Kalman filters have the weakness, they can only be applied to the average triangulation point, and Yamada et al. have proposed a tracking strategy using Gaussian sum filtering. By assuming triangulation points as a Gaussian mixture, we can hold the triangulation points as candidates of the sound source location and suppress the impact of outliers [11]. There are also other approaches that estimate the source location with multiple arrays but do not perform triangulation as the source location is estimated through particle filtering as a latent variable. The particle weights can be determined by the direction estimation error [31,32] or the estimated locational likelihood distribution [12]. Since source direction estimation methods are highly sensitive to drone noise and also have their weakness against discretization error, it is still necessary to develop tracking methods that are robust to both

drone noise and discretization error. Especially discretization error harms the tracking when the drones are close to each other and the estimated direction to the source is parallel to each other. Therefore, the weakness of using multiple arrays might be overcome by conducting proper placement of drones. Sensor placement have been studied for target localization. For example, in the field of audio signal processing, microphone arrangement in a microphone array has been studied to improve direction estimation performance [33,34]. In the field of image processing, optimizing the camera placements is an area of active research that influences sight coverage [35] and object tracking [36]. The performance improvement of sound source tracking by optimizing the placement of multiple microphone arrays has not yet been demonstrated. However, it is expected that the drone's mobility will enable it to optimize the placement of microphone arrays for tracking moving sound sources, which is examined in this study.

2.3. Robot/Drone Activities in Robot/Drone Audition

In most of the studies described in the previous section, the placement of the microphone array or layout of the microphone arrays have not been considered. In the research field of robots and drones, the placement of robots/drones is hardly tied together with the tracking results. However, tracking performance will be affected by how the microphone arrays are placed [13], and robots/drones should move around to improve their tracking, especially when sound source tracking is their main task. Therefore, this paper discusses about "active sound source tracking" where robots/drones move by themselves in order to improve their tracking performance. In active sound source tracking, robots/drones have their action rules tied with their localization results.

For direction estimation, there are previous studies that move the orientation of the microphone array to improve direction estimation results [37–40]. A binaural robot would adjust its head positioning to eliminate front-back ambiguity. For location estimation in robot/drone audition, as said in the Section 2.2, there are research that control robots to move around the sound source to observe the sound source from multiple perspectives and map the sound source location [10,14–16]. To perform this move-around-the-sound-source approach, either providing prior information to the robot or allowing the robot to explore the entire field to capture an approximate acoustic energy map is required [41]. Although these methods are helpful to localize the source location, robot/drone actions are not linked to their estimation results in this research, which still cannot be said to be "active" in this paper. Unplanned movements or exhaustive exploration may not be effective for moving sound sources as the robot or drone may lose track of the sound source. This is why a targeted activity is necessary for sound source tracking. Motion planning methods for robot audition have been reported to simulate making multiple hypotheses of future robot positions and taking the optimal position that minimizes the entropy of the belief of the source location [42–44]. However, these methods are implemented for stationary sound sources, and active sound source tracking still presents a wide scope for research.

In response to these previous studies, the contribution of this paper is introducing an active sound source tracking system for drone audition, designed to improve tracking performance by optimizing drone placement over time. A brief summary of the above related work of active audition and the positioning of this paper is described in Table 1.

Table 1. Related work classified by application and utilization of robot/drone activity

	Robot Audition	**Drone Audition**
Not utilizing activity	• Source localization (direction) [45–47] • Source localization (location) [14,15]	• Source localization (direction) [5,48] • Source localization (location) [10–12,24]
Utilizing activity	• Source localization (direction) [37,38,40] • Sound source separation [49]	• Source localization (location) [This paper]

3. Method

In this section, we introduce a placement optimization method for active sound source tracking using multiple microphone arrays. We consider a scenario where N drones will track S sound sources. Hereafter, $i, j \in \{1, \ldots, N\}$ is the index of the microphone array mounted on the drone, and $k \in \{1, \ldots, S\}$ is the index of the sound source. The entire procedure of the tracking is illustrated in Figure 2. Note that this placement optimization method can be applied to any kind of sound source tracking method that uses multiple microphone arrays. In this paper we use PAFIM as an example to understand the whole procedure [12]. The proposed method applied to other sound source tracking methods is shown in the evaluation. PAFIM is a sound source tracking method that estimates the locational likelihood distribution by integrating MUSIC spectrum, a directional spectrum computed through the MUSIC method [9]. The following sections will explain each procedure described in Figure 2.

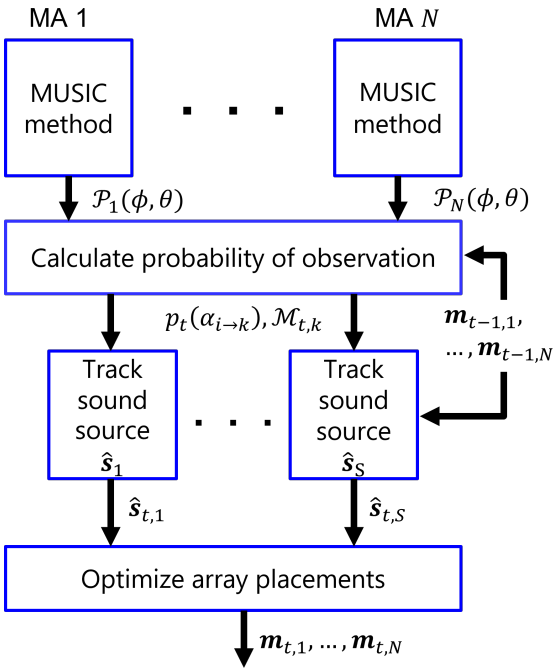

Figure 2. Procedure of sound source tracking with the proposed placement planning method for one time step. The sound source trajectory will be obtained by repeating this procedure.

3.1. MUSIC Method

With multiple channels of audio signals, we can estimate the phase difference or time difference between channels and we can also estimate the direction from them. Existing work has proposed several approaches to process multiple channels of audio signals. The generalized cross-correlation phase transform method (GCC-PHAT) [8], also known as cross-power spectrum phase (CSP) analysis, calculates the correlation between two channel recordings and obtains the time difference of arrival (TDOA) between channels, where the source direction can be estimated from the TDOA. The beamforming method [7], or steered-response power phase transform (SRP-PHAT) method in other words, is a linear spatial filtering method that calculates the steered-response power (SRP), which peaks at the sound source direction. GCC-PHAT is a direction estimation method using two microphones, while SRP-PHAT can use more microphones, and mathematically speaking, SRP-PHAT can be expressed as the sum of GCC-PHAT for all pairs of arrays of microphones. Multiple signal classification (MUSIC) [9] is known as a subspace method that analyzes the correlation matrix derived in the time-frequency domain and calculates the orthogonality between the noise subspace and the signal with the assumption that the noise subspace

is orthogonal to the target signal space. The MUSIC method shares similarities with beamforming, which produces a spatial spectrum known as the MUSIC spectrum. This spectrum lies within the azimuth-elevation plane, with the direction of the sound source corresponding to the peak value of the MUSIC spectrum. Within drone audition, the MUSIC method is commonly applied, with various adaptations proposed to handle the challenge of intense drone noise. The simplest form of MUSIC, known as standard eigenvalue decomposition MUSIC (SEVD-MUSIC), is widely used. To implement the MUSIC method, the acoustic signal recorded in the time domain must be converted into a time–frequency domain signal using short time Fourier transform (STFT). Given $\boldsymbol{z}(\omega, f) \in \mathbb{C}^L$ as the recorded signal in the time–frequency domain, where ω signifies the frequency bin index and f represents the time frame, the correlation matrix $\boldsymbol{R}(\omega, f)$ is then defined as

$$\boldsymbol{R}(\omega, f) = \frac{1}{T_R} \sum_{\tau=f}^{f+T_R-1} \boldsymbol{z}(\omega, \tau)\boldsymbol{z}^H(\omega, \tau) \tag{1}$$

In this equation, T_R is the number of frames used to create the correlation matrix $\boldsymbol{R}(\omega, f)$, while $\boldsymbol{z}^H$ denotes the complex conjugate transpose of $\boldsymbol{z}$. Using eigenvalue decomposition, the correlation matrix $\boldsymbol{R}(\omega, f)$ can be represented as

$$\boldsymbol{R}(\omega, f) = \boldsymbol{E}(\omega, f)\boldsymbol{\Lambda}(\omega, f)\boldsymbol{E}^{-1}(\omega, f) \tag{2}$$

After computing $\boldsymbol{E}$, the spatial spectrum is obtained as follows [9]:

$$\mathcal{P}(\omega, \phi, \theta, f) = \frac{|\boldsymbol{a}^H(\omega, \phi, \theta)\boldsymbol{a}(\omega, \phi, \theta)|}{\sum_{i=N_e+1}^{L} |\boldsymbol{a}^H(\omega, \phi, \theta)\boldsymbol{e}_i(\omega, f)|} \tag{3}$$

This is also referred to as the MUSIC spectrum, where $\boldsymbol{a}$ is the steering vector; ϕ, θ denote the source azimuth and elevation as seen from the microphone array center; N_e is the count of the target sound sources; and $\boldsymbol{e}_i$ is the i-th eigenvector. The eigenvectors are arranged such that $\lambda_1 \geq \ldots \geq \lambda_L$ is upheld, where λ_i is the corresponding eigenvalue to the eigenvector $\boldsymbol{e}_i$. If ϕ, θ precisely denote the direction of the sound source, the term $\boldsymbol{a}^H(\omega, \phi, \theta)\boldsymbol{e}_i(\omega, f)$ ideally equals zero, leading to a sharp peak in $\mathcal{P}(\omega, \phi, \theta, f)$ in that direction. As a result, the estimated direction of the sound source, $(\hat{\phi}(f), \hat{\theta}(f))$, can be determined by identifying the direction where the average MUSIC spectrum $\bar{\mathcal{P}}(\phi, \theta, f)$ peaks, which is defined as:

$$\bar{\mathcal{P}}(\phi, \theta, f) = \frac{1}{\omega_H - \omega_L + 1} \sum_{\omega=\omega_L}^{\omega_H} \mathcal{P}(\omega, \phi, \theta, f) \tag{4}$$

Here, $\bar{\mathcal{P}}(\phi, \theta, f)$ represents the average spatial spectrum of each frequency, while ω_L, ω_H respectively denote the minimum and maximum frequency bin indices. It is important to note that the steering vector $\boldsymbol{a}$ is generally defined for each potential direction. For instance, $\boldsymbol{a}$ is typically arranged as a table according to the direction, incrementing in 5 degree intervals.

3.2. Particle Filtering with Integrated MUSIC (PAFIM)

This section introduces a method for tracking a sound source based on location likelihood estimation, known as particle filtering with integrated MUSIC (PAFIM) proposed by Yamada et al. [12]. The equations in this section is adopted from the work of them. PAFIM works by integrating direction likelihood distributions derived from source direction estimation. Typically, the MUSIC method calculates the likelihood $\mathcal{P}(\phi, \theta)$ for each azimuth ϕ and elevation θ using Equation (4) and assumes the direction yielding the maximum $\mathcal{P}(\phi, \theta)$ to be the sound source direction. The unique aspect of PAFIM is that it transforms the direction likelihood $\mathcal{P}(\phi, \theta)$ from each microphone array into the likelihood of a three-dimensional location, thereby estimating the sound source location. Let $\mathcal{P}_i(\phi, \theta)$ be the

MUSIC spectrum derived from the i-th microphone array. The likelihood distribution of the sound location can be represented by simply summing $\mathcal{P}_i(\phi, \theta)$. Given an arbitrary three-dimensional location $\boldsymbol{x}$, and (ϕ_i, θ_i) as the direction from the i-th microphone array to point $\boldsymbol{x}$, the location likelihood L at location x can be described as

$$L(\boldsymbol{x}) = \sum_i \mathcal{P}_i(\tilde{\phi}_i^{\text{round}}, \tilde{\theta}_i^{\text{round}}) \tag{5}$$

$$\tilde{\phi}_i^{\text{round}} = \text{round}(\tilde{\phi}_i), \qquad \tilde{\theta}_i^{\text{round}} = \text{round}(\tilde{\theta}_i) \tag{6}$$

$$\begin{bmatrix} \cos\tilde{\phi}_i \cos\tilde{\phi}_i \\ \sin\tilde{\phi}_i \cos\tilde{\phi}_i \\ \sin\tilde{\phi}_i \end{bmatrix} = \boldsymbol{R}_i^{-1} \begin{bmatrix} \cos\phi_i \cos\phi_i \\ \sin\phi_i \cos\phi_i \\ \sin\phi_i \end{bmatrix} \tag{7}$$

In this context, round$(\cdot)$ denotes a function that rounds the direction according to the resolution of the transfer function $\boldsymbol{a}(\omega, \phi, \theta)$, and $\boldsymbol{R}_i$ represents the rotation matrix depicting the posture of the i-th microphone array. In essence, the location likelihood $L(x)$ is the summation of direction likelihoods corresponding to the directions pointing towards x as observed from each microphone array.

3.3. Active Placement Optimization

When using multiple microphone arrays for source location tracking, the placement of the microphone arrays may affect the tracking accuracy [13]. For example, when multiple microphone arrays are used to triangulate the source location, the source location estimation error may appear frequently at the array-to-source direction, especially if the microphone arrays are close to each other. In addition, when a microphone array is mounted on a drone, drone noise is always generated close to the microphone array, and the signal-to-noise ratio (SNR) may become significantly smaller as the drone moves away from the sound source. In this section, an algorithm to optimize the placement of microphone arrays is proposed in order to improve the performance of multiple source tracking using a drone swarm equipped with a microphone array. We assume that the following four policies will improve the tracking performance.

a. Microphone arrays should be allocated only to sound sources they can hear if there are multiple sound sources.
b. Pairs of array-to-source directions should be orthogonal with each other to have stable location estimation.
c. Microphone arrays should be close to the sound source to maintain high SNR.
d. Drones should make small movements as possible.

To satisfy these policies, each drone moves according to the Algorithm 1 which is briefly described as follows.

1. Initialize parameters (See Section 3.3.4)
2. Calculate MUSIC spectrum by MUSIC method [9]
3. Calculate probability of observation $p(\alpha_{i\to k}|\boldsymbol{z})$ from Equation (8) and assign microphone arrays to sources to be tracked according to probability of observation.
4. Perform sound source tracking for each sound source with the corresponding microphone array group.
5. Calculate next drone positions by estimated source locations with Equation (11).

Detailed explanation for each process is written in the following sections.

Algorithm 1 Proposed algorithm (For one time step).

Require: $\hat{\boldsymbol{m}^{\text{loc}}}_{i,t}$
for $i = 1, \ldots, N$ **do**
  $\boldsymbol{d}_i, \mathcal{P}(\phi, \theta) \leftarrow$ MUSIC spectrum derived by MUSIC method [9]
end for
for $k = 1, \ldots, K$ **do**
  $\mathcal{M}_k \leftarrow \emptyset$
  for $i = 1, \ldots, N$ **do**
    $p(\alpha_{i \to k} | \boldsymbol{z}) \leftarrow$ Equation (8)
    if $p(\alpha_{i \to k} | \boldsymbol{z}) \geq p_{\text{thre}}$ **then**
      Add i to $\mathcal{M}_k$
    end if
  end for
  $\hat{\boldsymbol{s}}_{k,t} \leftarrow$ Estimated source location by PAFIM (see Section 3.2) with microphone arrays belonging to $\mathcal{M}_k$
end for
$\boldsymbol{m}^{\text{loc}}_{t+1} \leftarrow$ Equation (11)

3.3.1. Probability of Observation Update

When we want to track multiple sound source, we need to be aware that it is almost impossible to capture all sound sources with one microphone array. Therefore, a parameter named probability of observation is defined, and track a sound source k only with arrays that have high probability of observation. Let the event of microphone array i observing sound source k to be $\alpha_{i \to k}$. $\alpha_{i \to k} = 1$ means that array i has observed source k, and $\alpha_{i \to k} = 0$ means that array i has not observed source k. Then, the probability of observation $p(\alpha_{i \to k} | \boldsymbol{z})$ is defined as a probability updated by Equation (8).

$$p(\alpha_{i \to k} | \boldsymbol{z}) = \frac{p(\alpha_{i \to k}) p(\boldsymbol{z} | \alpha_{i \to k})}{\sum p(\alpha_{i \to k}) p(\boldsymbol{z} | \alpha_{i \to k})} \tag{8}$$

In short, probability of observation $p(\alpha_{i \to k} | \boldsymbol{z})$ is the probability whether microphone array i has observed sound source k given the observation $\boldsymbol{z}$. The prior $p(\alpha_{i \to k})$ is the posterior in the previous time step, and likelihood $p(\boldsymbol{z} | \alpha_{i \to k})$ is calculated by Equations (9) and (10).

$$p(\boldsymbol{z} | \alpha_{i \to k}) = \begin{cases} \mathcal{P}_{\text{norm}}(\phi_{i \to k}, \theta_{i \to k}) & (\alpha_{i \to k} = 1) \quad (9) \\ \dfrac{1}{N_\phi N_\theta} & (\alpha_{i \to k} = 0) \quad (10) \end{cases}$$

$\mathcal{P}_{\text{norm}}(\phi, \theta)$ is the MUSIC spectrum normalized so that the sum is 1, and N_ϕ, N_θ are the azimuth bin and elevation bin, respectively. Moreover, $(\phi_{i \to k}, \theta_{i \to k})$ are respectively the azimuth and elevation of sound source k seen from microphone array i, hence $\mathcal{P}_{\text{norm}}(\phi_{i \to k}, \theta_{i \to k})$ is the normalized MUSIC spectrum of the according direction. Since MUSIC spectrum will have a peak at the source direction if correct direction estimation is done, $\mathcal{P}_{\text{norm}}(\phi, \theta)$ is defined as the likelihood distribution when $\alpha_{i \to k} = 1$. When $\alpha_{i \to k} = 0$, the likelihood is defined to be the value when the MUSIC spectral takes the same value for each direction, since there is no general form of MUSIC spectrum when there is no sound source. Hence, the probability of observation $p(\alpha_{i \to k} | \boldsymbol{z})$ will increase to 1 when the MUSIC spectrum of microphone array i peaks towards the direction of source k and vice versa if not. In this paper, probability of observation $p(\alpha_{i \to k} | \boldsymbol{z})$ is calculated for each possible pair of a microphone array and a sound source, and the group of microphone arrays to track sound source k is decided. When the probability of observation of microphone array i to sound source k ($p(\alpha_{i \to k} | \boldsymbol{z})$) is higher than the threshold p_{thre}, microphone array i will be added to a group $\mathcal{M}_k \subseteq \{1, \ldots, N\}$, which will be used in the following tracking and placement planning.

3.3.2. Sound Source Tracking

Sound source tracking is performed by using the microphone array groups $\mathcal{M}_k$ obtained in Section 3.3.1. In this paper, we use PAFIM for the tracking method, but other kinds of tracking methods that use multiple arrays are applicable. In PAFIM, the location likelihood of sound source k is calculated only by the MUSIC spectrum calculated by the arrays belonging to group $\mathcal{M}_k$. See Section 3.2 for further information on PAFIM.

3.3.3. Optimization of Drone Placement

Under the assumption that sound source tracking will perform better when (i) the array-to-source direction are orthogonal to each other and (ii) the microphone arrays are close to the sound source, Equation (11) is formulated. The optimal microphone array placement is computed through Equation (11) and the placement of drones is also determined assuming that the microphone array is firmly fixed to the drone.

$$\underset{\boldsymbol{m}_{t+1}}{\operatorname{argmin}} f(\boldsymbol{m}_{t+1}) + \lambda_g g(\boldsymbol{m}_{t+1}) + \lambda_h h(\boldsymbol{m}_{t+1}) \tag{11}$$

$$\text{s.t.} z_i \geq z_{\text{lim}} \tag{12}$$

$$\boldsymbol{M} q(\boldsymbol{m}) = \boldsymbol{C} \tag{13}$$

$$f(\boldsymbol{m}^{\text{loc}}_{t+1}) = \sum_{k=1}^{S} \sum_{\{i,j\} \in \mathcal{M}_k \ (i \neq j)} \left(\frac{(\boldsymbol{m}^{\text{loc}}_{i,t+1} - \hat{\boldsymbol{s}}_{k,t})^\top (\boldsymbol{m}^{\text{loc}}_{j,t+1} - \hat{\boldsymbol{s}}_{k,t})}{\|\boldsymbol{m}^{\text{loc}}_{i,t+1} - \hat{\boldsymbol{s}}_{k,t}\|_2 \|\boldsymbol{m}^{\text{loc}}_{j,t+1} - \hat{\boldsymbol{s}}_{k,t}\|_2} \right)^2 \tag{14}$$

$$g(\boldsymbol{m}^{\text{loc}}_{t+1}) = \sum_{k=1}^{S} \sum_{i \in \mathcal{M}_k} \|\boldsymbol{m}^{\text{loc}}_{i,t+1} - \hat{\boldsymbol{s}}_{k,t}\|_2 \tag{15}$$

$$h(\boldsymbol{m}^{\text{loc}}_{t+1}) = \sum_{i=1}^{N} \|\boldsymbol{m}^{\text{loc}}_{i,t+1} - \boldsymbol{m}^{\text{loc}}_{i,t}\|_2 \tag{16}$$

where $\boldsymbol{m}_{i,t} \in \mathbb{R}^6$ is the 6D state of the microphone array i at time t which consists of its 3D location $\boldsymbol{m}^{\text{loc}}_{i,t} \in \mathbb{R}^3$ and 3D posture $\boldsymbol{m}^{\text{pos}}_{i,t} \in \mathbb{R}^3$.

$$\boldsymbol{m}_{i,t} = \left[\left(\boldsymbol{m}^{\text{loc}}_{i,t} \right)^\top, \left(\boldsymbol{m}^{\text{loc}}_{i,t} \right)^\top \right]^\top \tag{17}$$

In addition, $\boldsymbol{m}_t = \{\boldsymbol{m}_{1,t}, \ldots, \boldsymbol{m}_{N,t}\}$ is the group of 6D state for all microphone arrays. Equation (12) is the restriction so that the drones will not be too close to the objects on the ground, where z_i is the z-coordinate of drone i and z_{lim} is the lower height limit. Besides, Equation (12) is a constraint of microphone array positions which will be necessary if a drone has more than one array, where $q(\boldsymbol{x})$ represents the respective positions between arrays, and $\boldsymbol{M}, \boldsymbol{C}$ are matrices or vectors that represent the relative position between microphone arrays. Therefore, if there are many constraints of array positions, there is another way to represent Equation (11) by optimizing the drone positions rather than the array positions. In both ways, the optimal microphone array position can be optimized with the almost same computational cost. $\boldsymbol{s}_{k,t} \in \mathbb{R}^3$ is the 3D location of sound source k at time t. Function $f(\boldsymbol{m}^{\text{loc}})$ is the sum of square of cosines where the angles are two array-to-source directions. $f(\boldsymbol{m}^{\text{loc}})$ will be minimized when each pair of directions are orthogonal to each other, which will not let the location likelihood to extend in a particular direction. $g(\boldsymbol{m}^{\text{loc}})$ is the sum of array-to-source distance, and minimizing $g(\boldsymbol{m}^{\text{loc}})$ means drones will try to be close to the sound sources. $h(\boldsymbol{m}^{\text{loc}})$ is the sum of distance between the current location and the previous location of drones, and minimizing means drones will try not to move too much in one time step. λ_g, λ_h are the coefficients on $g(\boldsymbol{m}^{\text{loc}}), h(\boldsymbol{m}^{\text{loc}})$, respectively. The value of $f(\boldsymbol{m}^{\text{loc}})$ is at most limited to $S \cdot {}_N C_2$, while $g(\boldsymbol{m}^{\text{loc}}), h(\boldsymbol{m}^{\text{loc}})$ is the sum of 3D distances, so in general $f(\boldsymbol{m}^{\text{loc}})$ is significantly smaller than $g(\boldsymbol{m}^{\text{loc}}), h(\boldsymbol{m}^{\text{loc}})$. Therefore, coefficients λ_g, λ_h are set to balance the terms; for example they are set to $\lambda_g = 0.01, \lambda_h = 0.0001$ in the evaluation which are decided heuristically. In this paper,

the microphone array is assumed to be mounted protruding from the drone so that the propeller emits drone noise behind the microphone array. Therefore, it is preferable that the sound source is in front of the drone not behind it, so the drone is positioned so that it faces the estimated source location $\hat{\boldsymbol{s}}_k$, and if the microphone array i exceeds the confidence threshold p_{thre} for multiple sources, it faces the average direction to the corresponding sound sources.

3.3.4. Initialization

When executing the above algorithm, the number of sources must be known and the initial value of probability of observation $p(\alpha_{i\to k})$ is required. Therefore, in this paper, S is assumed to be known, and the initial value of probability of observation $p(\alpha_{i\to k})$ is set to 0.5 for all source–microphone array pairs.

4. Evaluation

In this paper, we conducted a series of numerical simulations to evaluate the proposed sound source tracking framework. The objective of these simulations is to evaluate whether the proposed drone placement optimization is improving the tracking performance compared to that of a framework that does not update its drone positions. The simulation environment was set up using MATLAB, and two types of scenarios are considered: (1) single-source tracking and (2) multiple-source tracking.

4.1. Single-Source Tracking

In this simulation, we evaluate the proposed placement planning method by numerical simulation about tracking a single moving sound source. This simulation is focused on evaluating the tracking performance compared with stationary drone placements. In addition, the proposed placement planning is evaluated against various tracking methods that use multiple microphone arrays.

4.1.1. Simulation Settings

Consider using three drones where each drone has a microphone array. The target sound source is moving in a circle with an 11 m radius at a 1.5 m height (Figure 3). The sound source is emitting a 1 kHz sine wave for 46.2 s, which is also just enough time for the sound source to go around the circle. Each drone is equipped with a 16-channel spherical microphone array (Figure 4) recording at 24-bit, 16 kHz. Prerecorded drone noise is added to the recordings so that the SNR is set to -20 dB.

With the simulation settings above, we tested five types of drone placements for comparison. To evaluate the difference between using active placement planning or not, we prepared four placements where all drones are hovering still (See Figure 5).

Inner Place at the inner part of the source trajectory.

Rand1, 2 Uniformly random placements within the area $-15 \geq x, y \geq 15$.

Surround Place at the outer part of the source trajectory.

For all stationary placements, the height of the microphone arrays is $z = z_{\min}$. Finally, we implemented the proposed placement planning method. The initial location of the drones are listed below, and all drones are facing to the sound source.

MA1 $\boldsymbol{x}_1 = [9.0295,\ -2.0000,\ 6.5000]^T$

MA2 $\boldsymbol{x}_2 = [11.0295,\ -2.0000,\ 6.5000]^T$

MA3 $\boldsymbol{x}_3 = [13.0295,\ -2.0000,\ 6.5000]^T$

Note that the drones change their placements for each time step considering the estimated sound source location while the compared placements don't. In this simulation λ and $z_{\min}$ is set to $\lambda_g = \lambda_h = 0.01, z_{\min} = 0.48$.

In addition, to evaluate whether the proposed method is applicable to various tracking methods, four kinds of sound source tracking methods are implemented.

Potamitis04 Calculates the triangulation points based on estimated directions from each microphone array and applies Kalman filtering to the average point of them [29].

Lauzon17 Tracks the sound source by particle filtering where the weight of particles are determined by the angular difference between the direction estimation and the expectation [31].

Yamada20 Converts multiple triangulation points into a Gaussian mixture and applies Gaussian sum filtering to track the sound source [11].

Yamada21 PAFIM introduced in Section 3.2.

While each method uses different sound source direction estimation methods, we applied the MUSIC method [9] to all tracking methods as a direction estimation method. The update of the sound source location is 0.5 s for all methods.

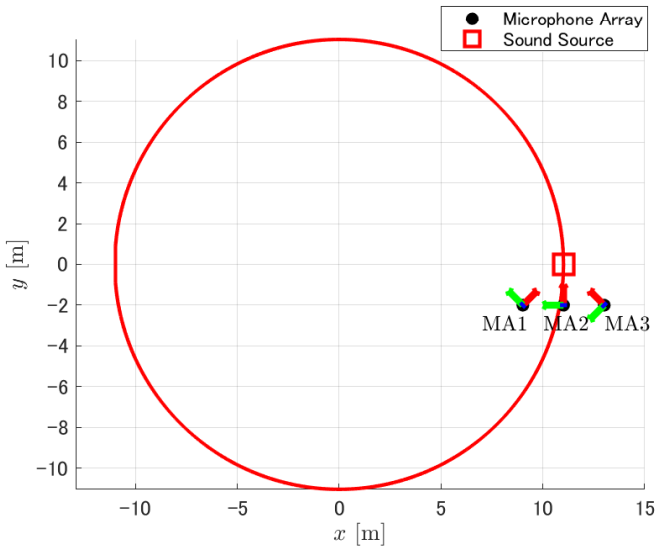

Figure 3. Top view of ground truth sound source trajectory and the initial state of microphone arrays. Each microphone array is indexed as MA1, MA2, and MA3. The red and green arrows starting from each microphone array indicate the longitudinal direction and lateral direction, respectively.

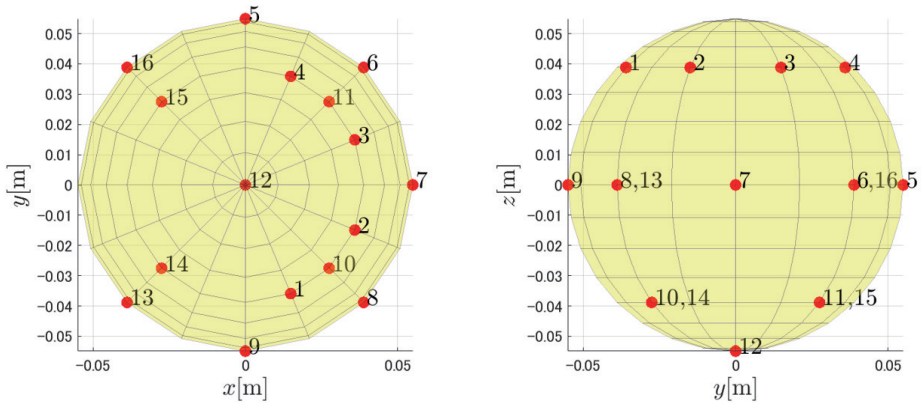

Figure 4. Microphone arrangement of a microphone array: 16 microphones are placed on the surface of a sphere with 110 mm radius.

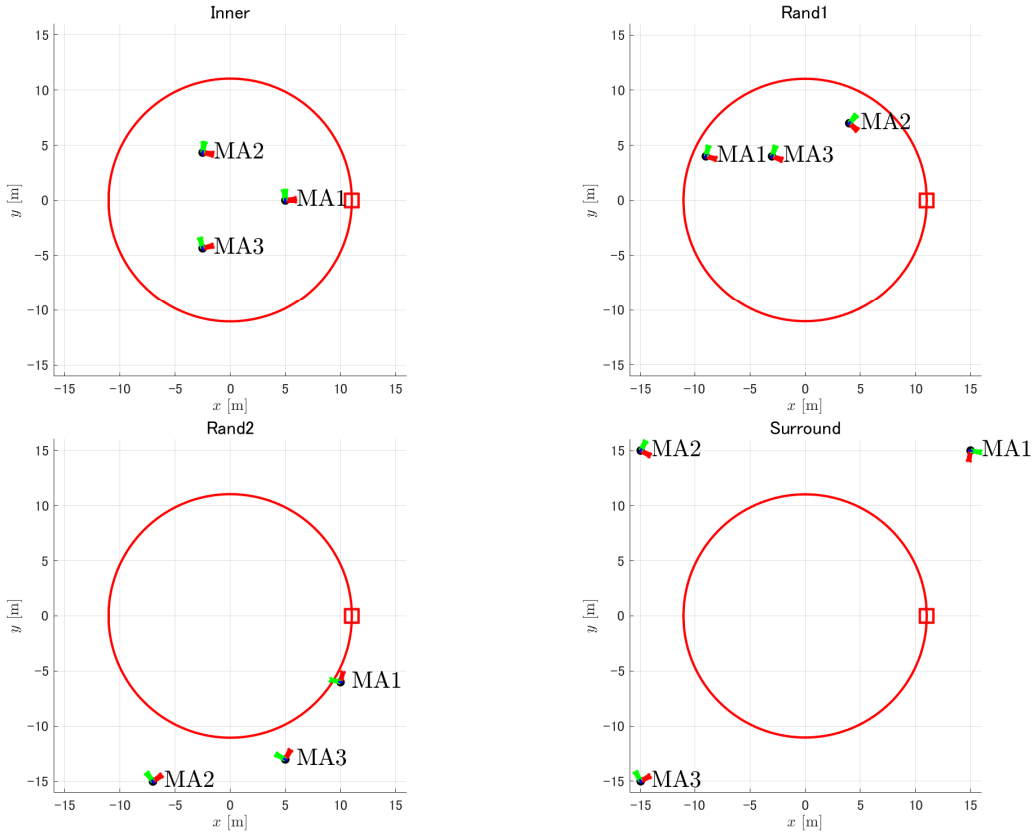

Figure 5. Microphone array placements to compare with. The red circle indicates the sound source trajectory. Each microphone array is indexed as MA1, MA2, and MA3. The red and green arrows starting from each microphone array indicate the longitudinal direction and lateral direction, respectively. The description of each placement is described in Section 4.1.1.

4.1.2. Results and Discussion

Based on the settings from the previous section, sound source tracking was performed using four different methods while changing the microphone array placement. Figures 6 and 7 show the microphone array placement and the trajectory calculated by the proposed method, and the sound source tracking results using the PAFIM are also illustrated together. From Figure 7, it can be seen that when the microphone array placement is optimized using the proposed method, where the three microphone arrays are placed to form an orthogonal coordinate system centered around the sound source. Through the simulation, the maximum cosine of the angles formed by the pairs of vectors extending from each microphone array to the sound source was 0.46 (approximately 63 degrees when converted to angle). The average distance between each microphone array and the sound source was 5.77 m, and the height of the microphone arrays converged to the lower limit $z_{\min}$. From these results, it can be considered that the proposed objective function attempts to shorten the distance between the microphone array and the sound source while maintaining orthogonality between the lines connecting the microphone array and the sound source.

From Figure 6, we can see that both the estimated trajectory and the microphone array trajectory trace an almost clean circle, mirroring the ground truth trajectory. This pattern is also maintained by other microphone arrays, as they depict a circular trajectory, conforming to the microphone array placement seen in Figure 7. This behavior is due to the smoothing factor λ_h, as indicated in Equation (11). When λ_h is set to 0, the drone carrying the microphone array needs to maintain the placement shown in Figure 7, revolving around the sound source. Depending on the optimization measure applied to Equation (11), this could lead to a potential non-smooth trajectory for the microphone arrays. However, one

must be careful not to set λ_h too high as this would limit the drone's range of movement. For this simulation, we have set λ_h to a low value of 0.0001 to balance smoothness and flexibility. Table 2 shows the RMSE (root-mean-square error) for each microphone array arrangement and each tracking method, which can be calculated using the following formula.

$$\mathrm{RMSE} = \sqrt{\sum_{k=1}^{K} \mathrm{error}_k^2} \tag{18}$$

Here, error_k is the Euclidean distance between the estimated sound source location and the true sound source location at time step k. In this simulation, the sound source tracking is performed every 0.5 s for a 46.2 s simulation, which means $K = 92$. From Table 2, it can be seen that the proposed placement planning method minimizes the error of triangulation points compared to other placements and also minimizes the tracking errors of the Yamada20 [11] and Yamada21 [12] tracking methods. This is thought to be because the proposed placement planning method has the effect of suppressing outliers of triangulation points by trying to orthogonalize the estimated directions. Conversely, the tracking error of the sound source is maximized in the inner arrangement, where the estimated sound source directions are always nearly parallel, highlighting the need for orthogonalization of the estimated sound source directions. With the minimization of the error of the average triangulation points, the RMSE of Yamada20 [11], which performs triangulation, is also minimized, and the RMSE of the tracking method Potamitis04 [29], which also uses triangulation, differed by only about 0.3 m compared to the arrangement that minimized the RMSE (surround). The tracking method Yamada21 [12] is not a triangulation-based tracking method, but a small tracking error was obtained using the proposed arrangement. Yamada21 [12] calculates the likelihood distribution of the sound source location by integrating the directional spectrum calculated when each microphone array estimates the direction, and this likelihood distribution tends to resemble the distribution where triangulation points appear. Therefore, the proposed arrangement designed to suppress the variance of triangulation is thought to have the effect of reducing the variance of the location likelihood distribution calculated by Yamada21 [12] and reducing the tracking error. On the other hand, it can be inferred that the proposed placement planning method did not lead to an improvement in the tracking performance of the tracking method Lauzon17 [31]. This is because Lauzon17 [31] is originally intended to be applied to a fixed group of microphone arrays and estimates the relative position of the sound source with respect to the microphone array group based on the estimated sound source direction. In other words, it was found that even when the proposed placement planning method is applied to tracking methods that assume that the microphone array does not move, its tracking performance is inferior to that of a fixed placement of the microphone arrays. In summary, through numerical simulations, the proposed arrangement method provided a placement that attempts to orthogonalize the estimated sound source directions of each microphone array while getting closer to the sound source. By sequentially realizing the optimal placement for a moving sound source, especially for tracking methods using triangulation, the tracking error was reduced. In addition, an improvement in tracking performance was observed for a tracking method that calculates a locational distribution which is similar to the distribution where triangulation points occur.

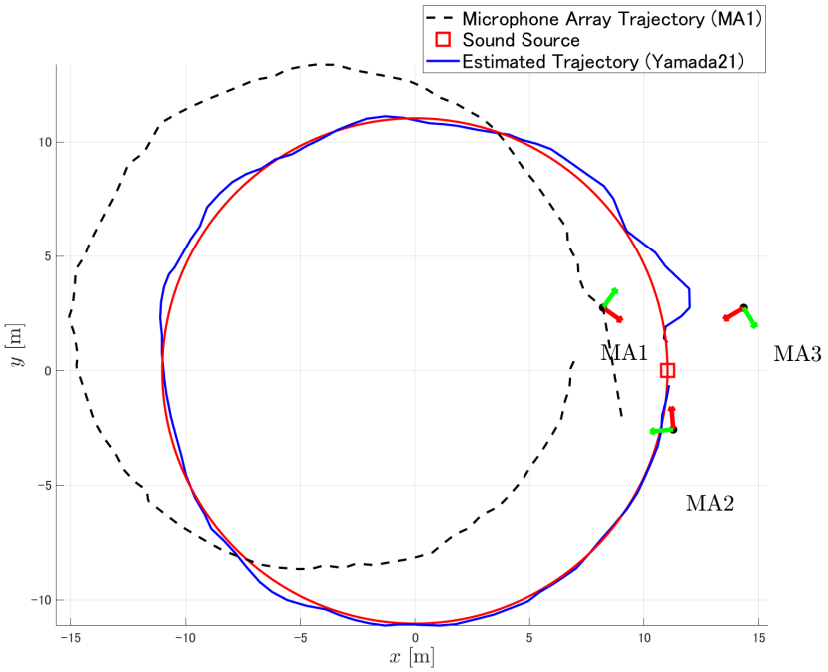

Figure 6. Top view of tracked sound source trajectory and microphone array trajectory of MA1. The red line is the ground truth of the sound source trajectory, the blue line is the estimated trajectory by PAFIM [12], and the black dashed line is the trajectory of MA1 that is derived from the proposed method. MA2 and MA3 are also moving in a trajectory similar to that of MA1.

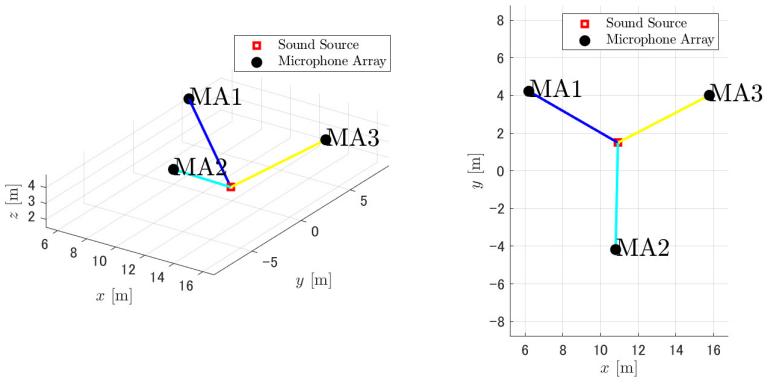

Figure 7. A drone placement at one moment in the simulation. (Time step $k = 2$). Since array-to-source direction is aimed to be orthogonal, the drones are placed in a formation that constructs a 3D Cartesian coordinate system with the sound source on the center. (**Left**: bird's-eye view, **Right**: top view).

Table 2. RMSE of tracking results for each tracking method and array placement. The leftmost column indicates the RMSE of the average of triangulation points. (The RMSE values presented in the table progressively increase in magnitude, as indicated by the color and font weight: they are largest for red bold letters, followed by red thin letters, black letters, and blue thin letters and are smallest for blue bold letters.)

	TripointAvr	**Potamitis04**	**Lauzon17**	**Yamada20**	**Yamada21**
Proposed	**0.37**	3.78	9.13	**0.55**	**1.64**
Inner	**7.08**	**4.52**	**38.05**	**9.29**	**20.22**
Rand1	1.52	3.76	13.14	4.67	4.88
Rand2	1.5	3.76	**3.04**	1.38	2.47
Surround	2.53	**3.46**	3.49	1.62	2.36

4.2. Multiple Source Tracking

In this simulation, we evaluate the strategy of allocating drones to sound sources through probability of observation.

4.2.1. Simulation Settings

In this simulation, two sound sources, Source 1 and Source 2, are placed, and 6 drones are used to track the position of each source (Figure 8). This section refers to the circular trajectory as "Source 1" and the rectangular trajectory as "Source 2" in Figure 8. Although Figure 8 is a 2D top view, the actual 3D location tracking is performed, and each sound source is located at a height of $z = 1.5$ m. The initial positions of each drone are placed at equal intervals on a 7 m radius arc centered at the point $(x, y, z) = (4, 0, 5)$ in the $z = 5$ plane, surrounding both sound sources. Each drone is equipped with a 16-channel spherical microphone array (Figure 4), recording at 24-bit, 16 kHz. The recording is performed for $T = 46$ s, and the sound sources are tracked for T seconds, making exactly one orbit around the red line in the figure. Source 1 and Source 2 are continuously emitting sine waves at 1000 Hz and 2000 Hz, respectively. In order to make the simulation environment similar to the actual outdoor environment, 16 channels of pre-recorded drone noise were added to each microphone array, and the SNR with the source signal was set to -35 dB. Each drone is moving following the update Equation (11), and the parameters required for the update were set to $\lambda_g = 0.01, \lambda_h = 0.0001, z_{\text{lim}} = 4.79$. The interior point method was used to compute the minimization of Equation (11). The MUSIC method [9] was applied to calculate the spatial spectrum in PAFIM, and it calculates from the signal between $\omega_L = 900$ Hz and $\omega_H = 2100$ Hz in 5-degree increments for both azimuth and elevation angle.

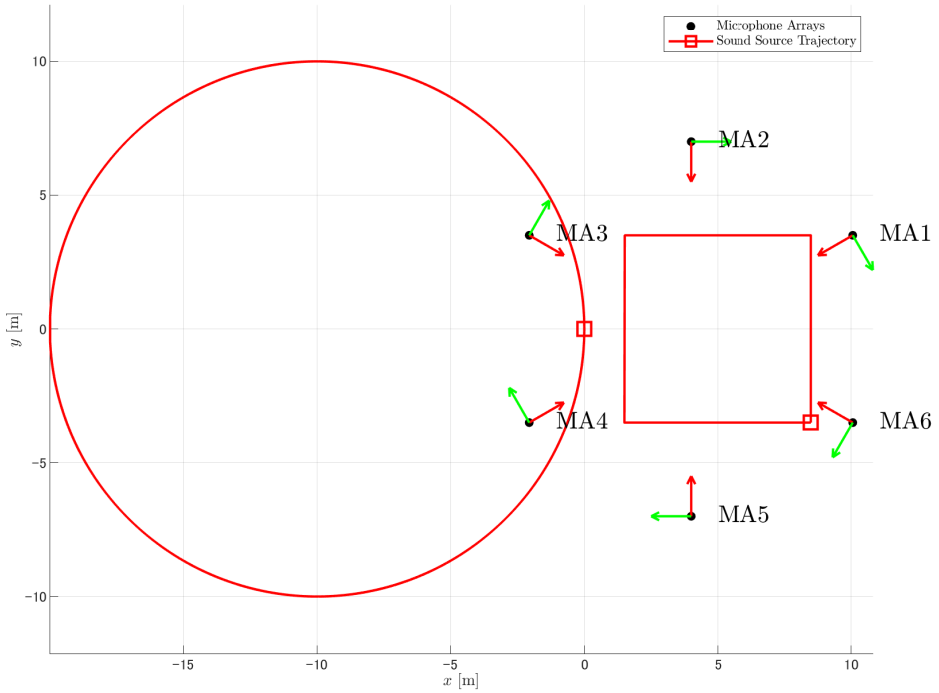

Figure 8. Top view of the simulation field for multiple source tracking. One sound source is emitting a 1000 Hz sine wave drawing a circular trajectory, and another is emitting a 2000 Hz sine wave drawing a square shaped trajectory. Each sound source starts from the red square markers and moves counter-clockwise. The red and green arrows indicates the initial position of the drones. (MA = microphone array).

4.2.2. Results and Discussion

Figure 9 shows the location tracking results and microphone array movements calculated during the 46 s of recording, indicating that the proposed method captures the approximate shape of the trajectory of both sound sources. In the middle of the simulation, MAs 2, 3, 4, and 5 surrounded Source 1 at equal intervals, and MAs 1, 4, and 6 moved to surround source 2 at equal intervals. This may be due to the fact that the estimated directions of each microphone array try to be closer to orthogonal to each other by optimization through the Equation (11). The height of each microphone array is always $z_{\text{lim}} = 4.79$ m, and the microphone arrays try to maintain a distance of about 10 m from the tracking sound source, which indicates that each microphone array is trying to get as close as possible to the target sound source while maintaining orthogonality between the estimated directions through optimization of the Equation (11). The RMSE (root-mean-square error) of Source 1 (circular orbit) was 2.15 m, and that of Source 2 (square orbit) was 0.65 m. The RMSE for Source 1 was relatively large due to the fact that the tracking results are pulled by the trajectory of Source 2, as seen in Figure 9f. In this section, when the filters that track the sound sources share the same location, no measures are taken to avoid tracking the same sound source, which may have caused the tracking results to drift to the trajectory of a nearby sound source. In fact, when the drone swarm was simulated with a different initial configuration, the two source tracking filters tracked the same source throughout the simulation, and there were cases where six drones surrounded only Source 2. Therefore, it is necessary to take measures to prevent the tracking filters from confusing the sound sources when the sound sources are close to each other. Figure 10 shows the transition of $p(\alpha_{i\to k})$ for each microphone array. For example, MA2 and 3, whose initial positions were closer to Source 1 and farther from Source 2, showed an increase in probability of observation of Source 1 and a decrease in probability of observation of Source 2. Therefore, MA 2 and 3 are concentrating on contributing to the source tracking of Source 1, and are moving around Source 1 from start to finish. The same kind of phenomenon also happened for MA 1 and 6 in the tracking of Source source 2. The initial positions of MA 4 and 5 were relatively close to both Sources 1 and 2, confirming that they were able to estimate the directions of both sources. Therefore, the probability of observation of MA 4 and 5 for both sources in the early stage of the simulation exceeded $p_{\text{thre}} = 0.3$, indicating that they were contributing to the tracking of both sources. However, Equation (11) let the microphone arrays to be positioned orthogonally to each other so that the microphone arrays move away from each other as much as possible. Therefore, MA 4 is always placed between sound Sources 1 and 2 to estimate the direction of both sources, while MA 5 is pulled toward the direction of Sound source 1 and eventually focuses only on the tracking of sound Source 1. The probability of observation of each microphone array increased when the source was audible (=close to source with high signal-to-noise ratio) and decreased when it was not, as designed. It can be seen that PAFIM tracked the source location only by the microphone arrays that have high probability of observation to the corresponding sound source.

Figure 11 shows the tracking result when the threshold p_{thre} is set to 0, which means that all microphone arrays are accounted for to track all sound sources whether each array can hear the sound source or not. Since all drones try to track both sound sources, the drones form a large shape that surrounds all sound sources. However, since this shape is too large, the source-to-array distance increases and the SNR decreases, which is why tracking of sound source 1 fails. These results shows us that the microphone arrays should be allocated only to sources that they can hear and greedy use of microphone arrays can make a formation that makes the SNR low.

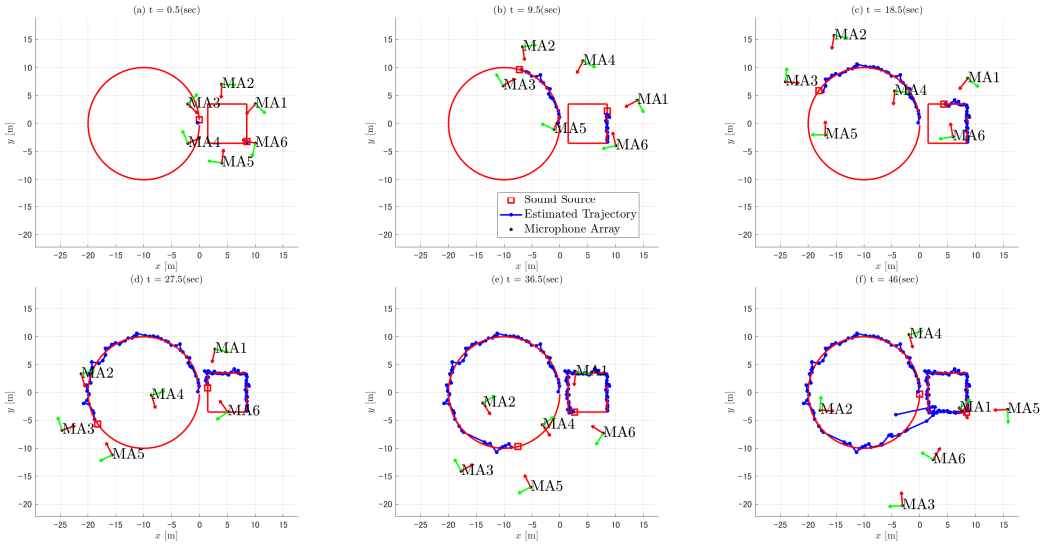

Figure 9. Tracking results through a 46.2 s simulation, each figure is the result of one moment in the simulation. $p_{\text{thre}} = 0.3$ in this simulation.

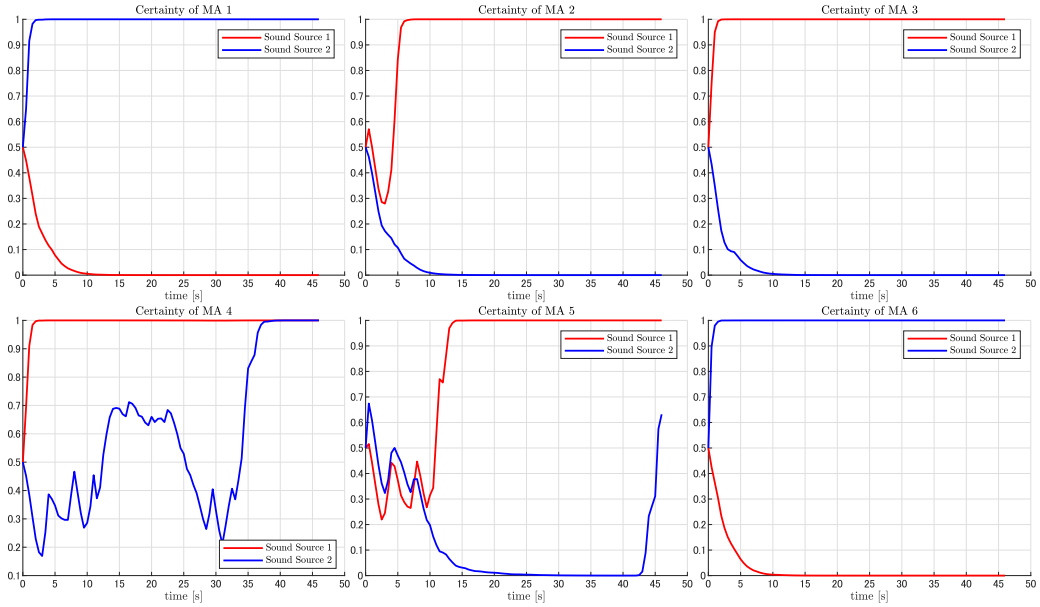

Figure 10. Transition of probability of observation for each microphone array. The red line is the probability of observation of sound source 1 (circular trajectory), and the blue line is the probability of observation of sound source 2 (square shape trajectory). Most microphone arrays focus and follow to one of the sources, but MA4 and MA5 tries to track both sound sources at the beginning of the simulation since both microphone arrays are between those sources.

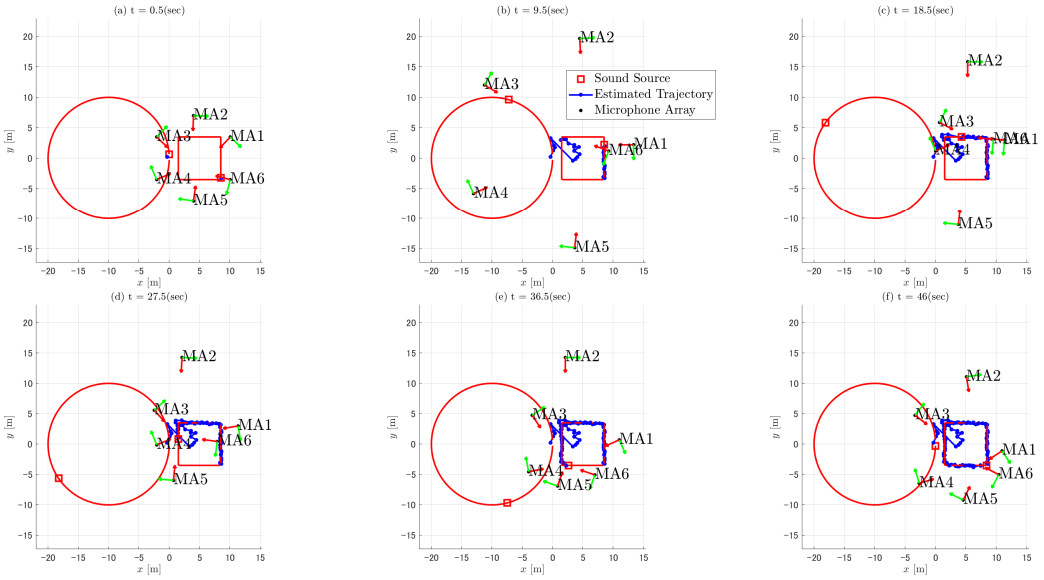

Figure 11. Tracking results through a 46.2 s simulation: each figure is the result of one moment in the simulation. $p_{\text{thre}} = 0.0$ in this simulation.

5. Conclusions

This paper has introduced active drone audition where microphone-array-equipped drones actively plan their placements to improve sound source tracking results. Generally, when multiple microphone arrays are used to estimate the location of a sound source, it is desirable that the direction of the sound source estimated by each array be orthogonal to each other, and at the same time, the microphone arrays should be close enough to the tracked sound source so that the sound source can be heard. To satisfy these two conditions, an algorithm is proposed to determine the drone and microphone array placement by optimizing an evaluation function (Equation (11)). In addition, the concept of probability of observation is introduced so that the estimation results are not affected by the microphone arrays that cannot hear multiple sources simultaneously and to perform the sound source tracking only using microphone arrays with high probability of observation. The proposed algorithm is evaluated by numerical simulation and shows that it can make small tracking error for triangulation based tracking methods. While the proposed algorithm was also able to track multiple sound sources in some cases, it was confirmed that the tracking results were confused when the sound sources were close to each other, which is our future work. In addition, this research has focused on numerical simulation which can understand the pure performance of the proposed method. However, unknown noise such as wind and self localization error of the drones will happen in the real world, and investigating the robustness of the proposed method in actual outdoor environments is also an important task for the future.

Author Contributions: Conceptualization, T.Y. and K.N. (Kazuhiro Nakadai); methodology, T.Y.; software, T.Y.; validation, T.Y.; formal analysis, T.Y.; investigation, T.Y.; resources, T.Y.; data curation, T.Y.; writing—original draft preparation, T.Y.; writing—review and editing, T.Y., K.I., K.N. (Kenji Nishida) and K.N. (Kazuhiro Nakadai); visualization, T.Y.; supervision, K.N. (Kazuhiro Nakadai); project administration, T.Y.; funding acquisition, K.I., K.N. (Kenji Nishida) and K.N. (Kazuhiro Nakadai). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by JSPS KAKENHI Grant No. JP19KK0260, JP20H00475, and JP23K11160.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

Abbreviations

The following abbreviations are used in this manuscript:

SNR	Signal-to-noise ratio
MUSIC	Multiple signal classification
PAFIM	Particle filtering with integrated MUSIC
RMSE	Root-mean-square error

References

1. Brown, G.J.; Cooke, M. Computational auditory scene analysis. *Comput. Speech Lang.* **1994**, *8*, 297–336. [CrossRef]
2. Okuno, H.G.; Nakadai, K. Robot audition: Its rise and perspectives. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, South Brisbane, QLD, Australia, 19–24 April 2015; pp. 5610–5614.
3. Hoshiba, K.; Washizaki, K.; Wakabayashi, M.; Ishiki, T.; Kumon, M.; Bando, Y.; Gabriel, D.; Nakadai, K.; Okuno, H. Design of UAV-embedded microphone array system for sound source localization in outdoor environments. *Sensors* **2017**, *17*, 2535. [CrossRef] [PubMed]
4. Nakadai, K.; Kumon, M.; Okuno, H.G.; Hoshiba, K.; Wakabayashi, M.; Washizaki, K.; Ishiki, T.; Gabriel, D.; Bando, Y.; Morito, T.; et al. Development of microphone-array-embedded UAV for search and rescue task. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, Vancouver, BC, Canada, 24–28 September 2017; pp. 5985–5990.
5. Hoshiba, K.; Sugiyama, O.; Nagamine, A.; Kojima, R.; Kumon, M.; Nakadai, K. Design and assessment of sound source localization system with a UAV-embedded microphone array. *J. Robot. Mechatronics* **2017**, *29*, 154–167. [CrossRef]
6. Sibanyoni, S.V.; Ramotsoela, D.T.; Silva, B.J.; Hancke, G.P. A 2-D acoustic source localization system for drones in search and rescue missions. *IEEE Sensors J.* **2018**, *19*, 332–341. [CrossRef]
7. Van Veen, B.D.; Buckley, K.M. Beamforming: A versatile approach to spatial filtering. *IEEE ASSP Mag.* **1988**, *5*, 4–24. [CrossRef]
8. Knapp, C.; Carter, G. The generalized correlation method for estimation of time delay. *IEEE Trans. Acoust. Speech Signal Process.* **1976**, *24*, 320–327. [CrossRef]
9. Schmidt, R. Multiple emitter location and signal parameter estimation. *IEEE Trans. Antennas Propag.* **1986**, *34*, 276–280. [CrossRef]
10. Wakabayashi, M.; Washizaka, K.; Hoshiba, K.; Nakadai, K.; Okuno, H.G.; Kumon, M. Design and Implementation of Real-Time Visualization of Sound Source Positions by Drone Audition. In Proceedings of the 2020 IEEE/SICE International Symposium on System Integration (SII), Honolulu, HI, USA, 12–15 January 2020; pp. 814–819.
11. Yamada, T.; Itoyama, K.; Nishida, K.; Nakadai, K. Sound Source Tracking by Drones with Microphone Arrays. In Proceedings of the 2020 IEEE/SICE International Symposium on System Integration, Honolulu, HI, USA, 12–15 January 2020.
12. Yamada, T.; Itoyama, K.; Nishida, K.; Nakadai, K. Sound Source Tracking Using Integrated Direction Likelihood for Drones with Microphone Arrays. In Proceedings of the 2020 IEEE/SICE International Symposium on System Integration (SII), Honolulu, HI, USA, 12–15 January 2020; pp. 394–399.
13. Yamada, T.; Itoyama, K.; Nishida, K.; Nakadai, K. Assessment of sound source tracking using multiple drones equipped with multiple microphone arrays. *Int. J. Environ. Res. Public Health* **2021**, *18*, 9039. [CrossRef] [PubMed]
14. Kagami, S.; Thompson, S.; Sasaki, Y.; Mizoguchi, H.; Enomoto, T. 2D sound source mapping from mobile robot using beamforming and particle filtering. In Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 19–24 April 2009; pp. 3689–3692.
15. Sasaki, Y.; Tanabe, R.; Takemura, H. Probabilistic 3D sound source mapping using moving microphone array. In Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, Republic of Korea, 9–14 October 2016; pp. 1293–1298. [CrossRef]
16. Misra, P.; Kumar, A.A.; Mohapatra, P.; Balamuralidhar, P. Droneears: Robust acoustic source localization with aerial drones. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, Australia, 21–25 May 2018, pp. 80–85.
17. Atmoko, H.; Tan, D.; Tian, G.; Fazenda, B. Accurate sound source localization in a reverberant environment using multiple acoustic sensors. *Meas. Sci. Technol.* **2008**, *19*, 024003. [CrossRef]
18. Ishi, C.T.; Even, J.; Hagita, N. Using multiple microphone arrays and reflections for 3D localization of sound sources. In Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, Japan, 3–7 November 2013; pp. 3937–3942.
19. Ishi, C.T.; Even, J.; Hagita, N. Speech activity detection and face orientation estimation using multiple microphone arrays and human position information. In Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 28 September–2 October 2015; pp. 5574–5579. [CrossRef]

20. Gabriel, D.; Kojima, R.; Hoshiba, K.; Itoyama, K.; Nishida, K.; Nakadai, K. Design and assessment of multiple-sound source localization using microphone arrays. In Proceedings of the 2019 IEEE/SICE International Symposium on System Integration (SII), IEEE, Paris, France, 14–16 January 2019; pp. 199–204.
21. Gabriel, D.; Kojima, R.; Hoshiba, K.; Itoyama, K.; Nishida, K.; Nakadai, K. 2D sound source position estimation using microphone arrays and its application to a VR-based bird song analysis system. *Adv. Robot.* **2019**, *33*, 403–414. [CrossRef]
22. Kaneko, S.; Gamper, H. Large-scale simulation of bird localization systems in forests with distributed microphone arrays. *JASA Express Lett.* **2022**, *2*, 101201. [CrossRef]
23. Washizaki, K.; Wakabayashi, M.; Kumon, M. Position estimation of sound source on ground by multirotor helicopter with microphone array. In Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, Republic of Korea, 9–14 October 2016; pp. 1980–1985.
24. Wakabayashi, M.; Okuno, H.G.; Kumon, M. Multiple sound source position estimation by drone audition based on data association between sound source localization and identification. *IEEE Robot. Autom. Lett.* **2020**, *5*, 782–789. [CrossRef]
25. Valin, J.M.; Michaud, F.; Rouat, J. Robust 3D Localization and Tracking of Sound Sources Using Beamforming and Particle Filtering. In Proceedings of the 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, Toulouse, France, 14–19 May 2006; Volume 4, p. IV. [CrossRef]
26. Portello, A.; Bustamante, G.; Danès, P.; Piat, J.; Manhes, J. Active localization of an intermittent sound source from a moving binaural sensor. In Proceedings of the Forum Acuticum, Krakow, Poland, 7–12 September 2014.
27. Evers, C.; Dorfan, Y.; Gannot, S.; Naylor, P.A. Source tracking using moving microphone arrays for robot audition. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 6145–6149. [CrossRef]
28. Brandstein, M.S.; Silverman, H.F. A practical methodology for speech source localization with microphone arrays. *Comput. Speech Lang.* **1997**, *11*, 91–126. [CrossRef]
29. Potamitis, I.; Chen, H.; Tremoulis, G. Tracking of multiple moving speakers with multiple microphone arrays. *IEEE Trans. Speech Audio Process.* **2004**, *12*, 520–529. [CrossRef]
30. Busset, J.; Perrodin, F.; Wellig, P.; Ott, B.; Heutschi, K.; Rühl, T.; Nussbaumer, T. Detection and tracking of drones using advanced acoustic cameras. In *Unmanned/Unattended Sensors and Sensor Networks XI; and Advanced Free-Space Optical Communication Techniques and Applications*; SPIE: Bellingham, WA, USA, 2015; Volume 9647, pp. 53–60.
31. Lauzon, J.; Grondin, F.; Létourneau, D.; Desbiens, A.L.; Michaud, F. Localization of RW-UAVs using particle filtering over distributed microphone arrays. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; pp. 2479–2484. [CrossRef]
32. Evers, C.; Habets, E.A.P.; Gannot, S.; Naylor, P.A. DoA Reliability for Distributed Acoustic Tracking. *IEEE Signal Process. Lett.* **2018**, *25*, 1320–1324. [CrossRef]
33. Gazor, S.; Grenier, Y. Criteria for positioning of sensors for a microphone array. *IEEE Trans. Speech Audio Process.* **1995**, *3*, 294–303. [CrossRef]
34. Rabinkin, D.V.; Renomeron, R.J.; French, J.C.; Flanagan, J.L. Optimum microphone placement for array sound capture. In *Advanced Signal Processing: Algorithms, Architectures, and Implementations VII*; SPIE: Bellingham, WA, USA, 1997; Volume 3162, pp. 227–239.
35. Akbarzadeh, V.; Gagne, C.; Parizeau, M.; Argany, M.; Mostafavi, M.A. Probabilistic sensing model for sensor placement optimization based on line-of-sight coverage. *IEEE Trans. Instrum. Meas.* **2012**, *62*, 293–303. [CrossRef]
36. Song, B.; Roy-Chowdhury, A.K. Robust tracking in a camera network: A multi-objective optimization framework. *IEEE J. Sel. Top. Signal Process.* **2008**, *2*, 582–596. [CrossRef]
37. Ma, N.; May, T.; Wierstorf, H.; Brown, G.J. A machine-hearing system exploiting head movements for binaural sound localisation in reverberant conditions. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 19–24 April 2015; pp. 2699–2703.
38. Tourbabin, V.; Barfuss, H.; Rafaely, B.; Kellermann, W. Enhanced robot audition by dynamic acoustic sensing in moving humanoids. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, South Brisbane, QLD, Australia, 19–24 April 2015; pp. 5625–5629.
39. Bustamante, G.; Danés, P.; Forgue, T.; Podlubne, A. Towards information-based feedback control for binaural active localization. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 6325–6329.
40. Schmidt, A.; Löllmann, H.W.; Kellermann, W. Acoustic self-awareness of autonomous systems in a world of sounds. *Proc. IEEE* **2020**, *108*, 1127–1149. [CrossRef]
41. Martinson, E.; Schultz, A. Discovery of sound sources by an autonomous mobile robot. *Auton. Robot.* **2009**, *27*, 221–237. [CrossRef]
42. 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, QLD, Australia, 19–24 April 2015; pp. 5630–5634.
43. 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.

44. Nguyen, Q.V.; Colas, F.; Vincent, E.; Charpillet, F. Motion planning for robot audition. *Auton. Robot.* **2019**, *43*, 2293–2317. [CrossRef]
45. Asano, F.; Goto, M.; Itou, K.; Asoh, H. Real-time sound source localization and separation system and its application to automatic speech recognition. In Proceedings of the Seventh European Conference on Speech Communication and Technology, Aalborg, Denmark, 3–7 September 2001.
46. Hioka, Y.; Kingan, M.; Schmid, G.; Stol, K.A. Speech enhancement using a microphone array mounted on an unmanned aerial vehicle. In Proceedings of the 2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC), Xi'an, China, 13–16 September 2016; pp. 1–5.
47. Nakadai, K.; Takahashi, T.; Okuno, H.G.; Nakajima, H.; Hasegawa, Y.; Tsujino, H. Design and Implementation of Robot Audition System'HARK'—Open Source Software for Listening to Three Simultaneous Speakers. *Adv. Robot.* **2010**, *24*, 739–761. [CrossRef]
48. Wang, L.; Sanchez-Matilla, R.; Cavallaro, A. Tracking a moving sound source from a multi-rotor drone. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 2511–2516.
49. Sekiguchi, K.; Bando, Y.; Itoyama, K.; Yoshii, K. Layout optimization of cooperative distributed microphone arrays based on estimation of source separation performance. *J. Robot. Mechatronics* **2017**, *29*, 83–93. [CrossRef]