

Article

A Performance Assessment on Rotor Noise-Informed Active Multidrone Sound Source Tracking Methods

Benjamin Yen ^{1,*}, Taiki Yamada ^{1,†}, Katsutoshi Itoyama ² 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; yamada@ra.sc.e.titech.ac.jp (T.Y.); nakadai@ra.sc.e.titech.ac.jp (K.N.)

² Honda Research Institute Japan Co., Ltd., 8-1 Honcho, Wako, Saitama 351-0188, Japan; itoyama@ra.sc.e.titech.ac.jp

* Correspondence: benjamin@ra.sc.e.titech.ac.jp

† Benjamin Yen is an International Research Fellow of the Japan Society for the Promotion of Science (JSPS) at the Nakadai Lab, Tokyo Institute of Technology.

‡ Taiki Yamada is now with TIER IV, Inc.

Abstract: This study evaluates and assesses the performance of recent developments in sound source tracking using microphone arrays from multiple drones. Stemming from a baseline study, which triangulates the spatial spectrum calculated from the Multiple Signal Classification (MUSIC) for each drone, otherwise known as Particle Filtering with MUSIC (PAFIM), recent studies extended the method by introducing methods to improve the method's effectiveness. This includes a method to optimise the placement of the drone while tracking the sound source and methods to reduce the influence of high levels of drone rotor noise in the audio recordings. This study evaluates each of the recently proposed methods under a detailed set of simulation settings that are more challenging and realistic than those from previous studies and progressively evaluates each component of the extensions. Results show that applying the rotor noise reduction method and array placement planning algorithm improves tracking accuracy significantly. However, under more realistic input conditions and representations of the problem setting, these methods struggle to achieve decent performance due to factors not considered in their respective studies. As such, based on the performance assessment results, this study summarises a list of recommendations to resolve these shortcomings, with the prospect of further developments or modifications to PAFIM for improved robustness against more realistic settings.

Keywords: drone audition; sound source tracking; rotor noise reduction



Citation: Yen, B.; Yamada, T.; Itoyama, K.; Nakadai, K. A Performance Assessment on Rotor Noise-Informed Active Multidrone Sound Source Tracking Methods. *Drones* **2024**, *8*, 266. <https://doi.org/10.3390/drones8060266>

Academic Editor: Petros S. Bithas

Received: 4 May 2024

Revised: 30 May 2024

Accepted: 11 June 2024

Published: 14 June 2024



Copyright: © 2024 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

Drones as robots have been widely used in many applications, including search and rescue and military operations [1]. For example, in search and rescue operations, victims are often stranded in areas that are challenging to search [2]. In contrast to conventional means, drones can take advantage of their small size and manoeuvrability to navigate through environments that are difficult to access or perform more challenging tasks, thereby improving the success rate of locating the victims. Recently, drones have been actively studied in the field of computational auditory scene analysis (CASA) [3,4], which is otherwise referred to as “drone audition” [5]. Here, drones equipped with auditory sensors help understand the acoustic scenes they are subjected to. This is particularly useful in search and rescue missions, as understanding the surrounding environment influences the efficiency of the search and rescue significantly, especially when the victim's visual presence is impaired due to factors such as unfavourable weather conditions, excessive vegetation coverage in mountainous terrain, and collapsed buildings [3,6].

Drone audition is often realised by attaching an array of microphones to a drone (known as a *microphone array* [7]). With audio from the sound source of interest arriving at each microphone at slightly different times (or phase difference), a microphone array can capture both the spectral and spatial information regarding the sound source. Sound source localisation (SSL), an essential aspect of CASA, primarily utilises spatial information to locate the whereabouts of the sound source [5]. To date, numerous studies have attempted to perform SSL using drones. Many of these studies are based on the MULTiple Signal Classification (MUSIC) algorithm [8], along with the inclusion of a noise correlation matrix (NCM) to reduce the influence of the drone's high levels of rotor noise contaminating the microphone recordings [9]. Common approaches include simply averaging segments of microphone input signals [10,11]. In contrast, more sophisticated approaches utilise multimodal information through various sensors, such as rotor speed, inertia measurement unit, etc., to build an NCM specifically for reducing rotor noise [12]. A real-life drone system was also developed in [13] demonstrating real-time SSL on a flying UAV. Other approaches include utilising generalised crosscorrelation phase transform (GCC-PHAT) as the baseline to perform SSL using drones. For example, the study from [14] utilised GCC-PHAT to perform SSL, along with statistics-based spatial likelihood functions regarding the target sound source and rotor noise, to facilitate the reduction of rotor noise and thus also perform source enhancement. On the other hand, the study from [15] employed Wiener filtering (WF) to directly reduce the influence of rotor noise at the microphone's input signal before applying GCC-PHAT. Recent studies also showed approaches using convolutional neural networks for source localisation, such as [16–18].

As such, studies regarding SSL for drone audition have shown increasing attention. Most studies to date are based on a single drone, single microphone array configuration, which is used to estimate the *direction* of the sound source. However, in search and rescue missions, it is important not only to identify the *direction* of the sound but also the actual *3D location* of the sound source, which is a limitation of most single drone (with single microphone array) approaches. Studies for SSL for 3D location estimation have been considered for robots, and various approaches have been proposed. For example, some studies continuously change the orientation of the robot (microphone array) to improve direction estimation results [19,20], or they navigate the robot around the sound source, thus obtaining location information of the sound source from multiple viewpoints to map the sound source location [21,22]. However, it should be noted that the robots in these approaches do not move in response to the estimated sound source location. More recently, some approaches utilise multiple robots, where the trajectory of a sound source can be obtained by calculating the triangulation points from the estimated directions of multiple microphone arrays (from each robot) prior to obtaining the sound source location through Kalman filtering [23–25].

Recently, studies on drone audition using multiple microphone arrays, either in the form of multiple arrays attached to a single drone [26] or multiple drones with a single array attached [27], have also been proposed, wherein most are based on triangulation of the direction estimations or other forms of SSL estimations to obtain the sound source's 3D location [27]. In this study, we focus on the approach using multiple drones. Notable studies include that in [27,28], which not only performed SSL of their 3D location, but in the case where the sound source was moving, they performed sound source *tracking* (SST). For example, the study in [27] triangulated direction estimations from each drone using MUSIC to form triangulation points, which were then processed through a Gaussian sum filter. This process ensures that only points highly likely to represent the target sound source are considered. The method is called Multiple Triangulation and Gaussian Sum Filter Tracking (MT-GSFT). This study was later extended, where instead of triangulating the estimated directions, the MUSIC spectrum from each drone was directly triangulated to form a direction likelihood distribution, where the location of the sound source could then be obtained. The method is known as Particle Filtering with Integrated MUSIC (PAFIM) [28]. This allows more information regarding the MUSIC spectrum to be considered, as much

can be lost when only considering the *peaks* of the MUSIC spectrum where the *direction* lies. Although the studies have successfully demonstrated their effectiveness in simulation or outdoor experiments [29], there is still room for improvement [28]. As expected, due to the high levels of rotor noise, sound source tracking performance degrades significantly when rotor noise dominates over the target source signal, which can easily occur when a) the target signal is too weak, or b) the target source is too far away from the drones. As such, placement of the drones relative to the target sound source is also an essential factor for consideration [28].

Considering these issues, the studies in [27,28] were further extended to consider the optimal placement of the drones/microphone arrays while performing SST [30]. Exploiting the fact that the drones are mobile and can be navigated, the study in [30] developed a drone placement algorithm to continuously optimise the placement of the microphone arrays by navigating the drones to maximise SST effectiveness (hereby referred to as *array placement planning*). Furthermore, the algorithm also enables the tracking of multiple sound sources by calculating the likelihood of the sound sources occurring based on the previous sound source location estimates of each sound source. In other words, the study in [30] is one of the first studies to perform *active drone audition* in which the drones become *active* in the SST process. We note that while autonomous navigation for the optimal placement of multiple drones is not yet a widely explored area of study for sound source tracking, other applications have found significant performance improvement, such as 3D visual map reconstruction from aerial images [31].

While optimising the placement of drones is a highly effective way to improve the SST performance, there is still a lack of direct treatment for the high levels of rotor noise itself. To address this issue, the study in [32] proposed an extension to MT-GSFT and PAFIM by introducing a WF designed using the drone's rotor noise power spectral density (PSD) [33,34]. This reduces the influence of rotor noise at each microphone channel in a manner similar to that from [15]. The WF approach was adopted due to its effectiveness not only in SSL but also in sound source enhancement for drone audition [35–37]. On the other hand, the study [32] proposed an NCM design by exploiting the fact that the drone's rotors and the microphone array are fixed in positions relative to each other. Therefore, using the steering vectors or impulse response measurements, nulls can be placed at the rotor locations to suppress the sound arriving from those directions. This is a well-proven approach in sound source enhancement, where the rotor noise's influence can be reduced using beamforming [33,38].

With several techniques proposed in recent years, particularly those from [30,32], this study aims to integrate these proposed techniques and evaluate their performance improvements in SST based on [27,28]. This includes SST not only for single but also multiple sound sources. In addition, it has also been mentioned in [32] that to date, the simulations conducted for multidrone SST that neglect the noise arriving from adjacent drones are irrelevant due to the elevation difference between the drones (typically ≥ 5 m above ground) and the target sound source (which is typically on the ground). However, this should not be ignored due to the loudness of the drones, particularly when they are grouped to track sound source(s). Therefore, this study also evaluates the degree of influence noise from adjacent drones on multidrone SST. Due to the range of techniques considered in this study, we only evaluate them using the PAFIM method.

Overall, this study contributes to the following:

- Evaluating the performance improvement brought by combining optimal microphone array placement and/or rotor noise reduction techniques to PAFIM;
- A more complete/realistic numerical simulation setting;
- Assessment of the evaluation results leading to design recommendations for further improving multidrone SST.

The rest of the paper is organised as follows. A brief description of the problem setting is introduced in Section 2, followed by an overview of PAFIM, the optimal drone placement algorithm, and the rotor noise reduction methods in Section 3. The simulation setup for

evaluation is shown in Section 4, with the results and discussion presented in Section 5. Finally, we conclude with some remarks in Section 6.

2. Setting

The problem assumes N microphone arrays $\text{MA}_1, \dots, \text{MA}_N$ consisting of M sensors, with each mounted to a drone. Each i th microphone array receives a number of mutually uncorrelated sound sources, including K target sources $S_k(\omega, f)$, *spatially coherent* noise generated by U rotors on the drone—with each u th rotor noise denoted as $N_u(\omega, f)$ —and ambient *spatially incoherent* noise, such as wind $\mathbf{v}(\omega, f)$, in an outdoor free field environment. For the multidrone scenario, we also consider the noise radiated by adjacent drones $N_{j \in N(j \neq i)}$ (i.e., all drones apart from the i th drone). Since the drones are, at most times, a distance away from each other, we assume the noise source of any neighbouring drone received by the i th microphone array to be a point source, and the noise itself propagates from the j th drone to the i th microphone array in accordance to the steering vector $\mathbf{a}_{i,j}(\omega, \phi_{i,j}, \theta_{i,j})$. ω and f denote the angular frequency and frame index, respectively. Here, we assume all drones carry the same number of rotors U . Using the $N \times M$ -channel noisy recordings, the system aims to localise and track the K target source signals [27]. Assuming overdetermined cases (i.e., $M \geq U + K + N - 1$ for each i th drone/microphone array), the short-time Fourier transform (STFT) of the i th microphone array's (MA_i) input signals are expressed in vector form as

$$\begin{aligned} \mathbf{x}_i(\omega, f) &:= [X_{i,1}(\omega, f), \dots, X_{i,M}(\omega, f)]^T \\ &= \sum_{k=1}^K \mathbf{a}_{i,S_k}(\omega, \phi_{i,S_k}, \theta_{i,S_k}) S_k(\omega, f) + \sum_{u=1}^U \mathbf{a}_{i,u}(\omega, \phi_{i,u}, \theta_{i,u}) N_u(\omega, f) \\ &\quad + \sum_{j \in N(j \neq i)} \mathbf{a}_{i,j}(\omega, \phi_{i,j}, \theta_{i,j}) N_j(\omega, f) + \mathbf{v}(\omega, f), \end{aligned} \quad (1)$$

where ^{T} denotes the transpose, and $X_{i,m}(\omega, f)$ is the STFT of the m th microphone input signal. ϕ_{i,S_k} and θ_{i,S_k} are the azimuth and elevation directions (i.e., in spherical coordinates) from the k th target sound source to the microphone array MA_i in its own body coordinates, respectively. Likewise, $\theta_{i,u}$ and $\phi_{i,u}$ indicate the azimuth and elevation directions from the u th rotor MA_i in its own body coordinates, respectively. $\mathbf{a}_i(\omega, \phi_i, \theta_i)$ and $\mathbf{v}(\omega, f)$ are the steering vector between the source located at directions θ_i, ϕ_i , and each microphone m in MA_i , as well as the incoherent noise vector observed by the microphone array.

In addition to the microphone input signals, the state of each microphone array MA_n is described as

$$\mathbf{m}_i(f) = [\mathbf{m}_{i,xyz}(f)^T, \mathbf{m}_{i,\phi\theta\psi}(f)^T]^T, \quad (2)$$

$$\mathbf{m}_{i,xyz}(f) = [x_i(f), y_i(f), z_i(f)]^T, \quad (3)$$

$$\mathbf{m}_{i,\phi\theta\psi}(f) = [\phi_i(f), \theta_i(f), \psi_i(f)]^T, \quad (4)$$

where $x_i(f), y_i(f)$, and $z_i(f)$ indicate the center of MA_n in the three-dimensional coordinates, and $\phi_i(f), \theta_i(f)$, and $\psi_i(f)$ indicate the yaw, pitch, and roll angles of MA_i . The location of the k th sound source is defined as

$$\mathbf{e}_k(f) = [x_k(f), y_k(f), z_k(f)]^T. \quad (5)$$

3. Sound Source Tracking Framework

In this section, we describe the SST method. Figure 1 shows a conceptual outline of the basic setup for SST, and Figure 2 shows a flowchart of the method, including the PAFIM baseline, as well as the components for array placement planning and rotor noise reduction. As shown in the figures, audio recordings from the drone's microphone array are first passed to the ground station, where most of the processing is performed. Upon

receiving the recordings, for every t th time step of duration T_t , sound source localisation is performed using the Multiple Signal Classification (MUSIC) algorithm [8] (Section 3.1), which produces a spatial spectrum containing information indicating the potential directions in which the target source is present. PAFIM uses the MUSIC spectrum to generate a likelihood distribution that is then used to infer the location of the sound source using particle filtering (Section 3.2). After obtaining the source location, a dedicated array placement planning algorithm calculates the optimal placement (including orientation) for each drone to be in. Finally, these coordinates are then relayed back to the drones for autonomous navigation. The following sections describe each process.

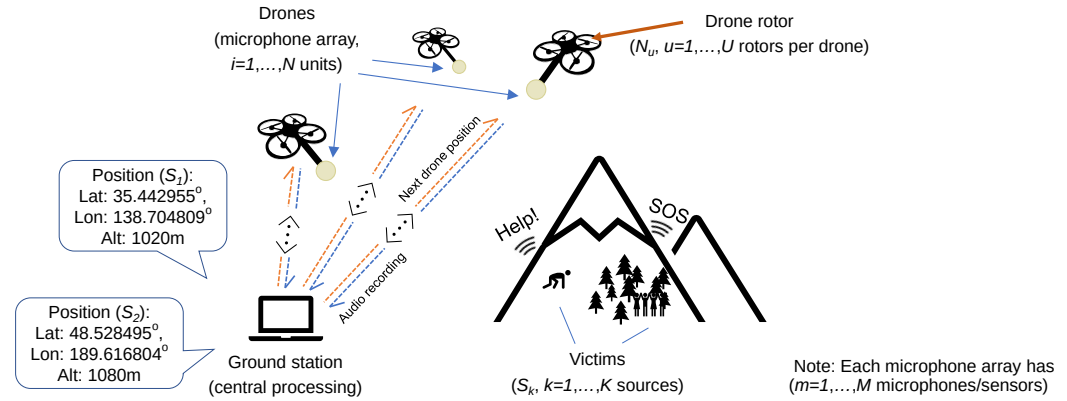


Figure 1. Overview of multidrone sound source tracking. Each drone has a microphone array attached.

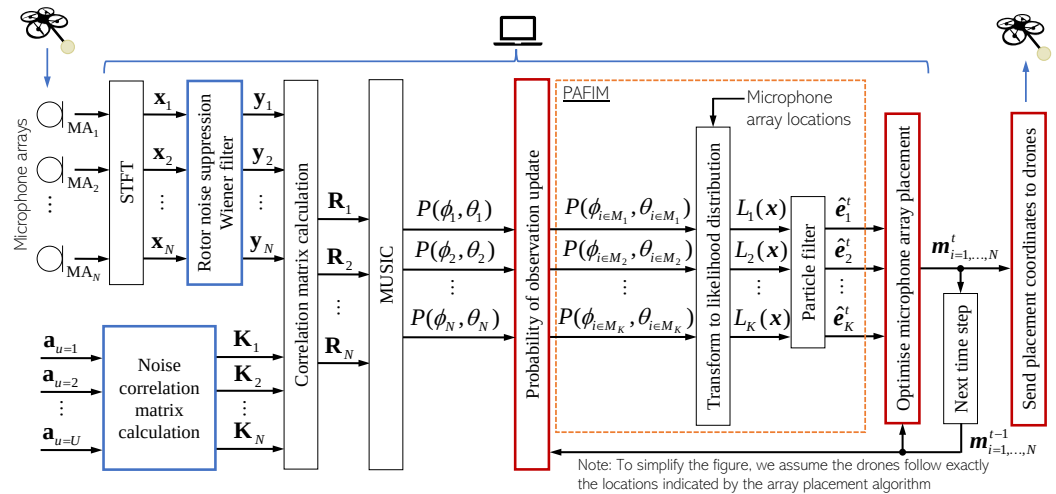


Figure 2. Flowchart describing the PAFIM method for each time step t . The array placement planning algorithm proposed in [30] (outlined in Section 3.3) is highlighted by the red boxes. Components of the rotor noise reduction method proposed in [32] (outlined in Section 3.4) are highlighted by the blue boxes.

3.1. Direction Estimation Using MUSIC

PAFIM, as the name suggests, uses MUSIC to obtain sound source localisation information for triangulation. MUSIC performs eigenvalue decomposition on the correlation matrix of the microphone array input signals $\mathbf{R}(\omega, t)$ as

$$\begin{aligned}\mathbf{R}(\omega, t) &= \frac{1}{T_R} \sum_{\tau=f}^{f+T_R-1} \mathbf{x}(\omega, \tau) \mathbf{x}^H(\omega, \tau) \\ &= \mathbf{E}(\omega, t) \mathbf{\Lambda}(\omega, t) \mathbf{E}^{-1}(\omega, t),\end{aligned}\quad (6)$$

where T_R is the number of frames to average to calculate $\mathbf{R}(\omega, t)$, and ^{H} denotes the Hermitian conjugate. $\mathbf{\Lambda}(\omega, t)$ and $\mathbf{E}(\omega, t)$ are the eigenvalue (diagonal) and eigenvector matrices, respectively. Using $\mathbf{E}$ calculated from (6), the spatial spectrum of MUSIC is calculated as

$$P(\phi, \theta, t) = \frac{1}{\omega_H - \omega_L + 1} \sum_{\omega=\omega_L}^{\omega_H} \frac{\|\mathbf{a}^H(\omega, \phi, \theta) \mathbf{a}(\omega, \phi, \theta)\|}{\|\mathbf{a}^H(\omega, \phi, \theta) \mathbf{E}_N(\omega, t)\|}, \quad (7)$$

where $\mathbf{E}_N$ is the matrix of eigenvectors corresponding to the noise subspace in $\mathbf{R}(\omega, t)$. ϕ and θ denote the source azimuth and elevation direction relative to the centre of the microphone array, respectively. Finally, ω_L and ω_H denote the minimum and maximum frequency bins. Usually, the sharp peaks in the MUSIC spectrum $P(\phi, \theta, t)$ correspond to the directions in which a target sound source is present. We denote these directions as $(\hat{\phi}_S(t), \hat{\theta}_S(t))$, where $\hat{\cdot}$ denotes an estimate. Consequently, at each time step t , the corresponding state of each microphone array $\mathbf{MA}_i$ $\mathbf{m}_i(t)$, $\mathbf{m}_{i,xyz}(t)$, $\mathbf{m}_{i,\phi\theta\psi}(t)$, and $\mathbf{e}(t)$ (see Section 2) will also be expressed in terms of the t th time step instead of the frame index f .

A common technique to improve robustness against noise is to apply an NCM to $\mathbf{R}(\omega, t)$ as a way to whiten the noise power [39]. In this study, we apply the NCM using the *Generalized Singular Value Decomposition MUSIC* (GSVD-MUSIC) technique [39], which applies the NCM $\mathbf{K}(\omega)$ and is followed by singular value decomposition:

$$\mathbf{K}^{-1}(\omega) \mathbf{R}(\omega, t) = \mathbf{E}_l(\omega, t) \mathbf{\Lambda}(\omega, t) \mathbf{E}_r^H(\omega, t), \quad (8)$$

where $\mathbf{E}_l(\omega, t)$ and $\mathbf{E}_r(\omega, t)$ are singular vector matrices, and $\mathbf{E}_l(\omega, t)$ is used to calculate the MUSIC spectrum in (7).

3.2. PAFIM

Following the calculation of the MUSIC spectrum, this section describes how MUSIC is used in PAFIM [28]. The method primarily consists of two processes: (i) calculating the likelihood distribution using the MUSIC spectrum, followed by (ii) particle filtering to calculate the location of the sound source at each time t .

Let $P_i(\phi, \theta)$ be the MUSIC spectrum calculated from $\mathbf{MA}_i$, let an arbitrary three-dimensional location be $\mathbf{x}$, and let the direction from $\mathbf{x}$ to $\mathbf{m}_i$ be (ϕ_i, θ_i) ; thus, the location likelihood L at a location $\mathbf{x}$ can be obtained by summing $P_i(\phi, \theta)$ across the microphone arrays as

$$L(\mathbf{x}) = \sum_i P_i(\tilde{\phi}_i^{\text{round}}, \tilde{\theta}_i^{\text{round}}), \quad (9)$$

$$\tilde{\phi}_i^{\text{round}} = \text{round}(\tilde{\phi}_i), \quad \tilde{\theta}_i^{\text{round}} = \text{round}(\tilde{\theta}_i), \quad (10)$$

$$\begin{bmatrix} \cos \tilde{\phi}_i \cos \tilde{\phi}_i \\ \sin \tilde{\phi}_i \cos \tilde{\phi}_i \\ \sin \tilde{\phi}_i \end{bmatrix} = \mathbf{R}_i^{-1} \begin{bmatrix} \cos \phi_i \cos \phi_i \\ \sin \phi_i \cos \phi_i \\ \sin \phi_i \end{bmatrix}, \quad (11)$$

where $\text{round}(\cdot)$ is a function that rounds the direction according to the resolution of the transfer function $\mathbf{a}(\omega, \phi_S, \theta_S)$, and $\mathbf{R}_i$ is the rotation matrix representing the posture of $\mathbf{MA}_i$, which can be defined by $\mathbf{m}_{i,\phi\theta\psi}$. In other words, the location likelihood $L(\mathbf{x})$ is the summation of direction likelihoods corresponding to directions towards $\mathbf{x}$ seen from each microphone array.

Following that, PAFIM tracks the sound source location by particle filtering using $L(\mathbf{x})$. Let I_P be the number of particles, and let $\mathbf{x}_t^{i_P}, w_t^{i_P}$ be, respectively, the state and weight of the i_P th particle at a time step t . The state $\mathbf{x}_t^{i_P}$ includes the location and velocity of particle i_P .

$$\mathbf{x}_t^{i_P} = [x_t^{i_P}, y_t^{i_P}, z_t^{i_P}, \dot{x}_t^{i_P}, \dot{y}_t^{i_P}, \dot{z}_t^{i_P}]^T. \quad (12)$$

Initial particles are sampled from the following distribution:

$$\mathbf{x}_0^{ip} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \quad (13)$$

$$\boldsymbol{\mu}_0 = [\boldsymbol{\mu}_{0,pos}, 0, 0, 0]^T, \quad (14)$$

$$\boldsymbol{\Sigma}_0 = \begin{bmatrix} \sigma_{pos}^2 \mathbf{I} & \mathbf{O} \\ \mathbf{O} & \sigma_{vel}^2 \mathbf{I} \end{bmatrix}, \quad (15)$$

where $\boldsymbol{\mu}_{0,pos}$ is the mean of the initial distribution, which is derived by triangulation based on direction estimation [40]. Following that, we take the average of all triangulation points to obtain $\boldsymbol{\mu}_{0,pos}$, and we employ an excitation-damping prediction model [40].

$$\mathbf{x}_t^{ip} = F\mathbf{x}_{t-1}^{ip} + H\mathbf{v}, \quad (16)$$

$$F = \begin{bmatrix} \mathbf{I} & T\mathbf{I} \\ \mathbf{O} & a\mathbf{I} \end{bmatrix}, \quad H = \begin{bmatrix} \mathbf{O} \\ b\mathbf{I} \end{bmatrix}, \quad (17)$$

$$\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \quad (18)$$

Here, $\mathbf{I} \in \mathbb{R}^{3 \times 3}$ is an identity matrix, and $\mathbf{O} \in \mathbb{R}^{3 \times 3}$ is a zero matrix. Variables a and b determine the ratio of velocity to carry on from the previous time step and excitation of particles, respectively. Each particle gain's weight is proportional to $L(\mathbf{x}_{t,pos}^{ip})$; hence, we have the following:

$$w_t^{ip} = w_{t-1}^{ip} \frac{L(\mathbf{x}_{t,pos}^{ip})}{\sum_i L(\mathbf{x}_{t,pos}^{ip})}, \quad (19)$$

where $\mathbf{x}_{t,pos}^{ip} = [x_t^{ip}, y_t^{ip}, z_t^{ip}]^T$. Note that if the effective particles are less than a threshold N_{thr} (i.e., $1 / \sum_{ip} (w_t^{ip})^2 \leq N_{thr}$), resampling would be necessary, with the weights of each particle reset to $1 / I_{PAFIM}$. Finally, the sound source location at time step t is estimated by taking the weighted average of the particles.

$$\mathbf{e}(t) = \sum_{ip=1} w_t^{ip} \mathbf{x}_t^{ip}. \quad (20)$$

3.3. Array Placement Planning

Previous studies showed that, when using multiple microphone arrays for source location tracking, the placement of the microphone arrays may affect the tracking accuracy [28,32]. This is reflected in both MT-GSFT and PAFIM in [32]. Fortunately, the drone's ability to manoeuvre freely enables the attached microphone arrays to be placed in a more desired formation. By first estimating the location of the sound source, we can calculate optimal positions for the drones to be placed at and improve the estimation accuracy of SST iteratively.

Generally, when the sound source is surrounded (or bounded) by the drones, better tracking accuracy is achieved compared to when the drones are outside the vicinity of the drones [30], as the drones are less likely to be orthogonal with each other relative to the target sound source. Of course, the closeness of the drone's rotors to the microphone array also influences the input signal-to-noise ratio (SNR) given their high noise levels. This is especially true when the sound source is far from the drone(s) [30]. To mitigate such problems, the study in [30] proposed a drone (microphone array) placement planning algorithm that actively navigates the drones to the optimal position and orientation relative to the target sound source to optimise the SST performance. The algorithm is designed to (a) only allocate drones (microphone arrays) to sound sources that they can hear, (b) maximise the orthogonality between pairs of drones for stable location estimation, (c) minimise the

distance between the target sound source and each drone to maximise the input SNR, and (d) constrain drones movements to minimise erratic movements [30].

To satisfy these policies, in each time step t , the drones are instructed to move to their updated positions as follows:

1. Initialise parameters (Section 3.3.3);
2. Calculate MUSIC spectrum [8] (Section 3.1);
3. Calculate probability of observation $p(\alpha_{i \rightarrow k} | \mathbf{z})$ (Section 3.3.1) and assign microphone arrays to sources to be tracked according to the probability of observation;
4. Perform sound source tracking for each sound source with the corresponding microphone array group (Section 3.2);
5. Calculate the next drone positions using estimated source locations. (Section 3.3.2).

The following sections will detail the procedures specific to calculating the next drone's positions.

3.3.1. Probability of Observation Update

As mentioned in Section 3.3, it is necessary to identify which sound source to observe for each drone (microphone array) to perform SST on multiple sound sources. This is because while MUSIC provides the potential directions of a *known* number of sound sources, it does not necessarily identify them. Therefore, a probability of observation variable is established to use the MUSIC spectrum and estimate whether the sound source is likely to occur, provided that we expect the sound source to be present there. Assume $\alpha_{i \rightarrow k}$ is a binary variable indicating the event that microphone array i has observed sound source k , and provided observation $\mathbf{z}$, the probability of observation $p(\alpha_{i \rightarrow k} | \mathbf{z})$ is defined as

$$p(\alpha_{i \rightarrow k} | \mathbf{z}) = \frac{p(\alpha_{i \rightarrow k})p(\mathbf{z} | \alpha_{i \rightarrow k})}{\sum p(\alpha_{i \rightarrow k})p(\mathbf{z} | \alpha_{i \rightarrow k})}. \quad (21)$$

The prior $p(\alpha_{i \rightarrow k})$ is the posterior in the previous time step, and likelihood $p(\mathbf{z} | \alpha_{i \rightarrow k})$ is then calculated as

$$p(\mathbf{z} | \alpha_{i \rightarrow k}) = \begin{cases} P_{\text{norm}}(\phi_{i \rightarrow k}, \theta_{i \rightarrow k}) & \text{for } \alpha_{i \rightarrow k} = 1 \\ \frac{1}{N_\phi N_\theta} & \text{for } \alpha_{i \rightarrow k} = 0 \end{cases}, \quad (22)$$

where $P_{\text{norm}}(\phi_{i \rightarrow k}, \theta_{i \rightarrow k})$ is the normalised MUSIC spectrum in the direction $(\phi_{i \rightarrow k}, \theta_{i \rightarrow k})$, in which $\phi_{i \rightarrow k}$ and $\theta_{i \rightarrow k}$ indicate the azimuth and elevation of source k from the perspective of microphone array i , respectively. N_ϕ and N_θ are the azimuth bin and elevation bin, respectively. Effectively, if $\alpha_{i \rightarrow k} = 1$ indicates that the MUSIC spectrum generated by microphone array i has a peak at the direction of the target source k , and thus $P_{\text{norm}}(\phi_{i \rightarrow k}, \theta_{i \rightarrow k})$ corresponds to the likelihood distribution for the k th target source. On the other hand, $\alpha_{i \rightarrow k} = 0$ indicates that the sound source k is not successfully detected by the i th array. Thus, the entire spatial spectrum has an equal probability, since there will be no distinct character from the MUSIC spectrum when there is no sound source. The probability of observation $p(\alpha_{i \rightarrow k} | \mathbf{z})$ is calculated for all N microphone arrays for each sound source, and when $p(\alpha_{i \rightarrow k} | \mathbf{z})$ is higher than a probability threshold p_{thre} , microphone array i will be added to a group $M_k \subseteq \{1, \dots, N\}$ that is specifically used to track the sound source k .

3.3.2. Drone Placement Optimisation

Following the probability of observation update and grouping drones to track their designated sound source, the optimised drone (microphone array) placement for the next time step $t + 1$ is calculated as

$$\arg \min_{m_{t+1}} f(m_{t+1}) + \lambda_g g(m_{t+1}) + \lambda_h h(m_{t+1}) \quad \text{s.t.} \quad z_t \geq z_{\text{lim}}, \quad (23)$$

$$f_{\perp}(m_{t+1}^{\text{loc}}) = \sum_{k=1}^K \sum_{\{i,j\} \in M_k (i \neq j)} \frac{\left(m_{i,t+1}^{\text{loc}} - \hat{e}_{k,t}\right)^T \left(m_{j,t+1}^{\text{loc}} - \hat{e}_{k,t}\right)}{\left\|m_{i,t+1}^{\text{loc}} - \hat{e}_{k,t}\right\|_2 \left\|m_{j,t+1}^{\text{loc}} - \hat{e}_{k,t}\right\|_2}, \quad (24)$$

$$g(m_{t+1}^{\text{loc}}) = \sum_{k=1}^S \sum_{i \in M_k} \left\|m_{i,t+1}^{\text{loc}} - \hat{e}_{k,t}\right\|_2, \quad (25)$$

$$h(m_{t+1}^{\text{loc}}) = \sum_{i=1}^N \sum_{i \in M_k} \left\|m_{i,t+1}^{\text{loc}} - m_{i,t}^{\text{loc}}\right\|_2, \quad (26)$$

where z_i is the z coordinate of drone i to be limited to a minimum height of z_{lim} , and $e_{k,t} \in \mathbb{R}^3$ is the 3D location of sound source k at time t . $m_{i,t} \in \mathbb{R}^6$ is the 6D state of the microphone array i at the time step t consisting of its 3D location $m_{i,t}^{\text{loc}} \in \mathbb{R}^3$ and 3D posture $m_{i,t}^{\text{loc}} \in \mathbb{R}^3$:

$$m_{i,t} = \left[\left(m_{i,t}^{\text{loc}}\right)^T, \left(m_{i,t}^{\text{pos}}\right)^T \right]^T. \quad (27)$$

Here, $f_{\perp}(m^{\text{loc}})$ is the sum of squares of cosines, where the angles are two array-to-source directions, thus meaning that $f_{\perp}(m^{\text{loc}})$ will be minimised when each pair of directions is orthogonal to one other. $g(m^{\text{loc}})$ is the sum of the distance from the array to the target source, and thus, minimising $g(m^{\text{loc}})$ will drive the drones to be close to the designated sound source(s). $h(m^{\text{loc}})$ is the sum of distances between $m_{i,t+1}^{\text{loc}}$ and $m_{i,t}^{\text{loc}}$ of the drones, which limits the amount of movement of the drone per time step when minimised. Since $f_{\perp}(m^{\text{loc}})$ is a normalised quantity, its value is at most limited to $S \cdot N C_2$. In contrast, $g(m^{\text{loc}})$ and $h(m^{\text{loc}})$ are simply a sum of 3D distances. Therefore, $f_{\perp}(m^{\text{loc}})$ is significantly smaller than $g(m^{\text{loc}})$ and $h(m^{\text{loc}})$, and coefficients λ_g and λ_h for $g(m^{\text{loc}})$ and $h(m^{\text{loc}})$, respectively, are required to balance the terms. These quantities are decided heuristically. We note that, as shown in Figure 3, the microphone array is mounted at a distance forward from the drone, which means that drone rotor noise primarily radiates behind the microphone array. As such, the posture of the drones is updated so that it always faces the estimated location of the source $\hat{e}_k$. In certain situations, microphone array i may sometimes clearly record audio coming from multiple sound sources (i.e., $p(\alpha_{i \rightarrow k}) \geq p_{\text{thre}}$ for more than one sound source). In such situations, its posture will be calculated as the average (or midpoint) between the directions to the corresponding sound sources.



Figure 3. DJI Inspire 2 drone used for this study.

3.3.3. Initialisation

At the first time step of the array placement planning algorithm, some parameters required, such as the number of sources and the initial values for the probability of observation $p(\alpha_{i \rightarrow k})$, are not known. To mitigate this, we assume that K is known and that the initial value of the probability of observation $p(\alpha_{i \rightarrow k})$ is 0.5 for all source–microphone array pairs.

3.4. Rotor Noise Reduction

For the drone audition scenario, assuming a single target sound source is present, we expect not only a sharp peak in $P(\phi, \theta, f)$ at the target source direction but also that of the rotors. Since the rotor locations are fixed relative to the microphone array, the directions of their corresponding peaks are also expected to be fixed and can be ignored. However, since the rotors are loud and near the microphone array, it generally results in very low input SNRs. In addition, due to the size of the propeller (i.e., the dominant contributor to rotor noise), we cannot simply consider rotor noise as point sources [34]. Therefore, under very low input SNRs, rotor noise can still dominate over the target sound source despite their sounds arriving towards the microphone array in vastly different directions. Therefore, this section describes approaches for reducing rotor noise influence prior to conducting MUSIC. Specifically, we introduce a Wiener filter (WF) that is specifically designed to minimise rotor noise from $\mathbf{x}(\omega, f)$ in Section 3.4.1, followed by an NCM designed based on prior knowledge regarding the relative locations of each rotor to the microphone array in Section 3.4.2.

3.4.1. Rotor Noise Reduction Wiener Filter

The study from [34] showed that a WF designed using accurate estimations of the rotor noise's PSD can greatly reduce the influence of rotor noise when applied to beamformer output signals. However, in this study, retaining the spatial information of the observed sound sources is desirable by preserving the relative time delays (i.e., phase information) of the observed input across the microphone array for sound source localisation. Therefore, we design a WF for each m th microphone input signal as follows

$$\begin{aligned} \mathbf{h}(\omega, f) &= [H_1(\omega, f), \dots, H_M(\omega, f)], \\ H_m(\omega, f) &= \frac{\phi_m(\omega, f) - \hat{\phi}_{m, \Sigma_u}(\omega, f)}{\phi_m(\omega, f)}, \end{aligned} \quad (28)$$

where $\phi_m(\omega, f)$ and $\hat{\phi}_{m, \Sigma_u}(\omega, f)$ are the rotor noise PSDs for the M microphone inputs. The output signal $\mathbf{y}(\omega, f)$ is then calculated as

$$\mathbf{y}(\omega, f) = \mathbf{h}(\omega, f) \odot \mathbf{x}(\omega, f), \quad (29)$$

where $\odot$ denotes the Hadamard product. Finally, we use $\mathbf{y}(\omega, f)$ to obtain the rotor noise reduced correlation matrix for the MUSIC algorithm to perform source localisation as

$$\mathbf{R}'(\omega, f) = \frac{1}{T_R} \sum_{\tau=f}^{f+T_R-1} \mathbf{y}(\omega, \tau) \mathbf{y}^H(\omega, \tau). \quad (30)$$

Typically, the rotor noise PSD is an estimated quantity. However, as an evaluation for the application of WF itself, we directly use the rotor noise recordings to calculate the rotor noise PSDs, along with temporal smoothing across f and adding white Gaussian noise to make the PSDs more representative of an estimated quantity.

3.4.2. Noise Correlation Matrix

As mentioned in Section 3.4, in a drone audition setting, since the microphone array is fixed to the drone itself (see Figure 3), the relative positions between the rotors and the microphones are fixed and thus known. Therefore, apart from the WF, we can take advantage of this property and apply an NCM $\mathbf{K}_{\text{rot}}(\omega)$ from the steering vectors that point towards each of the U rotors $\mathbf{a}_u(\omega)$ to spatially whiten their influence in the correlation matrix $\mathbf{R}(\omega, f)$ as

$$\mathbf{K}_{\text{rot}}(\omega) = \sum_{u=1}^U \mathbf{a}_u(\omega, \phi_u, \theta_u) \mathbf{a}_u^H(\omega, \phi_u, \theta_u). \quad (31)$$

Upon obtaining $\mathbf{K}_{\text{rot}}(\omega)$, it can then be applied in a GSVD-MUSIC approach (see Section 3.1 and (8)) to spatially whiten the directions corresponding to the rotors. By doing so, we expect the singular value decomposition process to more effectively cast the residual rotor noise contained in the input correlation matrix (see Section 3.1) into the noise subspace, thus revealing the target source more clearly in the MUSIC spectrum.

4. Numerical Simulation

In this section, we evaluate the SST performance of the PAFIM algorithms. First, we outline the simulation details utilised throughout all simulations in Section 4.1. Following that, simulation details for both single- and multisource scenarios (i.e., two sound sources) are described in Sections 4.2 and 4.3, respectively.

4.1. General Simulation Details

This section outlines the general details used throughout the simulation. As mentioned in Section 4, we evaluated PAFIM's SST performance as a baseline reference. We also evaluated PAFIM with array placement planning (described in Section 3.3.2) applied, which we denote as *active PAFIM* (aPAFIM). Furthermore, we evaluated PAFIM with rotor noise reduction techniques (described in Section 3.4) applied, which we denote as *rotor noise-informed PAFIM* (rPAFIM). Finally, we evaluated PAFIM with both array placement planning and rotor noise reduction applied, which is denoted as *rotor noise-informed active PAFIM* (raPAFIM). Details about the audio recording and simulation parameters for MUSIC are shown in Table 1, while simulation parameters for the PAFIM algorithm are shown in Table 2.

Table 1. Audio recording parameters and simulation parameters for MUSIC.

Variable	Value
Sampling rate	16 kHz
Audio resolution	16 bit
# of channels (for each drone)	16
STFT length (overlap shift)	512 (160)
ω_L	200 Hz
ω_H	4000 Hz

Table 2. Simulation parameters for PAFIM.

Variable	Value
a	0.9
b	0.3
I_p	500
Initial σ_{pos}^2	2.5
Initial σ_{vel}^2	2.5

As mentioned in Section 4, this study evaluated SST performance for single- and two-target source scenarios. The target source was 40 s of male speech randomly selected from the *Centre for Speech Technology Research VKTS (CSTR VKTS)* corpus [41], with their locations simulated at each $T_k = 0.5$ s time step. Drone noise was recorded directly from a DJI Inspire 2 drone (see Figure 3) in flight, with a 16-channel microphone array attached (see Figure 4 for details). Details regarding the audio recording parameters are shown in Table 1. We evaluated the input SNR conditions between -30 dB and 20 dB in 10 dB increments for both scenarios. This is to cover both scenarios that reflect what is expected in a realistic flight [34], but they are also ideal recording scenarios to evaluate the validity of the methods themselves. For all scenarios, the flight time was 40 s, thus resulting in 80 time steps. Realistically, drone scenarios have an input SNR of ≤ 0 dB. However, we also included input SNRs up to 20 dB for feasibility evaluation of the methods.

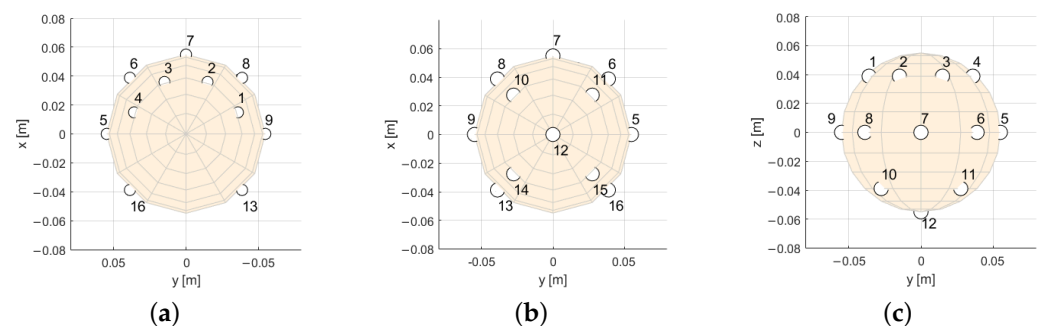


Figure 4. Geometrical details of the microphone array used in this study. Microphone positions are shown in (a) top view, (b) bottom view, and (c) front view.

We note that for the majority of the simulations, we omitted the rotor noise emitted from the neighbouring drones (see Section 2). This was done not only to simulate the input scenarios similar to previous studies but also to assess sound source tracking performance more progressively. This way, the input conditions where SST performance falls short for each technique applied towards PAFIM could be isolated. However, a performance evaluation, which includes the noise from the neighbouring drones, is included in the single source scenario as a separate assessment (see Section 4.2).

4.2. Simulation Detail: Single Sound Source

This section describes the simulation details for single-source scenarios. Specifically, we considered the scenario shown in Figure 5, which consists of a single target sound source moving at constant velocity in a circular trajectory 10 m in radius on the x - y plane. The sound source was tracked by four DJI Inspire 2 drones (i.e., $N = 4$) flying at a fixed height of 5 m. In practice, before the sound source's rough location is known, the sound source is often outside the drone's coverage. Therefore, the initial location of the sound source is placed a distance from the drone swarm. Details regarding the exact initial locations of the sound sources and the drones (microphone arrays) are shown in Figure 5.

In more realistic scenarios, each drone will receive not only the noise emitted by its own rotors but also the noise emitted by the neighbouring drones. Therefore, to evaluate how PAFIM and the technique applied (described in Sections 3.3 and 3.4) fare against these additional noise sources, the single source scenario also considers a simulation setting that includes neighbouring drone noise. Here, we assume that each neighbouring drone is distant enough to assume that their rotor noise is collectively a far-field point source.

Finally, the optimal array placement algorithm's scaling coefficients for the optimisation cost function were $\lambda_g = 0.2$ and $\lambda_h = 0.1$.

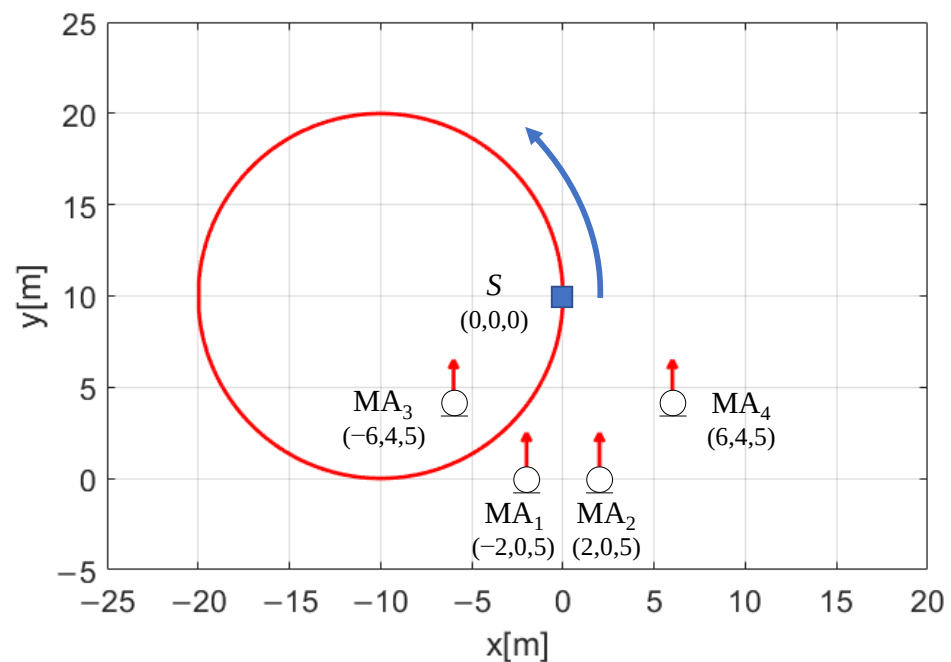


Figure 5. Simulation setup for single sound source scenarios. The x and y labels are the locations on the x axis and y axis in world coordinates. The z axis location for the sound source and microphone arrays MA₁ to MA₄ are 0 m and 5 m, respectively. The red line indicates the true trajectory of the sound source. The blue arrow indicates the direction in which the sound source moves, while the red arrows indicate the initial orientation of the microphone arrays. The blue square indicates the initial location of the sound source.

4.3. Simulation Detail: Two Sound Source

This section describes details regarding numerical simulation for the two source scenarios. Namely, we considered two scenarios shown in Figure 6, which consist of two target sound sources moving at a constant velocity. For scenario (a), one source moves in a circular trajectory 10 m in radius, while the other is a square trajectory with an edge of 7 m, with both on the x–y plane. Given that, in practice, it is very common to encounter situations where the two sources are far from each other, evaluating SST performance under such conditions is also of interest. As such, for scenario (b), we considered two sources placed further apart than scenario (a). Similar to scenario (a), we considered one sound source moving in a circular trajectory 4 m in radius on the x–y plane, while the other is a square trajectory with an edge of 4 m, again both on the x–y plane. Since the two source scenarios mean that some drones will concentrate their tracking on one of the two sound sources, fewer drones are tracking each sound source compared to the single source scenario. As such, here, the sound sources were tracked by six DJI Inspire 2 drones (i.e., $N = 6$) flying at a fixed height of 5 m. Similar to the single source scenario, we considered an initial sound source location placed a distance away from the drone swarm. Details regarding the exact initial locations of the sound sources and the drones (microphone arrays) for scenario (a) and scenario (b) are shown in Figure 6a and Figure 6b, respectively.

Finally, the optimal array placement algorithm's scaling coefficients for the optimisation cost function were $\lambda_g = 0.24$ and $\lambda_h = 0.01$.

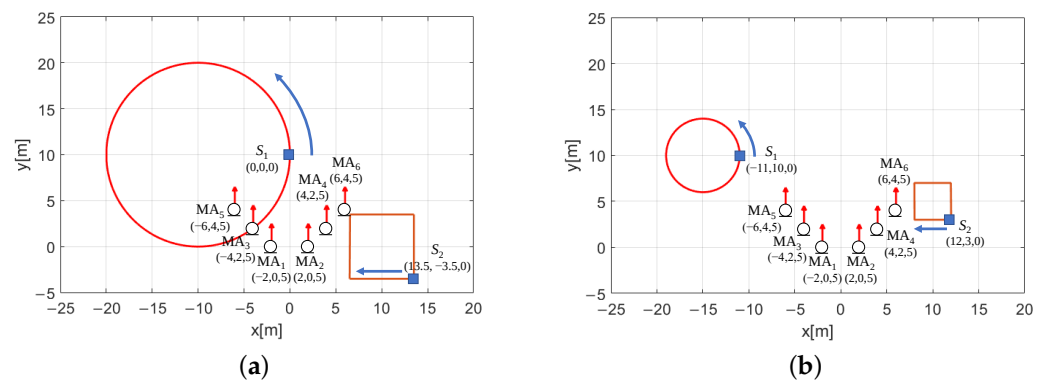


Figure 6. Simulation setup for two sound source scenarios, where (a) the sound sources are relatively close to each other at some point and (b) the sound sources are a distance away from each other. The x and y labels are the locations on the x axis and y axis in world coordinates. The z axis location for the sound source and microphone arrays MA₁ to MA₆ are 0 m and 5 m, respectively. The red and orange lines indicate the true trajectories of the sound sources S₁ and S₂, respectively. The blue arrows indicate the direction in which the sound source moves, while the red arrows indicate the initial orientation of the microphone arrays. The blue square indicates the initial location of the sound sources.

5. Results and Discussion

5.1. Single Sound Source Scenarios

Tables 3 and 4 show the evaluation results in terms of the Euclidean distance between the true and estimated locations for the single source scenario without and with neighbouring drone noise accounted for, respectively.

Table 3. Mean Euclidean distance between the estimated and true source trajectory for the single source scenario (in meters), **without accounting for the noise arriving from adjacent drones**. The lowest distance measures are highlighted in bold.

Method	Input SNR (dB)					
	−30	−20	−10	0	10	20
PAFIM (Baseline)	36.35	31.72	18.44	31.34	17.78	14.39
PAFIM w/WF [32]	8.69	7.17	8.01	8.48	7.42	6.53
PAFIM w/NCM [32]	27.15	20.43	18.76	33.19	28.03	22.11
rPAFIM [32]	15.27	12.99	14.62	6.00	14.38	8.70
aPAFIM [30]	41.89	36.43	41.11	28.67	14.81	2.19
aPAFIM w/WF [30,32]	2.96	1.98	2.14	3.03	1.98	1.72
aPAFIM w/NCM [30,32]	28.32	27.79	25.07	32.28	25.40	32.66
raPAFIM [30,32]	8.99	6.31	6.22	4.54	4.24	3.19

We first considered the case without neighbouring drone noise (i.e., Table 3), with Figure 7 showing an example of the estimated trajectories at an input SNR of −10 dB. Similar to the results shown in the previous study [32], the PAFIM baseline showed a dramatic improvement at input SNRs above −10 dB (despite the sudden drop in performance at 0 dB). However, with errors between the ground truth and the estimation reaching beyond 15 m for most input SNRs, it is clear that the performance of PAFIM alone was insufficient. This is expected, as most of the target sound source's travel is outside the vicinity of the drones and sometimes relatively far away from them, which can easily create and accumulate SST estimation errors over time. Coupled with the fact that PAFIM relies on its previous sound source location estimate as part of the location estimation for the next time step, if the general trajectory deviates, the modelling of the statistical distributions for particle filtering becomes increasingly skewed towards the erroneous trajectory, thus

leading to poorer subsequent location estimates [32]. As such, the influence of rotor noise on the SST performance was highly apparent.

Table 4. Mean Euclidean distance between the estimated and true source trajectory for the single source scenario (in meters), **with drone noise from adjacent drones accounted for**. The lowest distance measures are highlighted in bold.

Method	Input SNR (dB)					
	−30	−20	−10	0	10	20
PAFIM (Baseline)	35.04	24.16	28.21	35.19	29.91	13.96
PAFIM w/WF [32]	20.58	23.58	23.73	15.37	15.07	7.80
PAFIM w/NCM [32]	21.26	21.79	24.92	33.36	25.59	16.74
rPAFIM [32]	26.52	25.04	22.42	27.48	23.81	13.32
aPAFIM [30]	41.67	40.34	22.50	32.68	8.11	1.66
aPAFIM w/WF [30,32]	25.21	32.85	43.87	26.04	2.24	1.27
aPAFIM w/NCM [30,32]	36.88	25.73	28.22	31.57	31.62	27.64
raPAFIM [30,32]	21.48	18.79	46.89	20.27	4.53	2.19

Applying array placement planning to the PAFIM algorithm (i.e., aPAFIM) showed a significant change in performance. Under lower input SNRs, the aPAFIM method is outperformed by the PAFIM baseline (as shown by the higher errors in Table 3). However, when the input conditions start to suit aPAFIM (i.e., input SNR above −10 dB), a dramatic performance improvement, similar to that shown in the previous study [30], is shown. This is expected since with array placement planning under lower input SNRs, the drones follow the estimated trajectory of the target sound source. Therefore, when the estimation is erroneous, the subsequent placement of the drones will also be erroneous, causing the subsequent sound source location estimation to be increasingly less accurate. Nevertheless, at higher input SNRs (e.g., 20 dB), aPAFIM outperformed most compared methods (see Figure 7). However, the input SNR, where aPAFIM showed decent performance, was much higher than in [30]. The difference in performance shown here to that from [30] was most likely caused by a combination of two reasons: (i) in [30], sinusoidal tones of 1000–2000 Hz range were used as the target sound, which is continuous and has less spectral overlap with major components of drone noise compared to speech, and (ii) the initial placement of the drones were already relatively optimal compared to that of this study, which did not impose this assumption. However, despite the higher input SNR required for aPAFIM, the results suggest that aPAFIM can improve SST performance. However, to leverage the performance improvement, rotor noise reduction is required.

In contrast to the array placement planning algorithm, the rotor noise reduction methods generally showed consistent improvement in SST performance across the input SNRs. Overall, the WF showed significant improvement over the PAFIM baseline and was relatively consistent across most input conditions, similar to that of the previous study [32]. Furthermore, this performance improvement was apparent in PAFIM with or without the array placement planning algorithm. Therefore, it is unsurprising that PAFIM delivered the best tracking performance with WF and array placement planning. On the other hand, while the NCM showed some improvement over the PAFIM baseline, the error was still significant such that the results would not resemble the actual trajectory of the ground truth (see Figure 8, where PAFIM with WF showed tracking that resemble the ground truth. At the same time, PAFIM with NCM failed to follow any parts of the true trajectory). In fact, at input SNRs above 10 dB, the PAFIM baseline outperformed PAFIM with NCM. This also explains why rPAFIM had larger error figures than PAFIM with only WF applied, which is a result that is slightly different to that from [32]. However, overall, for the single sound source case (without taking into account adjacent drone noises), aPAFIM with WF delivered the highest level of improvement in tracking performance, and along with PAFIM with WF and raPAFIM, are the only cases that showed strong performance even under realistic (i.e., low input SNR) conditions. This indicates that both array placement planning and rotor

noise reduction techniques complement each other well and can significantly improve SST performance and each other's shortcomings.

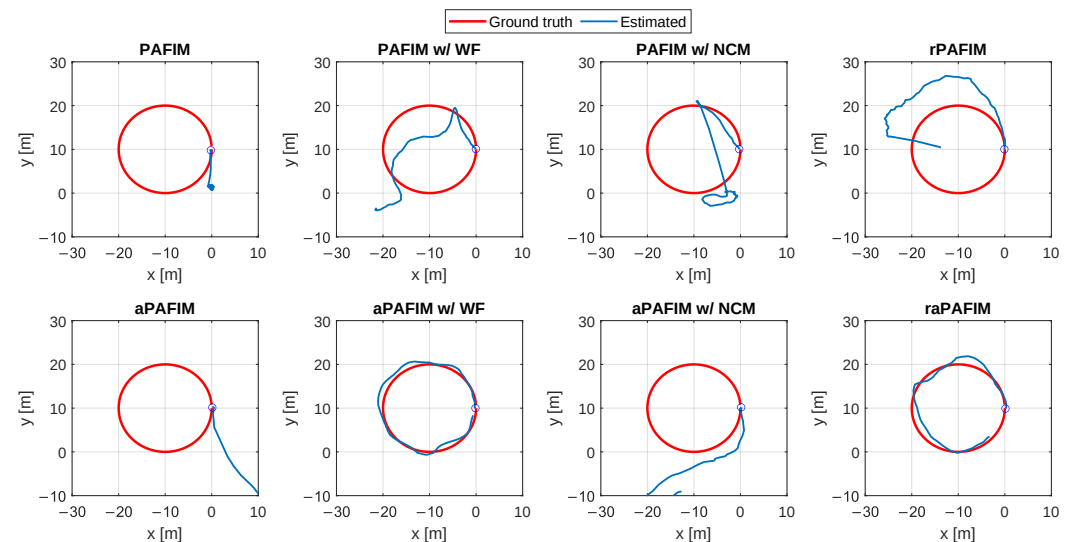


Figure 7. Example sound source trajectory at input SNR of -10 dB in 2D world coordinates (x - y plane) for the single sound source scenario, *without accounting for noise arriving from adjacent drones*. The blue circle indicates the initial location estimate.

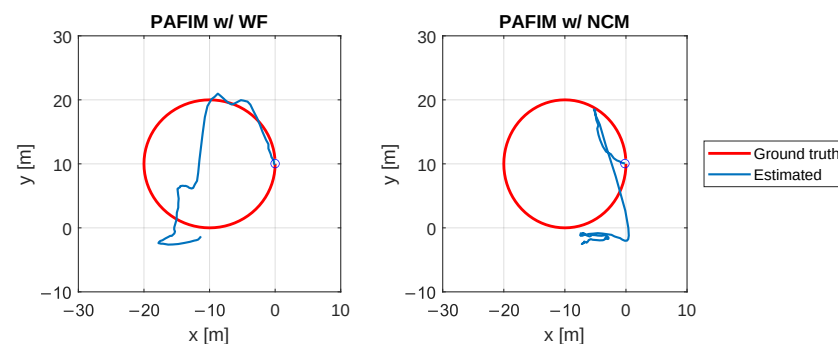


Figure 8. Example sound source trajectory at input SNR of -20 dB in 2D world coordinates (x - y plane) for the PAFIM with WF and PAFIM with NCM methods. The blue circle indicates the initial location estimate.

However, as mentioned in Section 1, due to the high levels of rotor noise in a multi-drone setting, the influence of their noise towards one another should not be overlooked. As such, we also compared the performance with these noises considered. Table 4 shows the evaluation results in terms of the Euclidean distance between the true and estimated locations for the single source scenario (in meters), with adjacent drone noise taken into account. As shown, the neighbouring drone noise significantly influenced the overall SST results, with most methods below an input SNR of 0 dB showing errors beyond 15 m. Again, the PAFIM method with WF delivered better SST improvement. Similar to single source scenario results, above an input SNR of 0 to 10 dB, the SST performance showed a significant increase. This was the case for PAFIM with or without rotor noise reduction, as well as array placement planning. However, above the 0 to 10 dB input SNR boundary, methods with array placement planning significantly outperformed methods without it.

On the contrary, for input SNRs below this boundary, it is apparent that the array placement planning algorithm was heavily impacted and could not function as desired, thus leading to higher errors than methods without array placement planning (see Table 4). This was particularly the case for methods with array placement planning, where, due to the influence of rotor noise, the drones seemed to draw into each other under lower-input

SNR cases. With the inclusion of rotor noise from other drones, the noise from adjacent drones will sometimes be accidentally recognised as the sound source, of which the array placement planning algorithm will instruct the drone to navigate towards it. Therefore, the drones within the drone group start to pick up each other's rotor noise, thus resulting in the drones drawing into each other. In a spiralling effect, the closer the drones get, the louder the drone noise becomes between each other, thus creating a heightened likelihood of the adjacent drone noise being treated as the target sound source while losing the true target source and leading to their eventual collapse (see Figure 9). While the rotor noise reduction methods (in particular the WF) could reduce the tracking error to some degree, it was still not enough for the tracking results to be meaningful, as shown by the high errors shown in Table 4 at input SNRs below 0 dB. This is expected even with the rotor noise reduction WF applied, since the WF only accounts for noise radiated from the drone's rotor noise but not those around it. Therefore, a means to explicitly reduce the influence of rotor noise arriving from drones other than itself is necessary. Alternatively, a policy that will force the drones to not arrive at some threshold distance closer to each other is needed. This distance would ultimately be a trade-off between the separation of the drones and the closeness to the target sound source (all factors that would eventually determine the input SNR). While it is desirable to separate the drones far enough to minimise the exposure to adjacent drone noise, doing so may increase the distance between the drones and the target source, thus decreasing the input SNR or its likelihood of detecting the target. Therefore, a means of monitoring this balance would be beneficial. This may also help with the placement planning process for each drone, as there is a possibility that the optimal positions planned may cause drones to cross paths, thus leading to collisions between the drones.

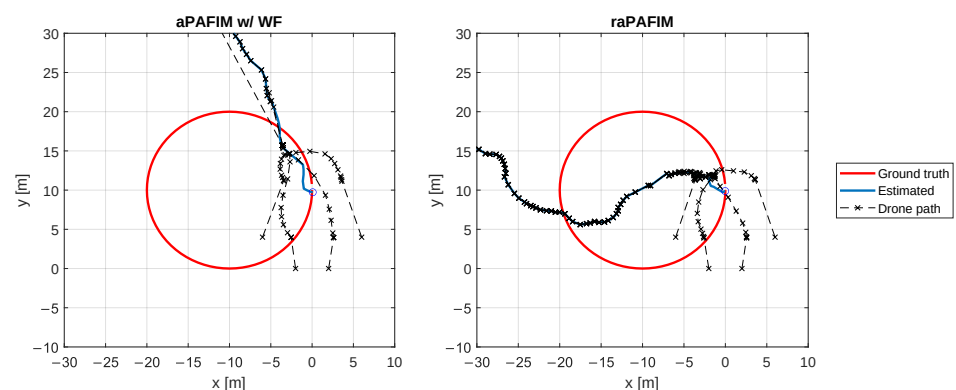


Figure 9. Example sound source trajectory at input SNR of 0 dB in 2D world coordinates (x–y plane) for the aPAFIM with WF and raPAFIM methods with adjacent drone noise accounted for. The blue circle indicates the initial location estimate. The black dashed lines are the travel paths of the four drones.

5.2. Two Sound Source Scenarios

Tables 5 and 6 show the evaluation results in terms of the Euclidean distance between the true and estimated locations for the two sound source scenarios (a) and (b) (see Section 4.3), respectively. Examples of the estimated trajectory for all methods are shown in Figure 10 and Figure 11 for scenario (a) and scenario (b), respectively.

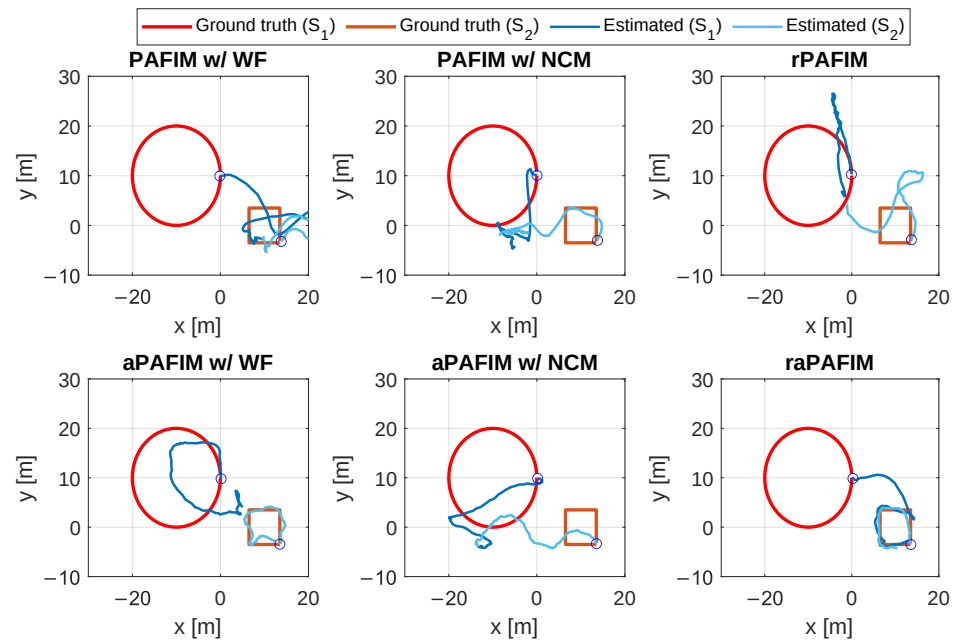


Figure 10. Example sound source trajectory at an input SNR of -10 dB in 2D world coordinates (x - y plane) for two-source scenario (a) (without accounting for noise arriving from adjacent drones). The blue circles indicate the initial location estimates. Trajectories from PAFIM and aPAFIM are not shown due to the methods failing to complete the simulation (see Table 5).

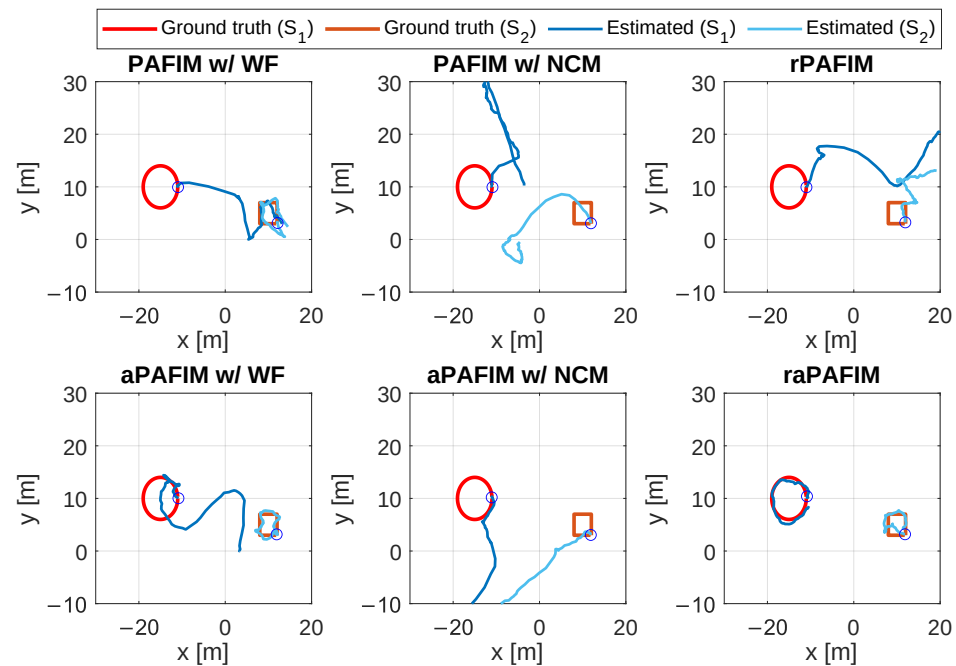


Figure 11. Example sound source trajectory at input SNR of -10 dB in 2D world coordinates (x - y plane) for two source scenario (b) (without accounting for noise arriving from adjacent drones). The blue circles indicate the initial location estimates. Trajectories from PAFIM and aPAFIM are not shown due to the methods failing to complete the simulation (see Table 6).

Table 5. Mean Euclidean distance between the estimated and true source trajectory for two source scenario (a) (in metres). Note that the results here **did not account for the noise that arrives from adjacent drones**. The lowest distance measures are highlighted in bold. N/A indicates that the simulation failed to complete.

Method	Input SNR (dB)					
	−30	−20	−10	0	10	20
Source 1 (S_1)						
PAFIM (Baseline)	36.10	28.77	36.78	37.82	37.26	20.40
PAFIM w/WF [32]	23.63	23.29	24.67	22.23	20.45	22.08
PAFIM w/NCM [32]	24.88	26.34	27.31	23.74	25.73	28.75
rPAFIM [32]	19.21	14.80	19.00	13.17	22.30	14.17
aPAFIM [30]	N/A	N/A	N/A	N/A	8.93	29.55
aPAFIM w/WF [30,32]	2.99	22.53	15.08	18.66	2.88	12.00
aPAFIM w/NCM [30,32]	36.94	33.17	30.05	27.77	27.37	21.19
raPAFIM [30,32]	21.31	21.79	20.35	20.78	18.87	21.11
Source 2 (S_2)						
PAFIM (Baseline)	25.21	33.12	25.97	24.34	21.36	11.20
PAFIM w/WF [32]	14.11	5.80	9.53	5.11	6.84	11.00
PAFIM w/NCM [32]	27.12	24.68	23.12	22.92	21.22	22.37
rPAFIM [32]	29.24	13.39	10.24	12.98	9.00	10.97
aPAFIM [30]	N/A	N/A	N/A	N/A	12.99	0.48
aPAFIM w/WF [30,32]	2.66	1.79	1.61	1.71	1.39	0.75
aPAFIM w/NCM [30,32]	27.92	24.33	15.24	24.14	10.97	0.84
raPAFIM [30,32]	2.13	1.89	1.87	2.11	1.37	0.79

Table 6. Mean Euclidean distance between the estimated and true source trajectory for two source scenario (b) (in metres). Note that the results here **did not account for noise arriving from adjacent drones**. The lowest distance measures are highlighted in bold. N/A indicates that the simulation failed to complete.

Method	Input SNR (dB)					
	−30	−20	−10	0	10	20
Source 1 (S_1)						
PAFIM (Baseline)	N/A	N/A	N/A	N/A	27.06	17.07
PAFIM w/WF [32]	8.69	13.43	22.22	8.68	6.56	3.33
PAFIM w/NCM [32]	27.84	27.39	17.14	21.90	21.51	27.83
rPAFIM [32]	22.45	24.02	23.98	21.86	17.54	25.11
aPAFIM [30]	N/A	N/A	N/A	N/A	3.32	1.09
aPAFIM w/WF [30,32]	17.24	15.27	14.75	18.18	8.67	12.04
aPAFIM w/NCM [30,32]	29.03	14.41	29.24	17.22	32.76	19.62
raPAFIM [30,32]	20.27	20.38	3.09	21.65	2.60	1.44
Source 2 (S_2)						
PAFIM (Baseline)	N/A	N/A	N/A	N/A	5.13	2.20
PAFIM w/WF [32]	3.10	4.52	1.92	2.33	2.28	5.45
PAFIM w/NCM [32]	26.75	25.56	24.03	27.13	29.04	4.02
rPAFIM [32]	15.23	7.57	7.46	8.03	5.20	3.01
aPAFIM [30]	N/A	N/A	N/A	N/A	9.80	0.56
aPAFIM w/WF [30,32]	1.16	0.98	1.26	0.93	1.10	0.80
aPAFIM w/NCM [30,32]	33.59	19.04	25.93	22.60	25.67	1.15
raPAFIM [30,32]	2.10	2.21	3.00	1.67	2.74	1.87

Similar to the single-source scenario results, PAFIM with WF showed the highest tracking performance improvement over the baseline for scenario (a). However, it is also

apparent that source S_2 was tracked much more accurately than source S_1 . This result is consistent with or without array placement planning. For methods without array placement planning, this is expected, as with the true trajectory of S_2 's travel being much smaller than that of S_1 for most times, the sound source was relatively close to the drones (microphone arrays), thus leading to better tracking performance. On the contrary, while methods with array placement planning generally outperformed methods without them, the contrast between the tracking performance of S_1 and S_2 was greater. This is due to the fact that the probability of observation update (see Section 3.3.1) relies on the sound source being apparent for the MUSIC method to recognise. When the sound source becomes far away relative to another sound source, the more apparent sound source will gradually cause the drones originally designated to track another sound source towards it. For example, as reflected in Figure 12a, drone MA_1 , which originally tracked S_1 , gradually transitioned its flight path towards S_2 , which became a less optimal placement for S_1 and caused its tracking accuracy to decrease. This phenomenon was apparent in many methods assessed in this study (see Table 5). In fact, several cases showed that all drones, after a certain time step, started to track only one of the sources, and it can be seen that the trajectory, which originally followed that of one sound source, transitioned to follow another part way through the tracking process (see Figure 13a). This transition from one source to another seemed to be influenced significantly by the input SNR between the two sound sources in relation to the drone during their respective positions at the time. For example, in a special case using raPAFIM, where S_1 was 15 dB louder than S_2 (at 0 dB input SNR, see Figure 14) relative to when both sound sources were both at 0 dB input SNR (see Figure 13), the almost immediate transition in the sound source was eliminated. It was only until the end that S_2 transitioned into the S_1 path, which was likely due to the two sound sources arriving close to each other. This is expected, because the triangulation process of PAFIM relies on the MUSIC spectrum calculated from each microphone array, which is influenced by the SNRs relative to each target source. This phenomenon is also reflected in the previous study [30], although to a much lesser degree, and the reasoning behind this difference is likely similar to that discussed with the single sound source scenario. Therefore, to ensure robust tracking of multiple sound sources, a way to enable the algorithm to fare against different levels of input SNRs between sound sources is necessary.

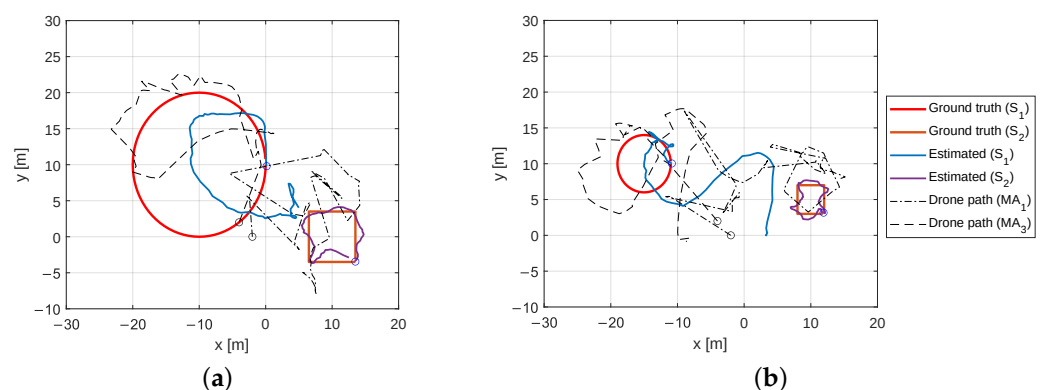


Figure 12. Example sound source trajectory at an input SNR of 0 dB in 2D world coordinates (x-y plane) using aPAFIM with WF for two-source scenarios (a,b). The blue circles indicate the initial location estimates. The black dashed trajectories are the travel paths of drones MA_1 and MA_3 , with the black circles indicating their starting positions.

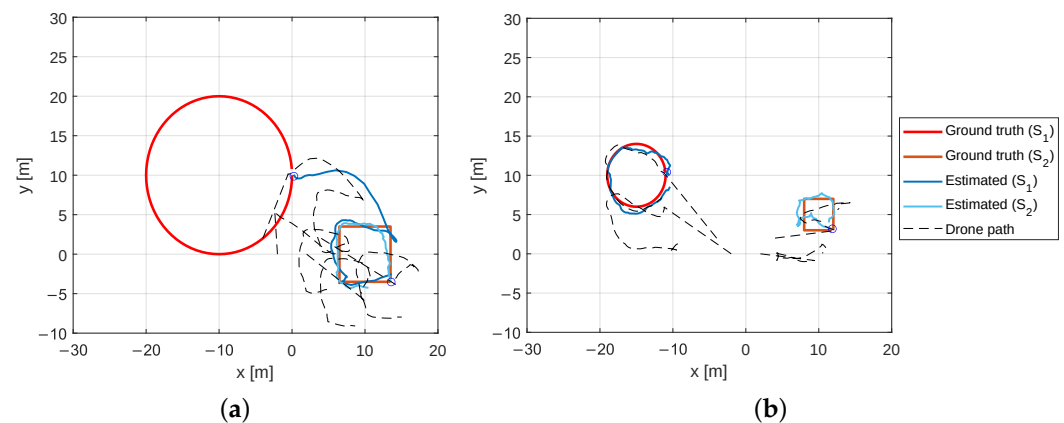


Figure 13. Example sound source trajectory at an input SNR of 0 dB in 2D world coordinates (x–y plane) using raPAFIM for the two sound source scenarios (a,b). The blue circles indicate the initial location estimates. The black dashed trajectories are the travel paths of the six drones.

The trend shown in the results for scenario (a) was also apparent in scenario (b) (see Table 6 and Figure 11), where again PAFIM with WF (with or without array placement planning) showed overall the best SST performance, which is consistent with that shown in [32]. Furthermore, the issue of some drones shifting towards tracking another sound source was also apparent in scenario (b). Therefore, regardless of the distance separation between the two sound sources, there is still a high possibility that drones that were to track one source may be driven towards another. As such, means to mitigate such shortcomings are vital in allowing the array placement planning to consistently track multiple sound sources successfully, which is a main component for future work.

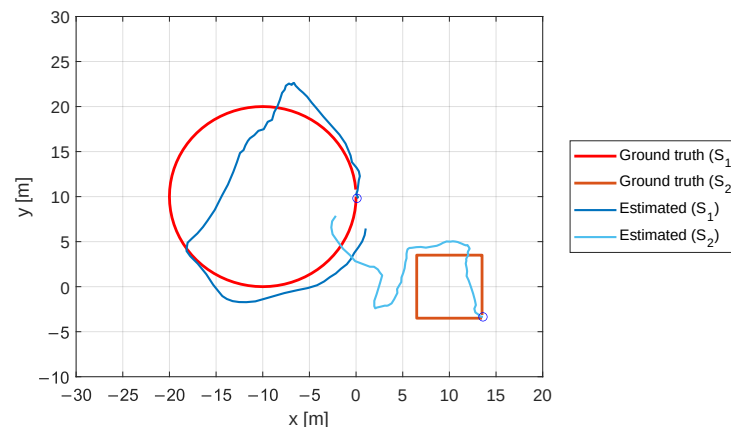


Figure 14. Special case example sound source trajectory in 2D world coordinates (x–y plane) for two source scenario (a) using the raPAFIM method. S_1 has an input SNR of 15 dB and 0 dB for S_2 (without accounting for noise arriving from adjacent drones). The blue circles indicate the initial location estimates.

It should be noted that in contrast to the single sound source scenario (see Section 5.1), aPAFIM failed to complete the simulation under lower input SNRs (as shown by the N/A entries in Tables 5 and 6). It appears that over a certain time step, no drones were able to achieve a probability of observation $p(\alpha_{i \rightarrow k} | \mathbf{z}) \geq p_{thre}$ for either one or both of the sound sources. This was the case for both two source scenario (a) and scenario (b), and is a result that is very different to that from [30], where all simulations were completed without issues. However, it is interesting to note that for scenario (b), PAFIM without array placement planning also failed to complete the simulation for most input SNRs, which is different from the results seen in most scenarios. Note that while the drones do not move in the

baseline PAFIM method, the probability of observation update is required when tracking multiple sources. In this case, it is the source S_1 in which PAFIM eventually fails to track due to no drones having a probability of observation that is above the p_{thres} threshold. Indeed, since S_1 is effectively placed out of the drone's vicinity and relatively far away, tracking it becomes much more challenging. With PAFIM relying on past location estimates to calculate the current sound source location, its tracking errors accumulate over time, thus causing the method to lose track of the sound source (see Figure 12b. Note that drone MA₁'s flight path, like that in Figure 12a, transitioned from S_1 to S_2). However, in contrast to scenario (a) for raPAFIM, the method managed to track both sources well in scenario (b), as shown in Figure 13b. Therefore, the sound source being farther apart seemed to help prevent the drones from drawing towards another sound source undesirably. It is also interesting to note that while for scenario (a), PAFIM appeared to fail to track both sources across most input SNRs, in scenario (b), S_2 was well tracked at an input SNR of 10 dB and above; this is a result that is similar to methods with array placement planning. This is expected, since the probability of observation update process is required for multiple sound sources, even if the drones remain static. Therefore, under low-input SNR conditions, the observation of the sound source would shift from one source to another. Thus, ways to allow more robust identification of sound sources are necessary.

Overall, it is apparent that similar to the single sound source scenarios, aPAFIM with WF and raPAFIM are the two methods that showed decent performance consistent across most input SNRs. This further emphasises that rotor noise reduction is crucial to ensuring that the PAFIM-based algorithm functions under realistic (low-input SNR) conditions and guaranteeing that the array placement planning's effectiveness is maximised.

5.3. Summary

Overall, despite the simulation scenario being more challenging than those shown in previous studies [30,32], both the array planning algorithm and rotor noise reduction methods showed significant improvement in SST performance. In particular, the WF for rotor noise reduction and the array placement planning algorithm are the more significant components of performance improvement, and they performed well, even under low-input SNRs in most cases. Therefore, based on the simulation results, a well-designed WF is paramount to improve SST performance, especially with the array placement planning algorithm. However, through the various scenarios tested, many aspects of the system still require further improvement. We summarise those considerations and suggestions as follows:

- A noise reduction WF that not only considers the rotor noise of its drone but also considers noise from other drones is required.
- A policy (design criteria) is required to limit drones from getting too close to each other.
- In a multisound source scenario with more assertive means to enforce drones only to track their designated sound source is required, especially when there is a significant difference in the sound levels perceived by the drones between the target sound sources.

6. Conclusions

In this study, we compared and performed a detailed assessment of the recent developments in multidrone sound source tracking algorithms, including a rotor noise reduction method and an array placement planning algorithm to navigate drones to their optimal recording positions. Both methods showed significant improvement in SST performance. However, the methods only worked well under input SNRs and were not representative of actual drone scenarios. When evaluated under more realistic settings, such as the consideration of noise from neighbouring drones and more challenging initial drone placements, most methods struggled with SST accuracy. This resulted in a list of additional design factors that require consideration to ensure a more robust sound source tracking performance. Future work includes addressing the shortcomings discovered in this perfor-

mance assessment and implementing the multidrone sound source tracking system in a real-life system.

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

Funding: This research was supported by the commissioned research fund provided by F-REI (JPFR23010102) and JSPS KAKENHI through Grant No. 22F22769 and Grant No. 22KF0141.

Data Availability Statement: Data will be made available on request.

Conflicts of Interest: Author Katsutoshi Itoyama was employed by the company Honda Research Institute Japan Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CASA	Computational auditory scene analysis
GCC-PHAT	Generalised crosscorrelation phase transform
GSVD-MUSIC	Generalised Singular Value Decomposition MUSIC
MT-GSFT	Multiple Triangulation and Gaussian Sum Filter Tracking
MUSIC	MULTiple Signal Classification
PAFIM	Particle Filtering with MUSIC
PSD	Power spectral density
SNR	Signal-to-noise ratio
SSL	Sound source localisation
SST	Sound source tracking
STFT	Short-time Fourier Transform
NCM	Noise correlation matrix
WF	Wiener filter

References

1. Koubaa, A.; Azar, A. (Eds.) *Unmanned Aerial Systems*; Elsevier: Amsterdam, The Netherlands, 2021. [\[CrossRef\]](#)
2. Karaca, Y.; Cicek, M.; Tatli, O.; Sahin, A.; Pasli, S.; Beser, M.F.; Tureli, S. The potential use of unmanned aircraft systems (drones) in mountain search and rescue operations. *Am. J. Emerg. Med.* **2018**, *36*, 583–588. [\[CrossRef\]](#) [\[PubMed\]](#)
3. 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. [\[CrossRef\]](#)
4. Martinez-Carranza, J.; Rascon, C. A review on auditory perception for unmanned aerial vehicles. *Sensors* **2020**, *20*, 7276. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Nakadai, K.; Okuno, H.G. Robot audition and computational auditory scene analysis. *Adv. Intell. Syst.* **2020**, *2*, 2000050. [\[CrossRef\]](#)
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 Sens. J.* **2018**, *19*, 332–341. [\[CrossRef\]](#)
7. Brandstein, M.; Ward, D. *Microphone Arrays: Signal Processing Techniques and Applications*; Digital Signal Processing; Springer: Berlin/Heidelberg, Germany, 2001.
8. Schmidt, R. Multiple emitter location and signal parameter estimation. *IEEE Trans. Antennas Propag.* **1986**, *34*, 276–280. [\[CrossRef\]](#)
9. Wang, L.; Cavallaro, A. Ear in the sky: Ego-noise reduction for auditory micro aerial vehicles. In Proceedings of the 2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Colorado Springs, CO, USA, 23–26 August 2016; pp. 152–158.
10. Okutani, K.; Yoshida, T.; Nakamura, K.; Nakadai, K. Outdoor auditory scene analysis using a moving microphone array embedded in a quadcopter. In Proceedings of the Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference, Vilamoura-Algarve, Portugal, 7–12 October 2012; pp. 3288–3293.
11. Strauss, M.; Mordel, P.; Miguet, V.; Deleforge, A. DREGON: Dataset and methods for UAV-embedded sound source localization. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 1–8.

12. Furukawa, K.; Okutani, K.; Nagira, K.; Otsuka, T.; Itoyama, K.; Nakadai, K.; Okuno, H.G. Noise correlation matrix estimation for improving sound source localization by multirotor UAV. In Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Tokyo, Japan, 3–7 November 2013; pp. 3943–3948.
13. 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), Vancouver, BC, Canada, 24–28 September 2017; pp. 5985–5990.
14. Wang, L.; Cavallaro, A. Acoustic sensing from a multi-rotor drone. *IEEE Sens. J.* **2018**, *18*, 4570–4582. [\[CrossRef\]](#)
15. Manamperi, W.; Abhayapala, T.D.; Zhang, J.; Samarasinghe, P.N. Drone audition: Sound source localization using on-board microphones. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 508–519. [\[CrossRef\]](#)
16. Choi, J.; Chang, J. Convolutional Neural Network-based Direction-of-Arrival Estimation using Stereo Microphones for Drone. In Proceedings of the 2020 International Conference on Electronics, Information, and Communication (ICEIC), Barcelona, Spain, 19–22 January 2020; pp. 1–5.
17. Yen, B.; Hioka, Y. Noise power spectral density scaled SNR response estimation with restricted range search for sound source localisation using unmanned aerial vehicles. *Eurasip J. Audio Speech Music. Process.* **2020**, *2020*, 1–26. [\[CrossRef\]](#)
18. Wang, L.; Cavallaro, A. Deep-Learning-Assisted Sound Source Localization From a Flying Drone. *IEEE Sens. J.* **2022**, *22*, 20828–20838. [\[CrossRef\]](#)
19. 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), Queensland, Australia, 19–24 April 2015; pp. 2699–2703.
20. Schmidt, A.; Löllmann, H.W.; Kellermann, W. Acoustic self-awareness of autonomous systems in a world of sounds. *Proc. IEEE* **2020**, *108*, 1127–1149. [\[CrossRef\]](#)
21. 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.
22. 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, Korea, 9–14 October 2016; pp. 1293–1298. [\[CrossRef\]](#)
23. Potamitis, I.; Chen, H.; Tremoulis, G. Tracking of multiple moving speakers with multiple microphone arrays. *IEEE Trans. Speech Audio Process.* **2004**, *12*, 520–529. [\[CrossRef\]](#)
24. 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. [\[CrossRef\]](#)
25. 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, 25–29 October 2020; pp. 10402–10407.
26. 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.
27. 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 (SII), Honolulu, HI, USA, 12–15 January 2020; pp. 796–801.
28. 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. [\[CrossRef\]](#) [\[PubMed\]](#)
29. Yamada, T.; Itoyama, K.; Nishida, K.; Nakadai, K. Outdoor evaluation of sound source localization for drone groups using microphone arrays. In Proceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan, 23–27 October 2022; pp. 9296–9301.
30. Yamada, T.; Itoyama, K.; Nishida, K.; Nakadai, K. Placement Planning for Sound Source Tracking in Active Drone Audition. *Drones* **2023**, *7*, 405. [\[CrossRef\]](#)
31. Suenaga, M.; Shimizu, T.; Hatanaka, T.; Uto, K.; Mammarella, M.; Dabbene, F. Experimental Study on Angle-aware Coverage Control with Application to 3-D Visual Map Reconstruction. In Proceedings of the 2022 IEEE Conference on Control Technology and Applications (CCTA), Trieste, Italy, 22–25 August 2022; pp. 327–333. [\[CrossRef\]](#)
32. Yen, B.; Yamada, T.; Itoyama, K.; Nakadai, K. Performance evaluation of sound source localisation and tracking methods using multiple drones. In Proceedings of the INTER-NOISE and NOISE-CON Congress and Conference Proceedings, Chiba, Japan, 20–23 August 2023; Volume 268, pp. 1926–1937.
33. 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.
34. Yen, B.; Hioka, Y.; Schmid, G.; Mace, B. Multi-sensory sound source enhancement for unmanned aerial vehicle recordings. *Appl. Acoust.* **2022**, *189*, 108590. [\[CrossRef\]](#)
35. Wang, L.; Cavallaro, A. Deep learning assisted time-frequency processing for speech enhancement on drones. *IEEE Trans. Emerg. Top. Comput. Intell.* **2020**, *5*, 871–881. [\[CrossRef\]](#)

36. Tengan, E.; Dietzen, T.; Ruiz, S.; Alkmim, M.; Cardenuto, J.; van Waterschoot, T. Speech enhancement using ego-noise references with a microphone array embedded in an unmanned aerial vehicle. In Proceedings of the 2022 24th International Congress of Acoustics (ICA 2022), Gyeongju, Republic of Korea, 24–28 October 2022; pp. 1–5.
37. Manamperi, W.N.; Abhayapala, T.D.; Samarasinghe, P.N.; Zhang, J.A. Drone audition: Audio signal enhancement from drone embedded microphones using multichannel Wiener filtering and Gaussian-mixture based post-filtering. *Appl. Acoust.* **2024**, *216*, 109818. [\[CrossRef\]](#)
38. Hioka, Y.; Kingan, M.; Schmid, G.; McKay, R.; Stol, K.A. Design of an unmanned aerial vehicle mounted system for quiet audio recording. *Appl. Acoust.* **2019**, *155*, 423–427. [\[CrossRef\]](#)
39. Ohata, T.; Nakamura, K.; Mizumoto, T.; Taiki, T.; Nakadai, K. Improvement in outdoor sound source detection using a quadrotor-embedded microphone array. In Proceedings of the 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, Chicago, IL, USA, 14–18 September 2014; pp. 1902–1907.
40. 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. [\[CrossRef\]](#)
41. Yamagishi, J.; Christophe, V.; Kirsten, M. CSTR VCTK Corpus: English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit (Version 0.92), [Sound]. University of Edinburgh. The Centre for Speech Technology Research (CSTR). 2019. Data Retrieved from Edinburgh DataShare. Available online: <https://datashare.ed.ac.uk/handle/10283/3443> (accessed on 9 March 2021).

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.