



Research Article

Acoustic-enhanced local bearing estimation using low-cost microphones for Micro Air Vehicle swarms

Aohua Li, Ye Zhou, Weijie Kuang, Hann Woei Ho*

School of Aerospace Engineering, Engineering Campus, Universiti Sains Malaysia, Nibong Tebal 14300, Malaysia

ARTICLE INFO

Article history:

Received 25 April 2025

Revised 4 July 2025

Accepted 8 August 2025

Available online 10 September 2025

Keywords:

Local bearing estimation

MAV swarms

Sound source localization

Microphone array

ABSTRACT

Micro Air Vehicle (MAV) swarms are often constrained by limited onboard processing capabilities and payload capacity, restricting the use of sophisticated localization systems. Lightweight ultra-wideband (UWB) ranging techniques are commonly used to estimate inter-vehicle distances, but they do not provide local bearing information—essential for precise relative positioning. Inspired by bat echolocation in low-visibility environments, we propose an acoustic-enhanced method for local bearing estimation designed for low-cost MAVs. Our approach leverages ambient acoustic signals naturally emitted by a target MAV in flight, combined with UWB distance measurements. The acoustic data is processed using the Frequency-Sliding Generalized Cross-Correlation (FS-GCC) method, enhanced with our analytical formulation that compensates for inter-channel switching delays in asynchronous, high-frequency sampling. This enables accurate Time Difference of Arrival (TDOA) estimation, even with compact microphone arrays. These TDOA values, along with known microphone geometry and UWB data, are integrated into our geometric model to estimate the bearing of the target MAV. We validate our approach in a controlled indoor hall across two experimental scenarios: static-bearing estimation, where the target MAV hovers at predefined angular positions (0° , $\pm 30^\circ$, $\pm 45^\circ$, $\pm 60^\circ$), and dynamic-bearing estimation, where it flies across angles at varying velocities. The results show that our method yields reliable TDOA measurements compared to classical and machine learning baselines, and produces accurate bearing estimates in both static and dynamic settings. This demonstrates the feasibility of our low-cost acoustic-enhanced solution for local bearing estimation in MAV swarms, supporting improved relative navigation and decentralized perception in GPS-denied or visually degraded environments.

© 2025 The Author(s). Published by Elsevier B.V. on behalf of Shandong University. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Micro Air Vehicles (MAVs) are rapidly transforming operations in challenging environments, such as tunnel inspection, archaeological exploration, and wildlife monitoring [1–3]. In these applications, MAV swarms offer enhanced robustness, scalability, and redundancy, enabling distributed decision-making and mission execution without relying on centralized control. A key capability that underpins such swarm autonomy is swarm perception—the ability of each MAV to estimate the relative position of neighboring agents in real time for navigation, collision avoidance, and coordination [4].

Swarm perception typically requires estimating both distance and direction (bearing) to nearby MAVs. A range of methods has been explored to tackle this challenge, including Visual-Inertial Odometry (VIO), optical flow, and camera-based methods [5–8]. While effective under favorable conditions, these techniques often degrade in low-visibility environments due to poor

lighting, occlusion, or a lack of distinctive visual features. This makes them less reliable for swarm navigation in dark, dusty, cluttered, or underground settings.

To address visibility-related limitations, Ultra-Wideband (UWB) ranging approaches have been increasingly adopted in MAVs [9] due to their lightweight form factor, low power requirements, and robustness in visually degraded conditions. UWB systems can provide accurate inter-MAV distance measurements, making them suitable for lightweight and distributed setups. However, they are fundamentally limited in that they do not provide bearing information, and thus do not fully resolve the swarm perception problem.

In this work, we draw inspiration from the natural world—specifically, from bats, which navigate in darkness using echolocation. Bats estimate the direction and distance of nearby objects by processing time delays and intensity differences between their two ears [10,11]. As shown in Fig. 1, a bat emits an acoustic pulse that reflects off a target and returns to its ears; the interaural time delay allows the bat to infer both distance and direction. Inspired by this biological strategy, we propose an

* Corresponding author.

E-mail address: aehannwoei@usm.my (H.W. Ho).

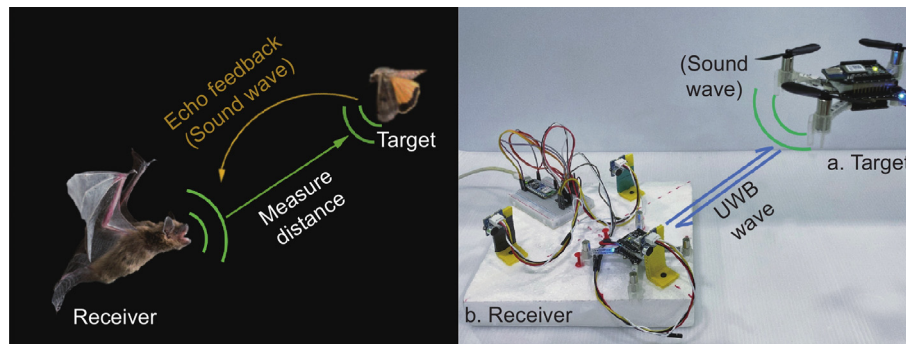


Fig. 1. Bio-inspired concept for local bearing estimation. *Left:* Bats emit sound pulses and process returning echoes to estimate the location of nearby objects using interaural time delays. *Right:* Our system emulates this principle using two MAVs. (a) The target MAV is equipped with a UWB module and emits sound during flight. (b) The receiver MAV integrates a microphone array and UWB module to acquire acoustic signals and distance data, enabling bearing estimation through TDOA processing.

acoustic-enhanced bearing estimation approach for MAV swarms. Our system emulates this behavior using a lightweight microphone array combined with a UWB ranging module, as illustrated on the right side of Fig. 1. While the UWB module provides distance information, the microphone array captures the directional acoustic cues necessary to estimate the bearing of a target MAV in flight.

The *main contributions* of this work are as follows:

- We propose an acoustic-enhanced framework for estimating the local bearing of neighboring MAVs by fusing sound-based TDOA with UWB distance measurements. This bio-inspired design motivated by the echolocation capabilities of bats, where a microphone array emulates spatial hearing and TDOA mimics interaural time delay processing, illustrated in Fig. 1.
- We introduce an analytical formulation to translate asynchronous high-frequency sample differences from multiple acoustic channels into accurate TDOA estimates by compensating for inter-channel switching delays. This improves TDOA estimation accuracy under high sampling rates and within compact microphone array configurations.
- We validate the proposed system in an indoor hall (50 m \times 40 m \times 15 m), where a receiver MAV estimates the bearing of a target MAV hovered at various angular positions (0° , $\pm 30^\circ$, $\pm 45^\circ$, $\pm 60^\circ$) and at multiple distances, as well as during continuous flight across these angles at different velocities. Experimental results demonstrate robust and accurate bearing estimation in both static and dynamic conditions using a lightweight, low-cost sensing setup, supporting deployment in GPS-denied and visually degraded environments.

The remainder of this paper is structured as follows: Section 2 reviews prior work on target bearing estimation techniques and TDOA estimation methods. Section 3 details the proposed methodology, including high-frequency acoustic signal acquisition, TDOA estimation using the improved FS-GCC algorithm, the geometric model for local bearing estimation, and the nominal bearing estimation under hardware constraints for accuracy evaluation. Section 4 presents and discusses the experimental results obtained from real-world indoor tests that validate the effectiveness of the proposed approach. Finally, Section 5 concludes the paper and outlines potential directions for future work based on current system limitations.

2. Related work

This section reviews prior work related to local bearing estimation, with a focus on techniques involving acoustic sensing

and TDOA methods. We categorize the literature into visual, beamforming-based, and learning-driven approaches, followed by a discussion on traditional and advanced methods for TDOA estimation.

2.1. Techniques for bearing estimation

Estimating the bearing of a target object is a fundamental challenge in many perception systems, including surveillance, robotics, and aerial navigation. Traditional approaches often rely on visual sensing. For instance, studies have proposed detection and tracking systems using high-resolution video cameras [12], distributed active sensing strategies using onboard cameras [13], and fiber-optic acoustic sensing systems for drone detection and localization [14]. While accurate, these methods depend heavily on visual or multimodal inputs, which limits their effectiveness in low-visibility environments and increases computational demands.

Beamforming-based techniques are widely used for acoustic bearing estimation. Classical Steered Response Power (SRP) methods have been enhanced with deep learning and visual inputs for improved accuracy [15]. Other approaches integrate beamforming with Generalized Cross-Correlation with Phase Transform (GCC-PHAT), differential evolution, or Fourier-based interpolation to estimate bearings [16–18]. These methods, however, often rely on signal weighting strategies that are sensitive to interference and require considerable computational resources.

In recent years, learning-based models have been developed to further enhance bearing estimation [19,20]. Some approaches leverage spiking neural networks with Hilbert-transform features [21], while others employ residual networks with channel attention mechanisms using log-Mel spectrogram and GCC-PHAT features [22]. These models primarily aim to improve TDOA estimation from microphone arrays—an essential aspect also addressed in our proposed method.

2.2. Time Difference Of Arrival (TDOA) estimation

TDOA is a fundamental technique for localizing sound sources by measuring the relative time lag between sensor inputs. The Generalized Cross-Correlation (GCC) method [23] remains one of the most established tools, with adaptations based on various weighting functions to improve performance across acoustic scenarios [24,25]. The GCC-PHAT variant is especially robust under noisy conditions and has seen widespread application [26,27].

To address the limitations of conventional GCC methods, the Frequency-Sliding GCC-PHAT (FS-GCC) technique was proposed, introducing a sliding window over the cross-power spectrum to

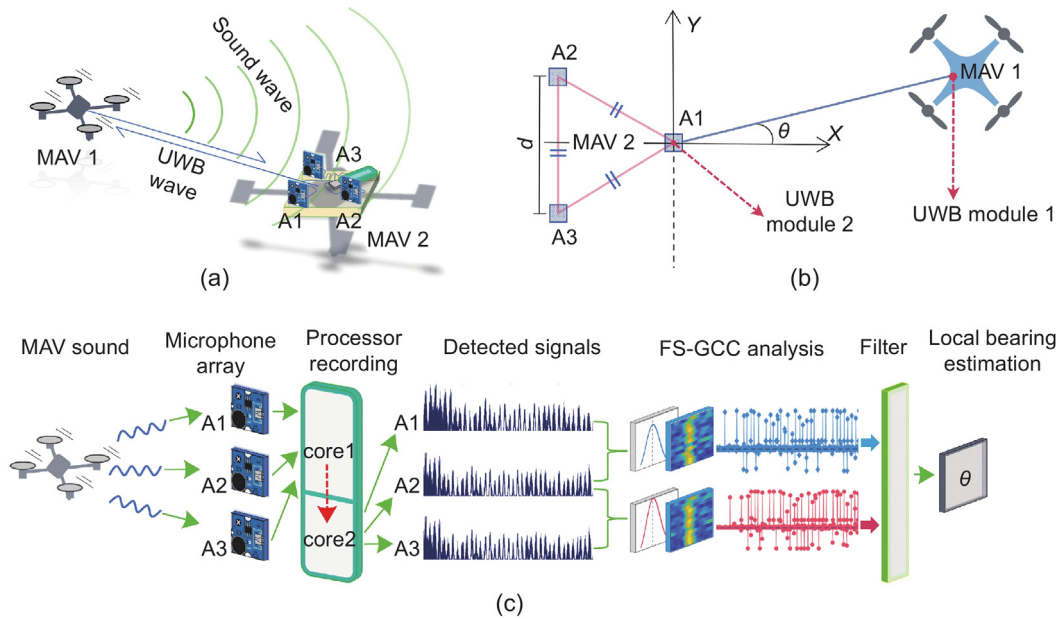


Fig. 2. Local bearing estimation by integrating MAV sound signals and UWB data. (a) MAV 1 and MAV 2 are both equipped with a UWB module. MAV 1 emits sound during flight and acts as the target MAV. MAV 2 is equipped with a microphone array which captures the sound and serves as the receiver MAV. (b) Microphones A1, A2, and A3 form the microphone array, arranged within a horizontal plane. (c) The overview of acoustic-enhanced local bearing estimation framework.

analyze sub-band correlations [28]. This provides improved temporal resolution and robustness against noise and reverberation. Additionally, techniques like weighted Singular Value Decomposition (SVD) have been explored to enhance the stability and accuracy of TDOA estimation [29].

In parallel, machine learning approaches have emerged as data-driven alternatives [30,31]. For example, LSTM networks have been trained to regress TDOA values directly from raw or preprocessed audio inputs, leveraging their ability to model temporal dependencies in sequential data [31]. These methods show potential in handling complex acoustic environments but often require large, labeled datasets and substantial computational resources.

Although prior research has primarily focused on refining TDOA accuracy through advanced signal processing or machine learning, relatively few have explored integrating acoustic TDOA with range-based sensing such as UWB, especially under constraints like inter-channel switching delays in low-cost systems. In this work, we extend these methods by fusing FS-GCC-derived TDOA with UWB distance measurements to recover both direction and relative position of neighboring MAVs, thereby enabling accurate local bearing estimation using lightweight and low-cost sensors.

3. Methodology

This section outlines our bio-inspired approach for estimating the local bearing of a target MAV using low-cost acoustic and ranging sensors. The overall system architecture is summarized in Fig. 2. In this setup as shown in Fig. 2(a), MAV 1 acts as the target, while MAV 2 functions as the receiver. Both are equipped with UWB modules to exchange inter-vehicle distance measurements. In addition, MAV 2 is outfitted with a lightweight microphone array to capture the acoustic signals emitted by MAV 1 during flight, as described in Section 3.1.

As illustrated in Fig. 2(b), the spatial arrangement of microphones on MAV 2 introduces measurable differences in the arrival times of the sound signal from the target MAV, due to variations in propagation paths. These time differences are extracted

using the Frequency-Sliding Generalized Cross-Correlation (FS-GCC) method [28], as detailed in Section 3.2. FS-GCC enables robust TDOA estimation under noisy and reverberant conditions. To improve reliability, a filtering step is employed to discard outlier or invalid TDOA values.

Finally, as shown in Fig. 2(c), the filtered TDOA data is combined with UWB distance measurements and known microphone geometry to compute the local bearing of the target MAV. A geometric model is introduced in Section 3.3 to fuse these inputs and recover the relative angle, enabling accurate and efficient bearing estimation in GPS-denied or visually degraded environments. The code and experimental dataset used in this study are publicly available on Github.¹

3.1. Acquisition of asynchronous high-frequency sound signals

The main purpose of acoustic signal acquisition is to record arrival times and ensure sound quality as the same segment reaches different microphones. As shown in Fig. 2(b), the microphone array is configured as a polygon with arbitrary side lengths, and microphones A1, A2, and A3 have fixed coordinates. As MAV 1 flies, sound from its rotors propagates to each microphone, differing based on distance. This leads to TDOA between the pairs of microphones A1-A2 or A1-A3, as expressed in the following equation:

$$\Delta T_{i-j} = T_i - T_j \quad (1)$$

where i and j represent different microphone indices, and ΔT_{i-j} is the TDOA between microphone pairs A_i-A_j . For the specific microphone pairs considered in our analysis, we have $i = 1$ and $j = 2, 3$.

Due to the high speed of sound (331.29 m/s in standard air) and the small distance differences between the microphone pairs and the target MAV, the system must be capable of high-precision acoustic signal acquisition.

Fig. 3 illustrates the amplitude variations of sound signals. These are obtained from asynchronous sampling by three microphones at frequency f . The sound amplitude trends received by

¹ <https://github.com/iAerialRobo/Acoustic-Bearing-Estimation.git>

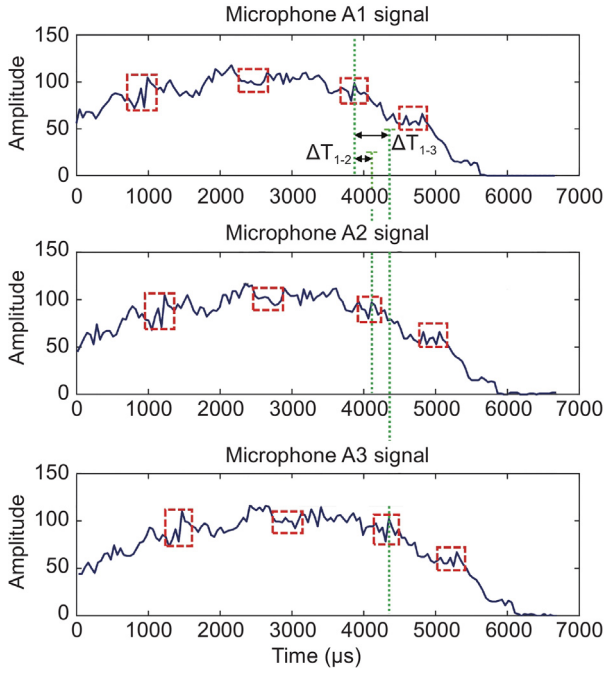


Fig. 3. The amplitude variation over time in MAV sound signals is asynchronously captured through three microphones.

different microphones are consistent, with several similar rapid shifts marked by the red dashed boxes. This suggests that the signals derive from the same sound segment and exhibit analogous amplitude variation properties. The green dotted line in the figure marks the similar rapid shift peaks, which serves as an example for estimating the TDOA between microphone pairs.

As the sound signals are captured as discrete samples, the process of acquiring data at a sampling frequency f introduces a fundamental time interval between consecutive samples, denoted as $\alpha = 1/f$ in μs , encompassing activation, data acquisition, and channel switching processes. Due to the sequential nature of sampling across 3 channels in the microphone array, inherent time differences arise when the sound wave arrives at different microphones in an array. Besides, there will be a 3α sampling time interval between consecutive data sample within each channel due to the sequential switching. Therefore, when determining the TDOA, ΔT_{i-j} , from the estimated sample difference, s_{i-j} , between a pair of microphones A_i - A_j , it is crucial to account for both the sampling time and the inter-channel delay, as follows:

$$\Delta T_{i-j} = \begin{cases} (s_{i-j}) \cdot 3\alpha + (j-i) \cdot \alpha, & \text{if } s_{i-j} \leq 0 \\ (s_{i-j}) \cdot 3\alpha - (3-j+i) \cdot \alpha, & \text{if } s_{i-j} > 0 \end{cases} \quad (2)$$

where $i = 1, j = 2, 3$. The sample difference, s_{i-j} , is an integer value, whose sign directly reflects the order of arrival of the sound wave at the microphone pair A_i and A_j . A negative value of s_{i-j} indicates that the sound wave reached microphone A_i before microphone A_j , whereas a positive s_{i-j} indicates that the arrival at A_j prior to A_i .

3.2. TDOA estimation using FS-GCC

To determine the TDOA between microphone pairs, this section employs the FS-GCC method [28]. Specifically, the FS-GCC technique will be utilized to estimate the sample difference, s_{i-j} , which is also known as TDOA in samples, between the signals received by the microphone pairs. Furthermore, specific parameter settings of the FS-GCC method have been adjusted to optimize its

performance for the characteristics of our acoustic signals and the requirements of our experimental setup.

As the target MAV 1 takes off, the sound generated by its rotors propagates through the experimental environment. The microphone pairs in the array capture these acoustic signals, and the initial representation of the recorded data for the selected pair A_i - A_j can be formulated by the following equation:

$$x_m[p] = \beta_m o[p - \eta_m] + w_m[p], \quad m = i, j \quad (3)$$

where $o[p]$ is the original sound signal from the MAV 1, $w_m[p]$ represents the noise part of the signal. p represents the discrete-time index, which identifies the sample position of the signals. $\beta_m \in \mathbb{R}^+$ denote an amplitude attenuation factor for microphone m , and η_m is the propagation delay. In the frequency domain representation using the discrete-time Fourier transform (DTFT), the microphone signals can be expressed as:

$$X_m(\omega) = \beta_m O(\omega) e^{-i\omega\eta_m} + W_m(\omega), \quad m = i, j \quad (4)$$

where $X_m(\omega)$ represents the received signal in the frequency domain. $O(\omega)$, $W_m(\omega) \in \mathbb{C}$ denote the DTFTs of the source and noise signals. The term $e^{-i\omega\eta_m}$ represents the phase shift caused by the delay η_m , and $i = \sqrt{-1}$ is the imaginary unit.

Sound signals often experience interference at specific frequencies. For instance, low frequencies may encounter echo issues due to wall reflections, while high frequencies might be masked by background noise. Instead of considering the full spectrum, we segment the sound into multiple sub-bands, with each sub-band processed individually using the following equation:

$$R[\tau, l] \triangleq \frac{1}{2\pi} \int_{-\pi}^{\pi} \Psi(\omega + \omega_l) \Phi(\omega) e^{i\omega\tau} d\omega = F^{-1}\{\Psi(\omega + \omega_l) \Phi(\omega)\} \quad (5)$$

where ω_l denotes the frequency offset for sub-band l , τ denotes the time delay, $\Psi(\omega + \omega_l)$ is the shifted cross-power spectrum that captures phase alignment in the target frequency range, and $\Phi(\omega)$ is a spectral window function used to suppress noise by isolating relevant frequency components. The inverse Fourier transform $F^{-1}\{\cdot\}$ converts the result to the time domain to obtain the delay profile for each sub-band. This sub-band processing enhances estimation accuracy by focusing on high-Signal-to-Noise Ratio (SNR) frequency regions while mitigating the influence of noise and reverberation.

Continuous-frequency signals can be transformed into discrete data for further processing. The next step involves using the Discrete Fourier Transform (DFT) to convert the waveform into a 'frequency spectrum' as follows:

$$\mathbf{X}_m = [X_m[0], X_m[1], \dots, X_m[N-1]]^T, \quad m = i, j \quad (6)$$

where \mathbf{X}_m is the column vector containing the DFT coefficients for microphone m , $X_m[\cdot]$ represents the Fourier coefficients at each of N different discrete frequencies, and the superscript T denotes the transpose operation, ensuring a column vector representation. In the discrete-frequency domain, the sub-band GCC is computed for each discrete sample time delay, which is then used to construct the GCC feature matrix $\mathbf{R} \in \mathbb{C}^{N \times L}$.

To prevent noise from affecting the assessment, Singular Value Decomposition (SVD) is applied to eliminate irrelevant information:

$$\mathbf{R} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^H \quad (7)$$

where $\mathbf{U} \in \mathbb{C}^{N \times L}$ contains the left singular vectors, $\mathbf{\Sigma} \in \mathbb{R}^{L \times L}$ is the diagonal matrix with ordered singular values, and $\mathbf{V} \in \mathbb{C}^{L \times L}$ contains the right singular vectors. $(\cdot)^H$ denotes the complex conjugate transpose.

To retain only the most significant components and suppress noise, a low-rank approximation is performed:

$$\mathbf{R}_k = \sum_{\ell=1}^k \sigma_{\ell} \mathbf{u}_{\ell} \mathbf{v}_{\ell}^T \quad (8)$$

where k is the number of retained singular values, σ_{ℓ} are the dominant singular values, and \mathbf{u}_{ℓ} and \mathbf{v}_{ℓ} are the corresponding singular vectors. This step effectively filters out noise and retains only the key structure of the GCC matrix for accurate TDOA estimation.

The first left singular vector \mathbf{u}_1 captures the dominant variation pattern in the GCC matrix along the sample difference dimension. It reflects the most significant structure of the signal – typically the primary amplitude trend – while suppressing components associated with background noise. This makes \mathbf{u}_1 highly effective for isolating the clean TDOA information from noisy data. To construct the characteristic value $\hat{\phi}_0[p]$, the real part of the first left singular vector \mathbf{u}_1 is processed. Specifically, the real component of each entry in \mathbf{u}_1 , denoted as $\Re\{u_1[p]\}$, is extracted, and the sign of the element corresponding to the maximum absolute value within $\Re\{\mathbf{u}_1\}$ is used to adjust the overall direction. The characteristic value $\hat{\phi}_0[p]$ is then defined as

$$\hat{\phi}_0[p] = \Re\{\mathbf{u}_1\} \cdot \text{sign}(\Re\{u_1[\gamma]\}) \quad (9)$$

where γ is defined as

$$\gamma = \arg \max_p |\Re\{u_1[p]\}| \quad (10)$$

indicating the index of the element with the largest absolute value in the real part of \mathbf{u}_1 . This characteristic value $\hat{\phi}_0[p]$ reflects the dominant component of the signal's variation at each sample point p . It is used to preserve the most significant pattern in the data while resolving any ambiguity related to the sign of the singular vectors.

To enhance robustness, a weighted low-rank approximation prioritizes reliable frequency bands. The approximation problem is then formulated as minimizing the weighted error norm

$$\min_{\hat{\mathbf{R}}} \left\| (\mathbf{R} - \hat{\mathbf{R}}) \odot \mathbf{W} \right\|_F, \quad \text{subject to } \text{rank}(\hat{\mathbf{R}}) \leq \ell \quad (11)$$

where $\hat{\mathbf{R}}$ represents the low-rank approximation of \mathbf{R} , $\mathbf{W} \in \mathbb{R}^{N \times L}$ is a weight matrix that provides different confidence levels to each sub-band GCC component. ℓ is the index for retained singular values and corresponding singular vectors, \odot denotes the Hadamard product (element-wise multiplication), and $\|\cdot\|_F$ represents the Frobenius norm, measuring the approximation error. This weighted optimization ensures that more reliable frequency bands contribute significantly to the final TDOA estimation, thereby improving robustness in noisy conditions.

After noise suppression, the final step is obtained by locating the peak of the processed GCC function:

$$\hat{s}_0 = \arg \max_p \hat{\phi}_0[p] \quad (12)$$

where $\hat{\phi}_0[p]$ is the GCC function obtained after noise suppression. By identifying the maximum peak in the processed GCC, we determine the most probable sample difference between the two microphone signals. The summarized procedure of the FS-GCC method for TDOA estimation is outlined in Algorithm 1.

3.3. A geometric model for local bearing estimation

This section combines the TDOA obtained from microphone pairs A_i - A_j with the distance information transmitted back from the UWB modules to estimate the local bearing. By analyzing the

Algorithm 1 FS-GCC Algorithm for TDOA Estimation

- 1: **Input:** Signals $x_i[p]$, $x_j[p]$; number of sub-bands L ; sub-band offsets ω_l ; spectral window $\Phi(\omega)$; weight matrix \mathbf{W}
- 2: **Output:** Estimated TDOA \hat{s}_0
- 3: Process data samples and compute the frequency spectrum of DFT (Eq. (6))
- 4: **for** Each sub-band $l = 0$ to $L - 1$ **do**
- 5: Shift spectrum by offset ω_l
- 6: Compute $\Psi(\omega - \omega_l)$ for different discrete frequencies
- 7: Apply spectral window $\Phi(\omega)$
- 8: Calculate sub-band GCC for each discrete sample time delay
- 9: **end for**
- 10: Stack all sub-band GCC to construct matrix \mathbf{R}
- 11: Apply weighted low-rank approximation with \mathbf{W} (Eq. (11))
- 12: Perform SVD on \mathbf{R} (Eq. (7))
- 13: Construct low-rank approximation \mathbf{R}_k using dominant components (Eq. (8))
- 14: Extract first left singular vector \mathbf{u}_1
- 15: Compute feature $\hat{\phi}_0[p]$ from \mathbf{u}_1 (Eq. (9))
- 16: Identify peak index of $\hat{\phi}_0[p]$ as TDOA estimate (Eq. (12))
- 17: **return** \hat{s}_0

geometry of the microphone array and the relative physical positions of MAV 1 and MAV 2 in a three-dimensional environment, we estimate the local bearing using a geometric model.

As illustrated in Fig. 2(c), microphone A1 is based on the coordinate origin, and the UWB module of MAV 2 was placed adjacent to A1. From this setup, the coordinates of each microphone in the two-dimensional plane can be determined. The relationship between the distance of MAV 2 from different microphones, r_1 , r_2 , r_3 , and the estimated TDOA, ΔT_{1-2} , ΔT_{1-3} , can be described by the following equation:

$$r_1 = d_{AP} \quad (13)$$

$$r_2 = v \cdot \Delta T_{1-2} + d_{AP} \quad (14)$$

$$r_3 = v \cdot \Delta T_{1-3} + d_{AP} \quad (15)$$

where v denotes the speed of sound, and d_{AP} represents the distance between MAV 1 and A1 on MAV 2 measured by the UWB modules. Next, we can determine the coordinates of the MAV 2 using the following equations:

$$x_s = \frac{2r_1^2 - r_2^2 - r_3^2 + 2x_2^2 + 2y_2^2}{4x_2} \quad (16)$$

$$y_s = \frac{r_3^2 - r_2^2}{4y_2} \quad (17)$$

where (x_s, y_s, z_s) are the coordinates of the sound source, $(0, 0, 0)$ is the coordinate of A1, $(x_2, y_2, 0)$ are the coordinates of A2, and $(x_3, y_3, 0)$ are the coordinates of A3. As shown in Fig. 2(b), we can determine that $x_3 = x_2$ and $y_3 = -y_2$, utilizing the symmetry of the microphone array. After calculating the coordinates of the sound source, the azimuth angle can be determined using the following equation:

$$\theta = \tan^{-1} \left(\frac{y_s}{x_s} \right) \cdot \frac{180}{\pi} \quad (18)$$

where θ represents the current local bearing of MAV 1 with respect to MAV 2 in the horizontal plane. The above summarizes the geometric model we developed, which incorporates microphone array positioning, TDOA, and the distances among MAVs. Ultimately, we estimated the local bearing of the target MAV.

3.4. Nominal bearing estimation under hardware constraints for accuracy evaluation

Our acoustic bearing estimation system operates within the constraints imposed by the limited sampling rate and the sequential switching mechanism of the microphone array. The system samples audio signals at 66.67 kHz, with each microphone requiring approximately 15 μ s for initialization and channel switching. This configuration introduces a quantization effect, as TDOAs, which are naturally continuous, must be represented in discrete sample intervals. Consequently, the bearing estimation process is inherently limited by the system's temporal resolution.

To evaluate the expected performance of the system under ideal conditions, we compute a *nominal bearing* estimate for each experimental configuration. This estimate represents the most likely output of the system, assuming perfect signal quality and known geometry, but constrained by the finite resolution of the sampling hardware.

The process begins with the known 3D coordinates of both the target MAV and the receiver MAV's microphones. Using this information, we calculate the precise distance from the target to each microphone in the array. By dividing these distances by the speed of sound, we derive the ground-truth signal arrival times at each microphone, from which the TDOA between pairs (A1–A2 and A1–A3) is obtained.

These TDOA values are then translated into expected sample differences by dividing them by the system's sampling period. Since the result is a floating-point value, it must be rounded to the nearest integer – a necessary discretization that directly influences the final bearing estimate. However, rather than treating all rounded values as equally likely, we exploit a key physical cue: the directional sequence of sound arrival across microphones, as presented in Eq. (2).

Given the spatial layout of the array and the target's location, we can know the order in which each microphone should receive the sound signal. For instance, if the source lies to the right of the array, microphone A3 is expected to detect the signal before A1 and A2. This expected arrival sequence provides a strong prior that guides the selection of the appropriate integer-valued sample difference (Note that, in real-world scenarios, this sequence is automatically inferred from the TDOA signs or peak positions in the FS-GCC output).

Once the rounded sample differences are selected in accordance with the expected arrival order, they are converted back to discrete TDOA values. These are then fed into the bearing estimation model (Eqs. (13)–(18)) to produce the nominal bearing – the most probable bearing value the system should produce under the given geometry and hardware constraints.

This nominal bearing serves as a benchmark reference against which experimental results are evaluated. It incorporates both the quantization imposed by the sampling rate and the deterministic structure of sound propagation, providing a principled ground truth tailored to the limitations of the system.

4. Experiment validation and results

In this section, we carried out experiments under specific environmental conditions to demonstrate that the proposed method reliably provides reasonable measurement results.²

² A video showcasing the proposed framework and MAV flight tests is provided as supplementary material.

Table 1

Experimental hardware components and their specifications.

Components	Specification
Microphones	Grove-sound sensors
MAV Platform	Crazyflie 2.0 with a UWB module
Processor	Arduino NanoESP32
Other	External ADC

4.1. Experiment set-up and algorithm parameters

We conducted experiments in an indoor hall (50 m \times 40 m \times 15 m) to minimize the effect of echoes and noise. The environment remained relatively quiet, apart from subtle noises like the footsteps of experimenters. A detailed list of the hardware components used in the experiments is presented in Table 1.

For the microphone array arrangement, the three microphones were fixed in an equilateral triangle with a side length of 0.15 m. The coordinates are as follows: A1 (0, 0, 0), A2 ($-0.15\sqrt{3}/2$, 0.15/2, 0), and A3 ($-0.15\sqrt{3}/2$, $-0.15/2$, 0). The equipment setup is shown in Fig. 1.

To evaluate the performance of our bearing estimation method, we designed two experimental scenarios: (i) static-bearing and (ii) dynamic-bearing estimation. In the static-angle estimation experiment, the target MAV (MAV 1) took off, hovered, and landed around predefined angular positions relative to the receiver MAV (MAV 2). These angular positions spanned from -90° to 90° , specifically including 0° , $\pm 30^\circ$, $\pm 45^\circ$, $\pm 60^\circ$, and $\pm 90^\circ$, as illustrated by the red landing pad markers in Fig. 4. The experiments were conducted at two fixed distances between MAV 1 and MAV 2: 1 m and 1.5 m.

In the dynamic-angle estimation experiment, MAV 1 continuously flew across these angular ranges along a straight-line trajectory. Unlike the static-bearing setup, where MAV 1 hovers at fixed angular positions, this experiment introduces additional complexity due to the continuously changing relative angle and distance between the two MAVs. Such a scenario better reflects real-world conditions, where MAVs move dynamically and do not maintain constant separation. To assess the robustness of our system under motion, MAV 1 was flown at three different horizontal velocities: 0.14 m/s, 0.24 m/s, and 0.28 m/s. These velocities were selected based on typical flight performance, with 0.24 m/s as the nominal speed. The lowest and highest values correspond to the minimum and maximum speeds at which the MAV could maintain stable flight under our experimental conditions.

To effectively analyze the signal characteristics, we apply a Hann window [28], defined as:

$$w(p) = 0.5 \left(1 - \cos \left(\frac{2\pi p}{N-1} \right) \right), \quad 0 \leq p < N \quad (19)$$

where the frame length is set to $N = 128$. To ensure smooth transition between adjacent frames, a 75% overlap is applied, leading to a hop size h :

$$h = (1 - 0.75)N = \frac{N}{4} = 32 \quad (20)$$

Subsequently, we perform the Fast Fourier Transform (FFT) with a transform length B given by $B = N = 128$. Since the FFT output is symmetric, only the positive-frequency components are retained, resulting in a single-sided spectrum with $N_f = B/2 = 64$. To enhance frequency resolution and reduce dimensionality, the single-sided spectrum is divided into $L = 13$ sub-bands. Finally, in the Singular Value Decomposition (SVD) process, we retain the first $M = 5$ singular values for weighting, optimizing noise suppression while preserving essential spectral features. This configuration ensures a balance between computational efficiency and spectral resolution, making it suitable for analyzing rapidly fluctuating signals.

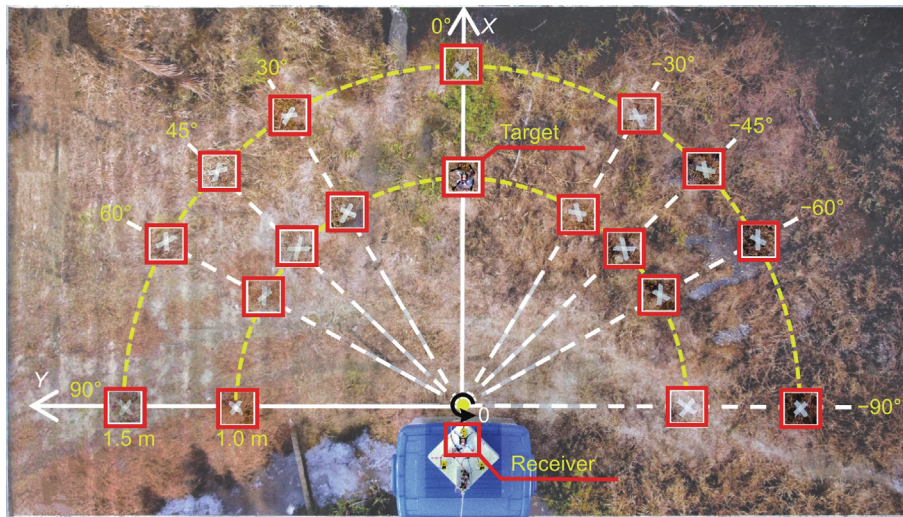


Fig. 4. Configuration of the experimental setup across various local bearings and distances. The receiver MAV is positioned at the coordinate origin, equipped with a microphone array and UWB module. Multiple landing pads, marked by red boxes, represent the target MAV's positions at various relative bearings and fixed distances from the receiver.

4.2. Results and discussion

In this section, to streamline the presentation, we first demonstrate the TDOA estimation outcome using a representative configuration: MAV 1 positioned 1 m away from MAV 2 at a local bearing of 0° . This case illustrates the typical signal behavior and FS-GCC processing observed across other angular setups and distances. Due to the similar pattern of TDOA estimation across configurations, we omit redundant plots for other individual angles. Instead, we aggregate and present the full bearing estimation results, including all static and dynamic configurations, in the final Sections 4.2.4 and 4.2.5, respectively. This collective comparison allows us to evaluate the overall performance and accuracy of the proposed method under diverse conditions.

4.2.1. Sound signals acquisition

Fig. 5 presents the collected sound signals emitted from MAV 1, illustrating distinct characteristics in both amplitude and signal density with the passage of time. We implemented a real-time operating system on a dual-core processor to improve sampling quality and distribute the workload efficiently, assigning Core 1 to handle sampling and Core 2 to manage data logging. To achieve higher-quality analog signal acquisition, an external ADC is required to support the processor. The red dashed box indicates significant amplitude fluctuations in the acoustic signal during MAV 1 takeoff. It is also observed that the signal density in this segment is the highest. This results from the rapid rotation of motors and propellers required for lift. The subsequent segment of the sound signal, recorded during MAV 1 ascent to 0.5 m height and hovering. Portions of the sound signal preceding the red dashed box originates from the measurement noise. This portion has significantly lower amplitude and signal density compared to the later segments.

4.2.2. Results of TDOA estimation using FS-GCC and its comparison with other classical and machine learning approaches

Fig. 6 shows the TDOA in samples between microphone pairs A1-A2 and A1-A3 for one of the frames (13 sub-bands) calculated from the sound signals during MAV 1 flight. We observed that in calculating the sample differences across various frequency subbands L , the maximum GCC $\hat{\phi}[p]$ is obtained at time delay s , as described in Eq. ((12)). For microphone pairs A1-A2 and A1-A3, different local bearings yield different values. The current result

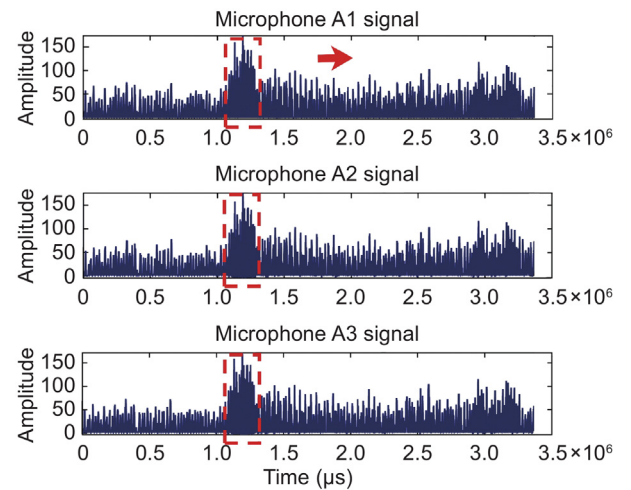


Fig. 5. Amplitude and density characteristics of sound signal samples of MAV 1 positioned 1 m away at the local bearing of 0° .

corresponds to the angle of 0° . In the FS-GCC matrix, yellow regions indicate high correlation, while blue regions indicate low correlation. It can be observed that the distribution of yellow and blue is not uniform across all sub-bands.

Fig. 7 presents the FS-GCC analysis of the entire MAV 1 sound signal segment between A1-A2 and A1-A3, displaying the results for each frame. The typical value shown in the figure corresponds to the maximum of the GCC $\hat{\phi}[p]$. A total of over 1000 frames are shown. This figure provides a visual representation, showing that the majority of the calculated results are accurate, as indicated by data points close to the ground-truth lines (i.e., -9 and -10 for pairs A1-A2 and A1-A3, respectively) derived from the known target geometry in the figure. However, there are several outliers, primarily resulting from the influence of noise. Filters will be utilized to process the data, aiming to extract and summarize the distinct features of the detected samples rather than relying on predefined threshold settings. Such features of the data are intended to be retained, while others will be treated as outliers.

Building upon this analysis, we compare the FS-GCC method against two other TDOA estimation techniques: the classical Generalized Cross-Correlation with Phase Transform (GCC-PHAT)

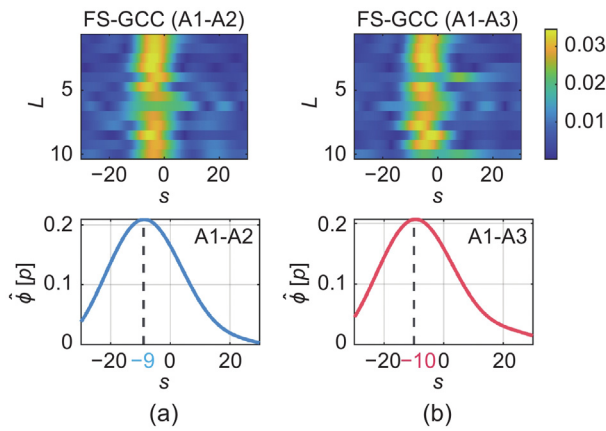


Fig. 6. FS-GCC calculation for one frame of data between microphone pairs (a) A1-A2 and (b) A1-A3 under the angle of 0° , 1 m configuration. L is the number of sub-bands, and $\hat{\phi}[p]$ is the GCC value at TDOA in samples s for sub-band L . The dashed line represents the typical TDOA estimation value.

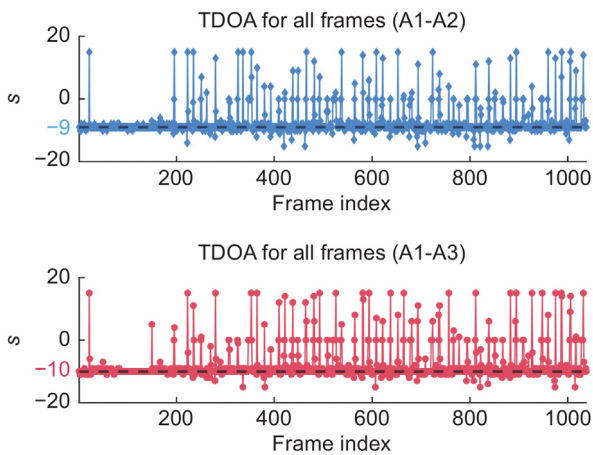


Fig. 7. TDOA (in samples) results for all frames are obtained through the FS-GCC method for microphone pairs A1-A2 and A1-A3 under the angle of 0° , 1 m configuration.

method and a data-driven approach using a LSTM neural network. GCC-PHAT is a widely used baseline in acoustic localization, known for its robustness to reverberation. It computes TDOA by identifying the peak in the cross-correlation of two normalized signal spectra. LSTM-based approach, on the other hand, leverages temporal dependencies in audio sequences to predict TDOA directly. We trained the LSTM neural network on a subset of our collected dataset comprising labeled audio sequences from various target MAV positions. These two methods were evaluated using the same dataset as the one used by the FS-GCC method for fair comparison.

Fig. 8 presents the TDOA estimates from GCC-PHAT and LSTM across multiple frames for the same microphone pairs (A1-A2 and A1-A3). Although GCC-PHAT and LSTM show general alignment with expected values, they also exhibit higher variability and more frequent deviations from the nominal estimate. In contrast, FS-GCC consistently produces tighter clusters around the typical TDOA value.

To quantify this, we compute the RMSE of each method against the nominal TDOA derived from ground-truth geometry. The results, shown in **Table 2**, validate that FS-GCC achieves the lowest RMSE across all pairs, indicating superior accuracy and robustness. These findings validate our choice of FS-GCC as the preferred method for reliable TDOA estimation in this system.

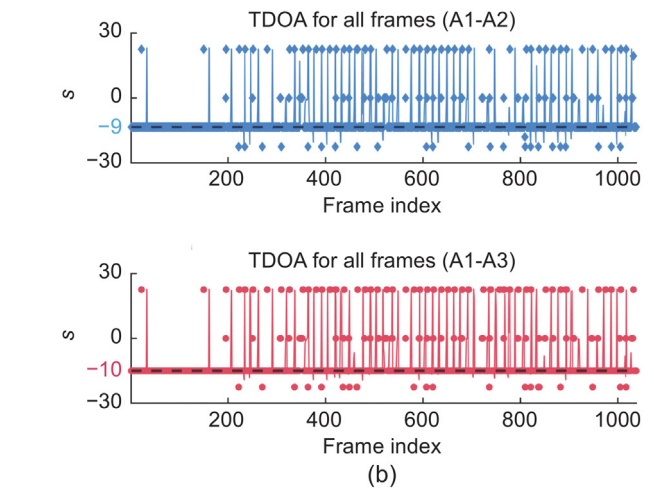
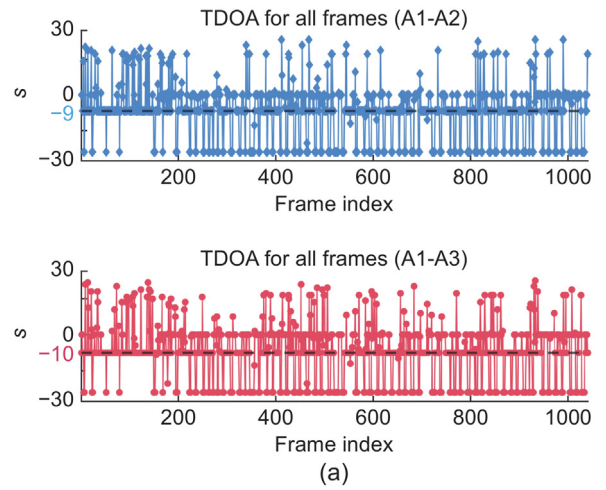


Fig. 8. (a) TDOA (in samples) estimated using the GCC-PHAT method. (b) TDOA (in samples) predicted by the LSTM neural network.

Table 2

Comparison of RMSE in sample units for TDOA estimation using a classical method (GCC-PHAT [27]), a machine learning approach (LSTM neural network [31]), and the method used in our work (FS-GCC [28]) across microphone pairs.

Estimation method	Microphone pair	RMSE (samples)
GCC-PHAT	A1-A2	14.6073
GCC-PHAT	A1-A3	14.7489
LSTM	A1-A2	5.8208
LSTM	A1-A3	6.1751
FS-GCC (<i>This work</i>)	A1-A2	5.0180
FS-GCC (<i>This work</i>)	A1-A3	5.4972

4.2.3. Filter box and filtered data

By combining a median filter with a low-pass filter, extreme outliers (such as sudden noise spikes) are effectively removed while smoothing TDOA estimates and reducing high-frequency jitter. A median filter with a window size of 5 effectively mitigates extreme outliers in TDOA data resulting from noise, interference, or errors. It preserves critical edge features without introducing smoothing, making it a robust choice for maintaining data integrity in signal processing. A fourth-order butterworth low-pass filter with a cutoff frequency of 0.05 effectively smooths TDOA data by reducing high-frequency noise without introducing oscillations. The sampling rate ensures that the filter's reference frequency adapts to the frame shift (hop) size, aligning the design with the current time scale of the data. Additionally, zero-phase

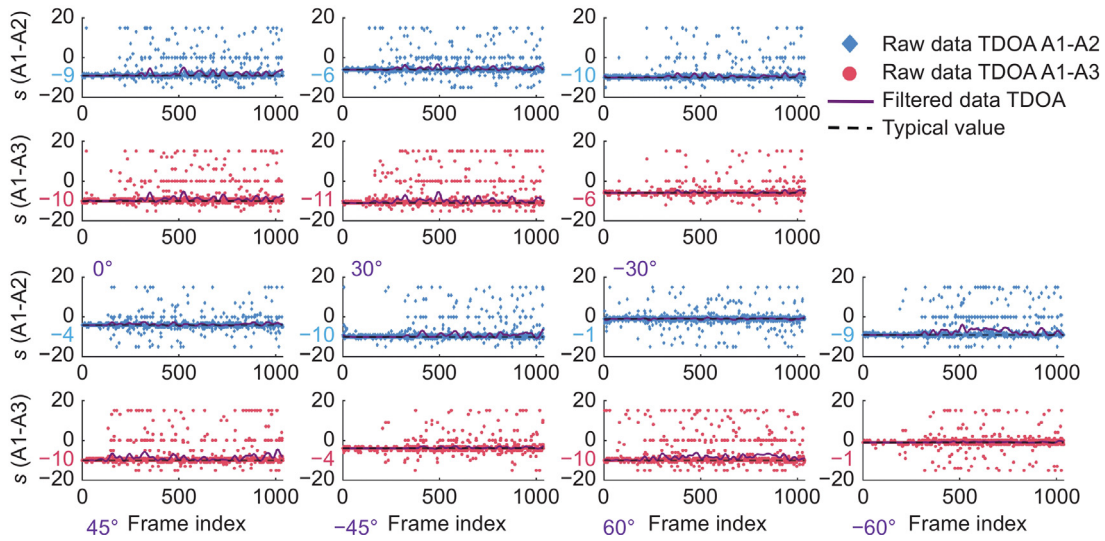


Fig. 9. Results of raw and filtered TDOA (in samples) data for 1 m configurations with various local bearings, after processing through the FS-GCC method.

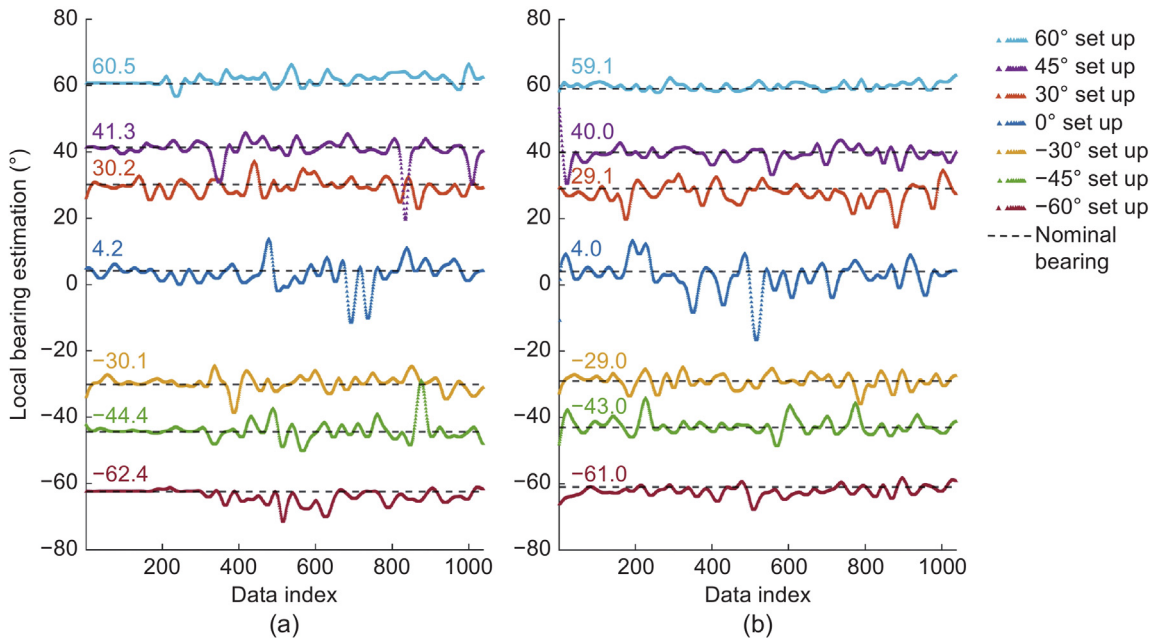


Fig. 10. Results of static-bearing estimation. Estimated local bearings at predefined angular positions (0° , $\pm 30^\circ$, $\pm 45^\circ$, $\pm 60^\circ$) and captured at a distance of (a) 1 m and (b) 1.5 m relative to the target. The horizontal dashed line represents the nominal (ground-truth) bearing calculated based on known MAV position, with the estimated bearings closely tracking it across all angular positions.

filtering maintains time alignment and prevents phase distortion, ensuring signal integrity.

After being processed through the filter box, we found that the filtered data is well-distributed around the line of typical value, as illustrated in Fig. 9. This figure features seven subplots, comparing the data before and after filtering at angles of 0° , $\pm 30^\circ$, $\pm 45^\circ$ and $\pm 60^\circ$ for MAV 1 at a distance of 1 m in the experimental setup. In the initial experimental setup, we arranged two distances of 1 m and 1.5 m. However, after analyzing the corresponding signals, we observed a phenomenon: as the distance increased to 1.5 m, the amplitude of the collected MAV sound was relatively smaller compared to that at 1 m. This is due to the relatively small size of the motors and propellers of the MAV ($9.5 \text{ cm} \times 9.5 \text{ cm}$) used in the experiment, which results in lower sound production. Consequently, the sound attenuation increases with distance. Although MAVs produce sound at lower volumes, they have the advantage

of being less affected by sound propagation phenomena, such as reflection, diffraction, and interference. Therefore, the experimental setup at a distance of 1 m better reflects the effectiveness of our method.

4.2.4. Performance at various static bearings

Fig. 10 presents the local bearing estimation results for the target MAV across multiple angular configurations. Each solid line reflects the output of the TDOA-based bearing estimation pipeline, as described in Section 3. To evaluate performance, we plot a nominal (dashed) line reflecting the most probable estimate derived from the known arrival sequence of sound across the microphones, established in Section 3.4.

As can be observed from Fig. 10, the FS-GCC outputs align closely with the nominal line, which indicates that the system reliably detects the correct sound arrival sequence among microphones. Additionally, the method provides a reliable directional

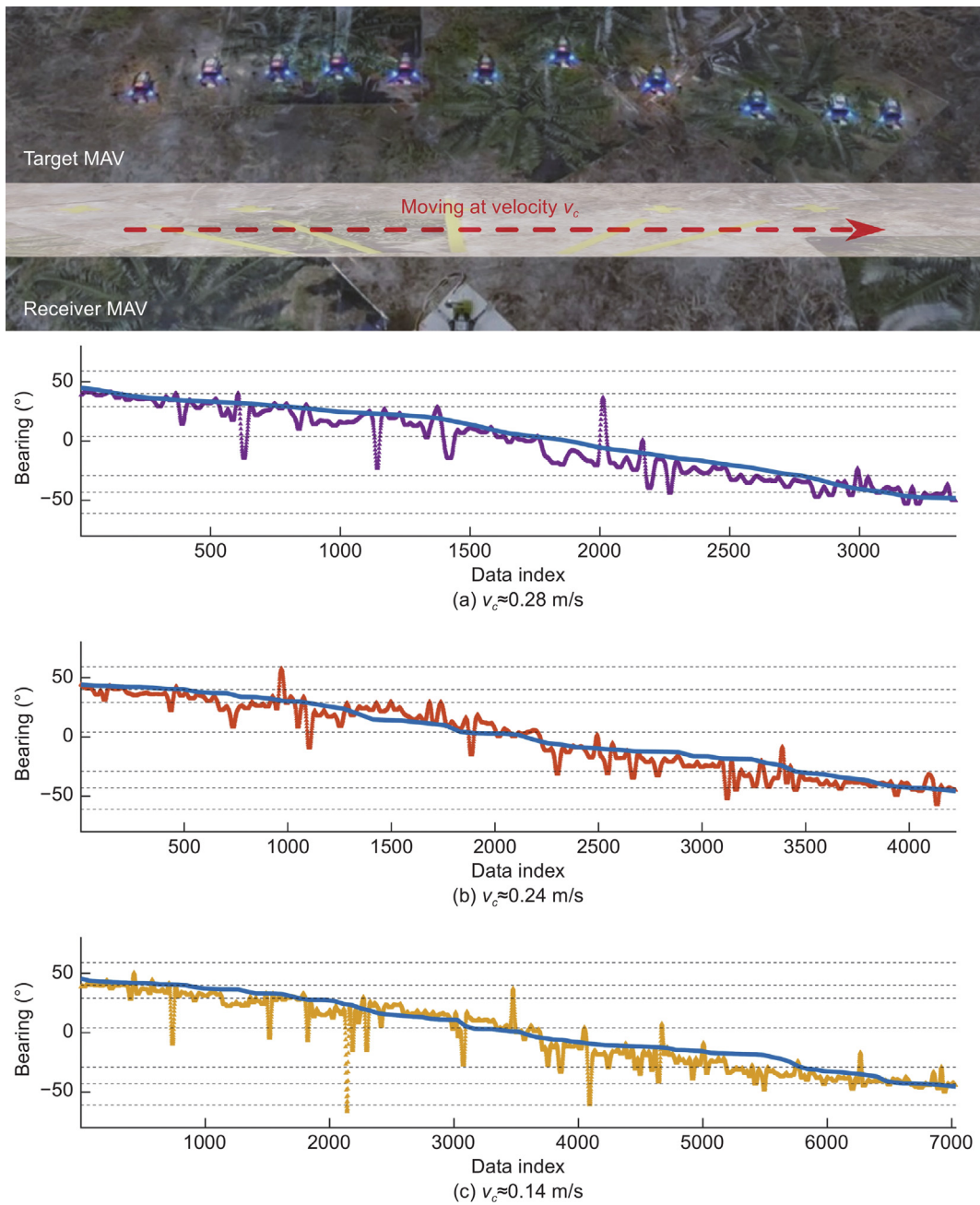


Fig. 11. Results of dynamic-bearing estimation. Estimated local bearings during continuous flights across various angular ranges in a straight-line trajectory at commanded velocities of (a) $v_c \approx 0.28$ m/s, (b) $v_c \approx 0.24$ m/s, and (c) $v_c \approx 0.14$ m/s. The blue line represents the nominal (ground-truth) bearing computed from known MAV positions, which the estimated bearings closely follow across all velocity settings.

estimate of the target MAV. Even in cases where the estimates slightly deviate from the true angle, they still fall within the correct directional regime—that is, the receiver MAV can consistently infer whether the target MAV is positioned in the left, right, or center bearing zone. This level of granularity is particularly useful for enabling coarse but effective decision-making in swarm navigation and coordination tasks under visually degraded conditions.

Notably, the $\pm 90^\circ$ configurations are excluded from the figure, as the received signals in those cases exhibited severe distortion due to the structural design of the microphones. The cylindrical acoustic port caused lateral wave reflections, degrading signal quality and making reliable TDOA estimation infeasible.

4.2.5. Performance at various dynamic bearings

Fig. 11 presents the local bearing estimates for the dynamic-bearing experiments, where MAV 1 was commanded to fly along a straight-line trajectory at three different velocity settings (i.e., $v_c \approx 0.28$ m/s, ≈ 0.24 m/s, and ≈ 0.14 m/s). Across all trials, the estimated bearings closely follow the ground-truth trajectories computed from the known positions of MAV 1. Despite minor fluctuations in the estimates, the overall trend aligns well with the true bearing progression, demonstrating the system's capability to track bearing changes in real time.

Importantly, although MAV 1 was flown along a straight line with nearly constant velocity, the relative distance to MAV 2 varied over time. As a result, the local bearing dynamically transitioned from a positive angle to a negative angle, deviating from a perfect linear trend. This curvature is apparent in both the

- [28] M. Cobos, F. Antonacci, L. Comanducci, A. Sarti, Frequency-sliding generalized cross-correlation: A sub-band time delay estimation approach, *IEEE/ACM Trans. Audio Speech Lang. Process.* 28 (2020) 1270–1281.
- [29] G. Strang, et al., *Linear Algebra and Learning from Data*, vol. 4, Wellesley-Cambridge Press Cambridge, 2019.
- [30] W. Zhao, J. Panerati, A.P. Schoellig, Learning-based bias correction for time difference of arrival ultra-wideband localization of resource-constrained mobile robots, *IEEE Robot. Autom. Lett.* 6 (2) (2021) 3639–3646.
- [31] J. Yu, J. Liu, Z. Peng, L. Gan, S. Wan, Localization of impact on CFRP structure based on fiber Bragg gratings and CNN-LSTM-attention, *Opt. Fiber Technol.* 87 (2024) 103943.