2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Asynchronous Microphone Array Calibration using Hybrid TDOA Information

Chengjie Zhang, Jiang Wang, and He Kong

***Abstract*— Asynchronous microphone array calibration is a prerequisite for many audition robot applications. A popular solution to the above calibration problem is the batch form of Simultaneous Localisation and Mapping (SLAM), using the time difference of arrival measurements between two microphones (TDOA-M), and the robot (which serves as a moving sound source during calibration) odometry information. In this paper, we introduce a new form of measurement for microphone array calibration, i.e. the time difference of arrival between adjacent sound events (TDOA-S) with respect to the microphone channels. We propose to use TDOA-S and TDOA-M, called hybrid TDOA, together with odometry measurements for bath SLAM-based calibration of asynchronous microphone arrays. Extensive simulation and real-world experiments show that our method is more independent of microphone number, less sensitive to initial values (when using off-the-shelf algorithms such as Gauss-Newton iterations), and has better calibration accuracy and robustness under various TDOA noises. Simulation results also demonstrate that our method has a lower Cramér-Rao lower bound (CRLB) for microphone parameters. To benefit the community, we open-source our code and data at https://github.com/AISLAB-sustech/Hybrid-TDOA-Calib.**

I. INTRODUCTION

Microphone arrays can equip robots with sound source localization and tracking abilities, etc [1]–[5]. A prerequisite for realizing the above functionalities is to calibrate the array geometric information accurately [6]. A common approach to the above calibration problem is to utilize the time difference of arrival between microphone pairs (TDOA-M) from a series of sound events to estimate both microphone and sound source locations. Earlier methods require the clock synchronization of all microphones [7], [8]. To overcome the limitation, recent studies, including [9]–[11], take into account the initial time offset between microphone channels.

During calibration, one can obtain the relative position measurements between adjacent sound events from the odometer onboard the robot (which acts as a moving sound source) and use them to improve the calibration accuracy. Following the above idea, based on TDOA-M and odometry measurements, an extended Kalman filter-based simultaneous localization and mapping (EKF-SLAM) method is proposed in [12] to estimate microphone positions, time offsets, and sound source positions simultaneously. However, the impact of the other asynchronous factor, i.e. clock drift rate, is not considered. In [13], [14], a batch SLAM-based method (see, e.g., [15]) is presented to jointly estimate microphone locations, time offsets, clock drift rates, and sound event positions. In our recent work, the framework in [13], [14] has also been generalized to multiple microphone arrays [16] and refined by the data compression method [17].

This work was supported by the Science, Technology, and Innovation Commission of Shenzhen Municipality, China, under Grant No. ZDSYS20220330161800001, the Shenzhen Science and Technology Program under Grant No. KQTD20221101093557010, and the National Natural Science Foundation of China under Grant No. 62350055. The authors are with the Shenzhen Key Laboratory of Control Theory and Intelligent Systems, Southern University of Science and Technology, Shenzhen 518055, China. Emails: 12332644@mail.sustech.edu.cn; 12132297@mail.sustech.edu.cn; kongh@sustech.edu.cn

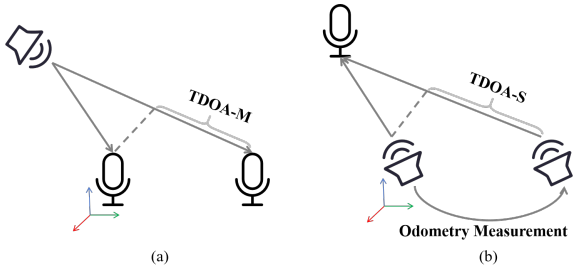


Fig. 1: Differences between TDOA-M (a) and TDOA-S (b).

In this paper, we introduce a new form of measurement for microphone array calibration, i.e. the time difference of arrival between adjacent sound events (TDOA-S) with respect to the microphone channels. Fig. 1 shows the difference between TDOA-M and TDOA-S in the calibration scene. We propose to combine TDOA-S with TDOA-M, called hybrid TDOA, with odometry measurements for microphone array calibration. Our main contributions are stated as follows.

- We introduce TDOA-S and present a simple method to extract TDOA-S from raw audio data. To our best knowledge, this is the first time TDOA-S has been proposed in the literature and used in calibrating robot audition systems. The idea can be useful for other sensing modalities.
- We propose a batch SLAM-based calibration method utilizing hybrid TDOA information and odometry measurements to jointly estimate the asynchronous microphone array parameters (microphone positions, time offsets, clock drift rates) and sound source positions. Extensive simulations and real-world experiments show that compared to [13], our method is more independent of microphone number, less sensitive to initialization, has higher accuracy and robustness under various TDOA noises, and has lower CRLB for microphone parameter estimates.

II. THE PROPOSED METHOD

Assume there are N microphones. Denote the i-th microphone location, time offset, and clock drift rate as $\mathbf{x}_i$, τ_i, and

δ_i respectively. There are K sound events and the j-th sound event location is $\mathbf{s}_j$. Without loss of generality, the coordinate frame is established by sound event positions, called $Sound$ frame, $\mathbf{s}_1 = \mathbf{0}, (\mathbf{s}_2)_y = (\mathbf{s}_2)_z = (\mathbf{s}_3)_z = 0$. The unknown microphone parameters in the $Sound$ frame are

$$\mathbf{x}_{mic} = [\mathbf{x}_1, \delta_1, \mathbf{x}_2, \tau_{2,1}, \delta_2, ..., \mathbf{x}_N, \tau_{N,1}, \delta_N]^T, \quad (1)$$

where $\tau_{i,1} = \tau_i - \tau_1$, $i > 1$. Sound source parameters that need to be estimated are

$$\mathbf{s} = [(\mathbf{s}_2)_x, (\mathbf{s}_3)_x, (\mathbf{s}_3)_y, \mathbf{s}_4, ..., \mathbf{s}_K]^T. \quad (2)$$

A. TDOA-S Derivation and Extraction

1) Derivation: TDOA-S is derived from the time of arrival (TOA) model that considers two asynchronous parameters: time offset and clock drift rate in microphones. In the absence of noise, the arrival time detected by i-th microphone for the j-th sound event, $T_{i,j}$ is shown below

$$T_{i,j} = (1 + \delta_i)(\frac{||\mathbf{x}_i - \mathbf{s}_j||}{c} + \tau_i + t_j), \quad (3)$$

where c is the known sound speed and t_j is the emitting time of j-th sound event. τ_i and δ_i represent the shift and scaling of the temporal frame of i-th microphone with respect to (w.r.t.) the absolute temporal frame, respectively. The former is caused by different startup moments in different microphones. The latter is caused by the sampling rate mismatch between the microphone's actual and absolute sampling rates [18], which can be modeled as a scale constant between a microphone's temporal frame and the absolute temporal frame. If the mismatch does not exist for i-th microphone ($\delta_i = 0$), the TOA model is the same as the common TOA that only considers time offset [6].

In indoor calibration scenarios, the distance between the microphone and sound events does not generally exceed 10 meters. In most cases (Table I in [13]), the clock drift rate and time offset are less than 10^{-4} and 0.1s respectively. Therefore, the term $\delta_i(\frac{||\mathbf{x}_i - \mathbf{s}_j||}{c} + \tau_i)$ is negligible and can be ignored. After this simplification, $T_{i,j}$ becomes

$$\tilde{T}_{i,j} = \frac{||\mathbf{x}_i - \mathbf{s}_j||}{c} + \tau_i + (1 + \delta_i)t_j. \quad (4)$$

Therefore, TDOA-S is expressed as $T^S_{i,j} = \tilde{T}_{i,j+1} - \tilde{T}_{i,j}$. The measurement model of $T^S_{i,j}$ is

$$T^S_{i,j} = \frac{||\mathbf{x}_i - \mathbf{s}_{j+1}|| - ||\mathbf{x}_i - \mathbf{s}_j||}{c} + (1 + \delta_i)\Delta t_j, \quad (5)$$

where $j < K$ and $\Delta t_j = t_{j+1} - t_j$ can be obtained accurately since the speaker installed on the robot is controllable.

2) Extraction: There are two steps in obtaining TDOA-S, each visualized in Fig. 2. Initially, short-time energy [19] is used to obtain the rough left endpoint of each calibration signal in a single-channel microphone. The rough time delay of adjacent calibration signals (T_{rough}) is equal to the difference between the corresponding rough left endpoints (Fig. 2a). Next, each window containing a calibration signal is extracted based on the signal length and rough left endpoint, and we align the adjacent windows and perform GCC-PHAT [20] to obtain the precise delay (T_{pre}) (see Fig. 2b). Finally, one combines the rough delay and precise delay to obtain the overall delay ($T_{rough} + T_{pre}$), which is equal to the difference between two consecutive moments of arrivals.

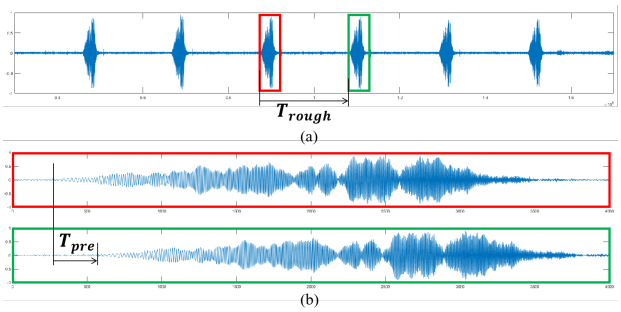


Fig. 2: Visualization of acquiring the rough delay: T_{rough} (a) and precise delay: T_{pre} (b). The red/green box represents the capture window obtaining the current/next recorded calibration signal.

B. Calibration using Hybrid TDOA

1) Hybrid TDOA Measurements: The TDOA-S formulation is derived in (5) and here we also derive the TDOA-M model based on (4). If we select the first microphone as a reference, TDOA-M becomes $T^M_{i,j} = \tilde{T}_{i,j} - \tilde{T}_{1,j}$,

$$T^M_{i,j} = \frac{||\mathbf{x}_i - \mathbf{s}_j|| - ||\mathbf{x}_1 - \mathbf{s}_j||}{c} + \tau_{i,1} + \delta_{i,1}t_j, \quad (6)$$

where $\delta_{i,1} = \delta_i - \delta_1$ ($i > 1$). Assume $t_1 = 0$ without loss of generality, for $j > 1$, $t_j = t_j - t_1 = \sum_{k=2}^{j} \Delta t_{k-1}$. The TDOA-M formula (6) is equivalent to the TDOA formula in [13]. Hence, without noise, the total hybrid TDOA measurements are

$$\mathbf{T}^H = [\mathbf{T}^S, \mathbf{T}^M]^T, \quad (7)$$

where $\mathbf{T}^S = [\mathbf{T}^S_1, \mathbf{T}^S_2, ..., \mathbf{T}^S_N]^T$, $\mathbf{T}^S_i = [T^S_{i,1}, T^S_{i,2}, ..., T^S_{i,K-1}]^T$ and $\mathbf{T}^M = [\mathbf{T}^M_1, \mathbf{T}^M_2, ..., \mathbf{T}^M_K]^T$, $\mathbf{T}^M_j = [T^M_{2,j}, T^M_{3,j}, ..., T^M_{N,j}]^T$.

Considering i.i.d Gaussian noises, the real TDOA-M and TDOA-S measurements are $t^M_{i,j} = T^M_{i,j} + w^M_{i,j}$ ($i > 1$) and $t^S_{i,j} = T^S_{i,j} + w^S_{i,j}$ ($j < K$), respectively, with $w^M_{i,j}, w^S_{i,j} \sim N(0, \sigma^2_{tdoa})$. The real hybrid TDOA measurements are

$$\mathbf{t}^H = [\mathbf{t}^S, \mathbf{t}^M]^T, \quad (8)$$

where $\mathbf{t}^S = [\mathbf{t}^S_1, \mathbf{t}^S_2, ..., \mathbf{t}^S_N]^T$, $\mathbf{t}^S_i = [t^S_{i,1}, t^S_{i,2}, ..., t^S_{i,K-1}]^T$ and $\mathbf{t}^M = [\mathbf{t}^M_1, \mathbf{t}^M_2, ..., \mathbf{t}^M_K]^T$, $\mathbf{t}^M_j = [t^M_{2,j}, t^M_{3,j}, ..., t^M_{N,j}]^T$.

Under Gaussian noise $\mathbf{v}_j \sim N(\mathbf{0}, \sigma^2_{odo}\mathbf{I_3})$, the odometry measurements are $\mathbf{m} = [\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_{K-1}]^T$ with $\mathbf{m}_j$ being defined as follows

$$\mathbf{m}_j = \Delta\mathbf{s}_j + \mathbf{v}_j = \mathbf{s}_{j+1} - \mathbf{s}_j + \mathbf{v}_j, \quad (9)$$

where $j < K$. Denote $\Delta\mathbf{s} = [\Delta\mathbf{s}_1, \Delta\mathbf{s}_2, ..., \Delta\mathbf{s}_{K-1}]^T$.

2) Nonlinear Least Squares solved by Gauss-Newton: From the perspective of batch SLAM, nodes are the locations of a series of sound events (robot pose without orientation) and microphone array (landmark) with positions and asynchronous parameters, while edges are odometry measurements and hybrid TDOA measurements. Note that during the calibration process, any microphone observes every sound event. One can then construct the corresponding nonlinear least squares based on maximum likelihood estimate (MLE) and then use the Gauss-Newton (GN) method to estimate microphone array positions, time offsets, clock drift rates, and the sound event locations simultaneously. Denote the unknown parameters $\mathbf{x} = [\mathbf{x}_{mic}, \mathbf{s}]^T$, measurements $\mathbf{z} = [\mathbf{t}^H, \mathbf{m}]^T$ and measurement function $\mathbf{f}(\mathbf{x}) = [\mathbf{T}^H, \Delta\mathbf{s}]^T$. The minimum of the nonlinear least squares (LS) is shown below

$$\min_{\mathbf{x}} (\mathbf{f}(\mathbf{x}) - \mathbf{z})^T \mathbf{W}^{-1} (\mathbf{f}(\mathbf{x}) - \mathbf{z}), \tag{10}$$

where $\mathbf{W} = diag(\sigma_{tdoa}^2 \mathbf{I}_{N(K-1)+K(N-1)}, \sigma_{odo}^2 \mathbf{I}_{3K-3})$. For solving the above optimization problem, the GN method is usually used. Moreover, for performing source localization tasks after calibration, we need to convert $\mathbf{x}_{mic}$ in $Sound$ frame to $\mathbf{x}_{mic}$ in $Mic.$ frame. The details of the transformation are shown in Appendix A.

C. Computation of CRLB

CRLB is a popular and powerful tool for analyzing parameter estimation errors, as it provides a lower bound on the estimated parameter variance for any unbiased estimator. For nonrandom vector parameters, the CRLB states that the covariance matrix of an unbiased estimator is bounded as follows [21],

$$E[(\hat{\mathbf{x}}(\mathbf{z}) - \mathbf{x}_0)(\hat{\mathbf{x}}(\mathbf{z}) - \mathbf{x}_0)^T] \geq \mathbf{C}, \tag{11}$$

where $\hat{\mathbf{x}}(\mathbf{z})$ is an unbiased estimator of $\mathbf{x}$ given measurement $\mathbf{z}$, $\mathbf{x}_0$ is the true value of vector parameter of $\mathbf{x}$ and $\mathbf{C}$ is the CRLB matrix w.r.t. parameters $\mathbf{x}$. $\mathbf{C} = \mathbf{F}^{-1}$ and $\mathbf{F}$ is the Fisher information matrix,

$$\mathbf{F} = E[[\nabla_{\mathbf{x}} ln\mathbf{L}(\mathbf{x})][\nabla_{\mathbf{x}} ln\mathbf{L}(\mathbf{x})]^T]|_{\mathbf{x}=\mathbf{x}_0}. \tag{12}$$

Furthermore, the Fisher information matrix is shown below,

$$\mathbf{F} = \mathbf{J}^T \mathbf{W}^{-1} \mathbf{J}. \tag{13}$$

In our method, we consider the GN solver for the nonlinear least squares (10) as the unbiased estimator and the CRLB matrix of $\mathbf{x}_{mic}$ in $Sound$ frame, called $\mathbf{x}_{mic}^S$ here, is defined as $\mathbf{C}_{\mathbf{x}_{mic}^S} = \mathbf{C}(1:5N-1, 1:5N-1)$, which is the submatrix of $\mathbf{C}$ w.r.t. $\mathbf{x}$. Then, we need to obtain CRLB for $\mathbf{x}_{mic}$ in $Mic.$ frame, called $\mathbf{x}_{mic}^M$. The affine transformation between $\mathbf{x}_{mic}^M$ and $\mathbf{x}_{mic}^S$ is represented below

$$\mathbf{x}_{mic}^M = \mathbf{A}_S^M \mathbf{x}_{mic}^S + \mathbf{b}_S^M, \tag{14}$$

where the expression of $\mathbf{A}_S^M$ and $\mathbf{b}_S^M$ are shown in Appendix A. According to [22] in Section 3.8, the CRLB matrix of $\mathbf{x}_{mic}^M$, $\mathbf{C}_{\mathbf{x}_{mic}^M}$ is shown below,

$$\mathbf{C}_{\mathbf{x}_{mic}^M} = \mathbf{A}_S^M \mathbf{C}_{\mathbf{x}_{mic}^S} (\mathbf{A}_S^M)^T. \tag{15}$$

In $\mathbf{C}_{\mathbf{x}_{mic}^M}$, we extract diagonal elements corresponding to the CRLB for $\mathbf{x}_{mic}^M$. Then we define an indicator D_{CRLB}

$$D_{CRLB} = \sqrt{\frac{\sum_{i=2}^{N} D_{CRLB_i}}{N-1}}, \tag{16}$$

where D_{CRLB_i} is represented as CRLB for i-th microphone location, offset, or clock drift rate in $Mic.$ frame.

III. SIMULATIONS

We next present simulations to validate the advantages of our method: independence of the number of microphones (Part A), less insensitivity to initial values (Part B), better calibration accuracy and robustness under various TDOA noises (Part C), and lower CRLB for microphone parameters (Part D). For comparative analysis, we use the calibration method [13] using TDOA-M in 3D.

1) Setup: We design three motion trajectories of a sound source. The first one has the space of 3m×3m×3m with 8 sound events (trajectory 1), the second has the space of 2m×6m×2m with 10 sound events (trajectory 2), and the third one possesses the space of 4m×4m×2m with 14 sound events (trajectory 3).

TABLE I: SIMULATION SETTINGS

Setup	Part A	Part B	Part C/D
N	4,6,8,10	6	6
K	8/10/14	8/10/14	8/10/14
True $\mathbf{x}_{mic}$	random		
Initial $\mathbf{x}$	random	σ_{init}	random
σ_{tdoa}	0.1ms	0.1ms	0.05,0.1,0.5ms
σ_{odo}	0.01m		

In "True $\mathbf{x}_{mic}$", "random" means microphone locations are randomly generated in the corresponding trajectory space and $|\tau_{i,1}| \leq 0.1s$, $|\delta_i| \leq 10^{-4}s$. In "Initial $\mathbf{x}$", "random" means both microphone and sound event locations are randomly generated in the corresponding trajectory space, and asynchronous parameters are set to be zero. σ_{init} are standard deviations (SDs) of zero-mean Gaussian noises added into the true positions as the initial values of both microphone and sound event locations. In trajectory 1, $\sigma_{init} = 0m, 1m, 2m, 3m$ and in trajectory both 2 and 3, $\sigma_{init} = 0m, 2m, 4m, 6m$. Simulation under different numbers of microphones (Part A), various initial value noises (Part B), and several TDOA noises (Part C/D) repeat 200 times in each trajectory and the results of three trajectories are combined to analyze.

2) Evaluation Metric: The average root mean square errors of the estimated microphone locations (Loc. err.), time offset (Off. err.), and clock drift rates (Dri. err.) are evaluated in the $Mic.$ frame, whose definition is in Appendix A.

3) Results: It can be observed from Fig. 3a-c that as the number of microphones changes, the calibration performance for microphone parameters of our method remains basically unchanged. However, the performance of [13] shows significant changes and approaches that of our method as the number increases. Fig. 3d-f shows that the estimation performance for microphone parameters of our method remains

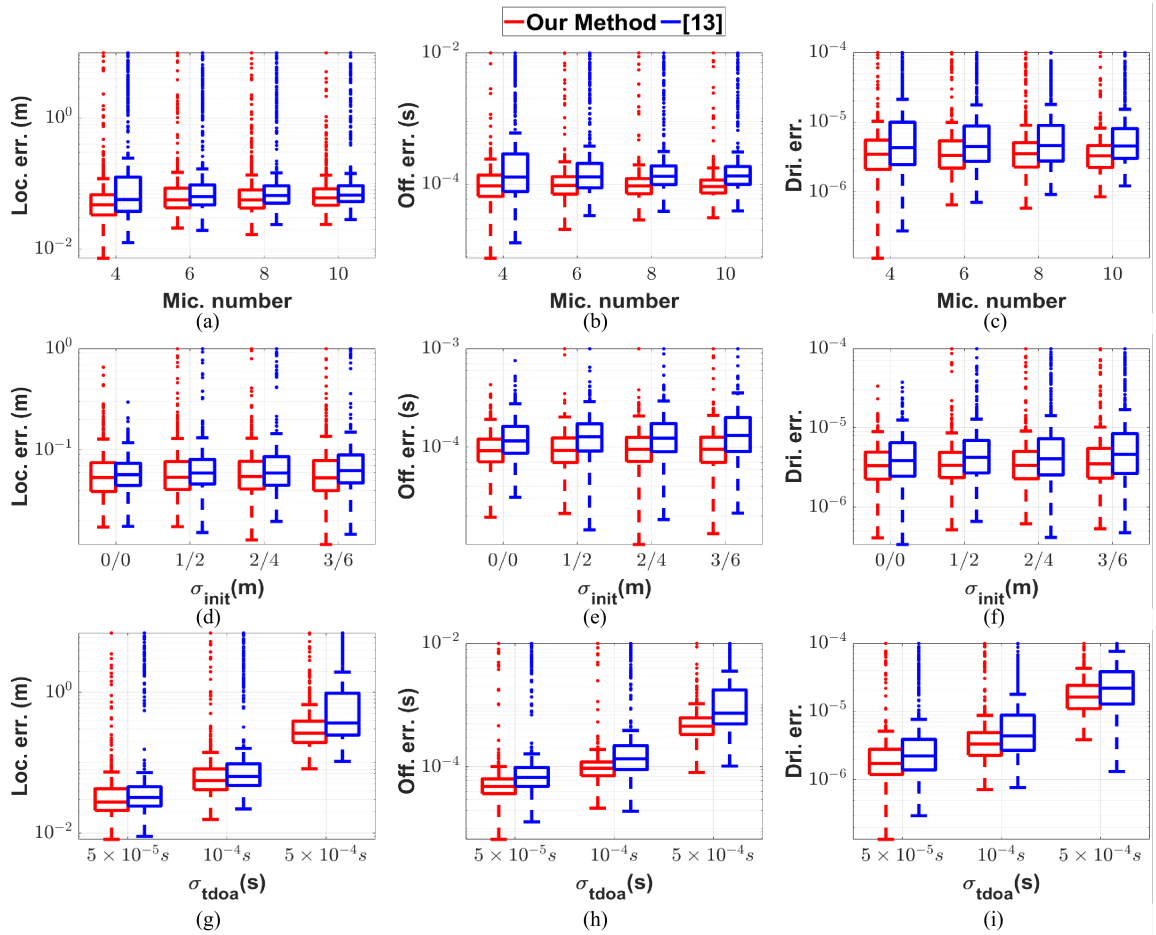


Fig. 3: Box plot of estimation errors for microphone parameters in simulations: microphone locations (a),(d),(g), time offsets (b),(e),(h), and clock drift rates (c),(f),(i) under four microphone numbers, i.e. 4, 6, 8, and 10, four initial values noise SDs under three trajectories, i.e. 0m/0m, 1m/2m, 2m/4m, and 3m/6m, and three TDOA noise SDs, i.e. $5 \times 10^{-5}s$, $10^{-4}s$, and $5 \times 10^{-4}s$, respectively. Here, "xm/ym" means combining the estimation result of trajectory 1 under $\sigma_{init} = x$m and results of both trajectory 2 and 3 under $\sigma_{init} = y$m.

unchanged under different initial value noises. However, microphone parameters estimated by [13] exhibit an increase in estimation error as the initial values noise increases. In Fig. 3g-i, we can observe that our method has better accuracy and robustness in estimating microphone parameters under three levels of TDOA noises, as we have lower median and interquartile range (IQR) values for each box. Table II confirms that our method estimates the CRLB for microphone parameters to be smaller under various TDOA noises.

TABLE II: CRLB results under various TDOA noises (Bold means better)

$\sigma_{tdoa} = 5 \times 10^{-5}s$	Loc. err. (m)	Off. err. (ms)	Dri. err. (10^{-6})
[13]	0.033	0.074	2.765
Our method	**0.027**	**0.055**	**2.191**
$\sigma_{tdoa} = 1 \times 10^{-4}s$	Loc. err. (m)	Off. err. (ms)	Dri. err. (10^{-6})
[13]	0.064	0.140	5.148
Our method	**0.044**	**0.102**	**3.989**
$\sigma_{tdoa} = 5 \times 10^{-4}s$	Loc. err. (m)	Off. err. (ms)	Dri. err. (10^{-6})
[13]	0.310	0.673	24.169
Our method	**0.191**	**0.488**	**18.681**

IV. REAL-WORLD EXPERIMENT

1) Calibration Scenario: The real-world calibration scenario is shown in Fig. 4. The robot (TurtleBot3) carrying a speaker moves around a given plane trajectory whose space is $1.6m \times 2m \times 1m$. When the robot reaches the marked point, the speaker sends out a calibration signal (chirp), and there are 14 sound event locations. On the robot, the speaker is installed on a rotatable pole to change the height of the sound source. Both TDOA-S and TDOA-M are obtained by the GCC-PHAT method [20] and odometry measurements are obtained by an efficient Monocular Visual-Inertial State Estimator (VINS-Mono) [23]. There are three microphone arrays inside the trajectory, each array uses IFLYTEK M160C, a circular array with six microphones.

2) Setup: We randomly set five microphone position configurations and each one is repeated three times. A certain

number of microphones are selected from the three arrays randomly to form a microphone array. The advantage of extracting microphones from multiple arrays to form an array is that it can generate a large amount of real data more conveniently. We conduct similar comparisons with Section III (Part A, Part B, and Part C). The real-world experiment settings are the same as shown in Table I, except that there is only one sound source trajectory with 14 sound events, and TDOA noises of real-world data need to be derived based on the real data. Also, in "True $\mathbf{x}_{mic}$", "random" means microphones are randomly selected from three microphone arrays. σ_{tdoa} is set to $10^{-4}s$.

3) TDOA Noises Evaluation: It's necessary to estimate the noises of TDOA-S and TDOA-M before conducting the real experiment. Because the true values of both microphone and sound locations are known, the estimated noise standard deviation of TDOA-S ($\tilde{\sigma}^S_{tdoa}$) and TDOA-M ($\tilde{\sigma}^M_{tdoa}$) are obtained based on MLE in Appendix B. In Part A and B, to ensure fairness, we select data satisfying $|\tilde{\sigma}^S_{tdoa} - \tilde{\sigma}^M_{tdoa}| < 10^{-5}s$. In Part C, data is divided into five cases with different estimated TDOA noises: $\tilde{\sigma}^S_{tdoa}, \tilde{\sigma}^M_{tdoa} < 10^{-4}s$ (Case A), $10^{-4}s < \tilde{\sigma}^S_{tdoa}, \tilde{\sigma}^M_{tdoa} < 1.5 \times 10^{-4}s$ (Case B), $1.5 \times 10^{-4}s < \tilde{\sigma}^S_{tdoa}, \tilde{\sigma}^M_{tdoa} < 5 \times 10^{-4}s$ (Case C), $|\tilde{\sigma}^S_{tdoa} - \tilde{\sigma}^M_{tdoa}| < 10^{-5}s$ (Case D) and all TDOA-S and TDOA-M without any conditions (Case E).

4) Results: Fig. 5 shows microphone location estimation results in real-world experiments and proves our method performs well and independently of the number of microphones (Fig. 5a), has low sensitivity to initial values (Fig. 5b), and is robust under different TDOA noise levels (Fig. 5c). In Case C of Fig. 5c, although the accuracy of our method is slightly lower than [13] due to the average $\tilde{\sigma}^S_{tdoa}$ is 80μs larger than that of $\tilde{\sigma}^M_{tdoa}$, our method remains much more stable with a smaller IQR.

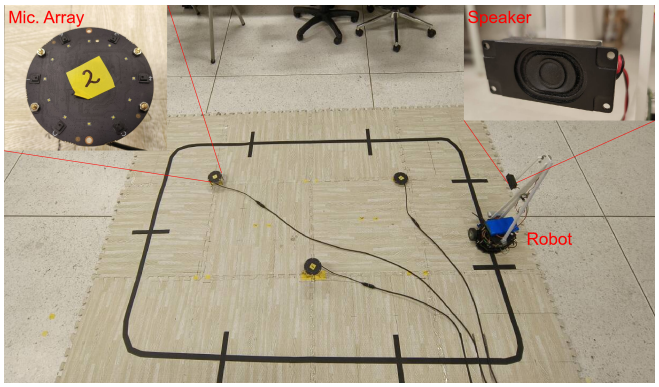


Fig. 4: The calibration scenario for real-world experiments.

V. DISCUSSION

The simulation and experimental results show that our calibration method, incorporating both TDOA-S and TDOA-M measurements, has several advantages compared to the case using only TDOA-M information. A major reason for the above outcome could be that TDOA-S actually has fewer timing parameters than TDOA-M, which helps to reduce the nonlinearity of non-linear LS and sensitivity to initial values. For lower CRLB, from an information theory perspective, compared to [13], we have more measurement information, which results in lower CRLB.

VI. CONCLUSIONS

This paper is concerned with asynchronous microphone array calibration using batch SLAM. More specifically, we have introduced a new type of measurement, i.e. TDOA-S, for the above calibration problem. We have presented a simple procedure to extract TDOA-S measurements from raw audio information. We proposed to use hybrid TDOA information (both TDOA-S and TDOA-M) for calibrating asynchronous microphone arrays. Both simulation and experiment results show that our proposed method leads to improved calibration results compared to the case using only TDOA-M information. In particular, the proposed method is more independent of the number of microphones, has lower sensitivity to initial values, and has higher accuracy and robustness under various TDOA noises. The focus of our current and future work is to explore the relationship between calibration observability and sound source trajectory configuration and generalize the ideas in this paper to the problem of calibrating multiple microphone arrays.

VII. APPENDIX

A. Affine Transformation from $\mathbf{x}^S_{mic}$ to $\mathbf{x}^M_{mic}$

$\mathbf{x}_{mic}$ in $Mic.$ frame which is established by assuming $\mathbf{x}_1 = \mathbf{0}$, $(\mathbf{x}_2)_y = (\mathbf{x}_2)_z = (\mathbf{x}_3)_z = 0$, is defined below:

$$\mathbf{x}_{mic} = [\mathbf{x}_1, \mathbf{x}_2, \tau_{2,1}, \delta_{2,1}, ..., \mathbf{x}_N, \tau_{N,1}, \delta_{N,1}]^T. \tag{17}$$

Given the definitions of the vectors $\mathbf{x}^S_{mic}$ and $\mathbf{x}^M_{mic}$, the details of this linear transformation relationship are as follows

$$\begin{aligned} \mathbf{x}^M_i &= \mathbf{R}\mathbf{x}^S_i + \mathbf{t}, \\ \tau^M_{i,1} &= \tau^S_{i,1}, \\ \delta^M_{i,1} &= \delta^S_i - \delta^S_1. \end{aligned} \tag{18}$$

where $\mathbf{R}$ and $\mathbf{t}$ are the rotation matrix and translation vector respectively and transfer $\mathbf{x}_i$ from $Sound$ frame into $Mic.$ frame. The construction of $\mathbf{A}^M_S$ and $\mathbf{b}^M_S$ are based on (1), (17) and (18).

B. Estimating Standard Deviation of TDOA Noise

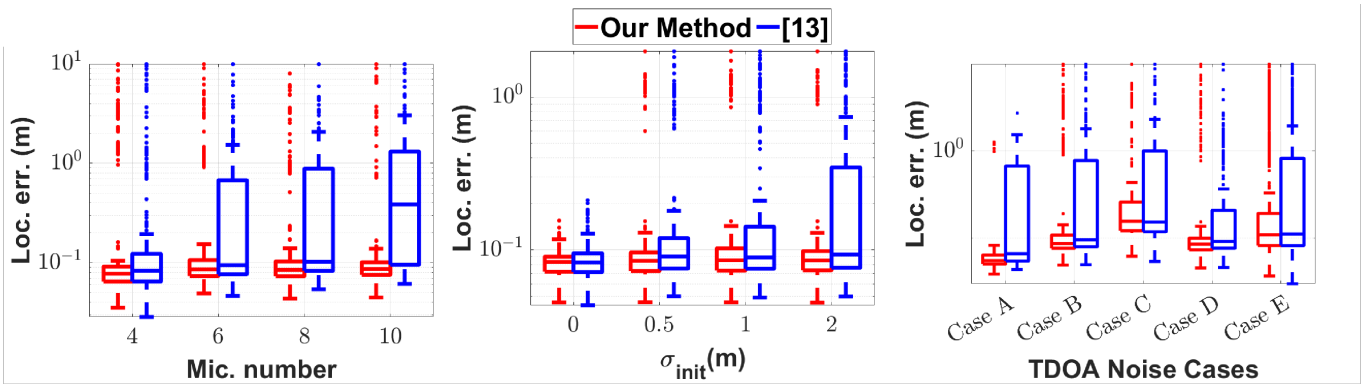
1) Computation of $\tilde{\sigma}^S_{tdoa}$: Given $t^S_{i,j}$, $\mathbf{x}_i$ and $\mathbf{s}_j$, $i = 1, 2, ..., N$ and $j = 1, 2, ..., K-1$. $\tilde{t}^S_{i,j}$ is shown below

$$\begin{aligned} \tilde{t}^S_{i,j} &= t^S_{i,j} - \frac{||\mathbf{x}_i - \mathbf{s}_{j+1}|| - ||\mathbf{x}_i - \mathbf{s}_j||}{c} - \Delta t_j \\ &= \delta_i \Delta t_j + w^S_{i,j}. \end{aligned}$$

Unbiased estimation based on MLE for δ_i is below

$$\min_{\delta_i} \sum_{j=1}^{K-1} (\tilde{t}^S_{i,j} - \delta_i \Delta t_j)^2 \Longrightarrow \hat{\delta}_i = \frac{\sum_{j=1}^{K-1} \tilde{t}^S_{i,j}}{\sum_{j=1}^{K-1} \Delta t_j}.$$

Therefore, $\tilde{w}^S_{i,j} = \tilde{t}^S_{i,j} - \hat{\delta}_i \Delta t_j$. $\tilde{\sigma}^S_{tdoa}$ is estimated unbiased based on $\tilde{w}^S_{i,j}$.



(a) (b) (c)

Fig. 5: Box plot of results in real-world experiment: microphone location estimation errors under various microphone numbers (a), i.e. 4, 6, 8, and 10, initial values noise SDs (b), i.e. 0m, 0.5m, 1m, and 2m, and five cases of TDOA noises (c).

2) Computation of $\tilde{\sigma}^M_{tdoa}$: Given $t^M_{i,j}$, $\mathbf{x}_i$ and $\mathbf{s}_j$, $i = 2, 3, ..., N$ and $j = 1, 2, ..., K$. $\tilde{t}^M_{i,j}$ is shown below

$$\tilde{t}^M_{i,j} = t^M_{i,j} - \frac{||\mathbf{x}_i - \mathbf{s}_j|| - ||\mathbf{x}_1 - \mathbf{s}_j||}{c} = \tau_{i,1} + \delta_{i,1} t_j + w^M_{i,j}.$$

Unbiased estimation based on MLE for $\tau_{i,1}, \delta_{i,1}$ are below

$$\min_{\tau_{i,1},\delta_{i,1}} \sum_{j=1}^{K} (\tilde{t}^M_{i,j} - \tau_{i,1} - \delta_{i,1} t_j)^2 \Longrightarrow \begin{bmatrix} \hat{\tau}_{i,1} \\ \hat{\delta}_{i,1} \end{bmatrix} = (A^T A)^{-1} A^T b,$$

$A = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_K \end{bmatrix}$ and $b = \begin{bmatrix} \tilde{t}^M_{i,1} \\ \tilde{t}^M_{i,2} \\ \vdots \\ \tilde{t}^M_{i,K} \end{bmatrix}$. Therefore, $\tilde{w}^M_{i,j} = \tilde{t}^M_{i,j} - \hat{\tau}_{i,1} - \hat{\delta}_{i,1} t_j$. $\tilde{\sigma}^M_{tdoa}$ is estimated unbiased based on $\tilde{w}^M_{i,j}$.

REFERENCES

[1] C. Rascon and I. Meza, "Localization of sound sources in robotics: A review," *Robotics and Autonomous Systems*, vol. 96, pp. 184–210, 2017.

[2] L. Fu, Y. He, J. Wang, X. Qiao, and H. Kong, "I-asm: Iterative acoustic scene mapping for enhanced robot auditory perception in complex indoor environments," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, accepted and to appear.

[3] M. Strauss, P. Mordel, V. Miguet, and A. Deleforge, "Dregon: Dataset and methods for uav-embedded sound source localization," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 5735–5742.

[4] C. Evers, H. W. Löllmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor, and W. Kellermann, "The locata challenge: Acoustic source localization and tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1620–1643, 2020.

[5] P.-O. Lagacé, F. Ferland, and F. Grondin, "Ego-noise reduction of a mobile robot using noise spatial covariance matrix learning and minimum variance distortionless response," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 3533–3538.

[6] A. Plinge, F. Jacob, R. Haeb-Umbach, and G. A. Fink, "Acoustic microphone geometry calibration: An overview and experimental evaluation of state-of-the-art algorithms," *IEEE Signal Processing Magazine*, vol. 33, no. 4, pp. 14–29, 2016.

[7] M. Crocco, A. Del Bue, and V. Murino, "A bilinear approach to the position self-calibration of multiple sensors," *IEEE Transactions on Signal Processing*, vol. 60, no. 2, pp. 660–673, 2012.

[8] Y. Kuang, S. Burgess, A. Torstensson, and K. Åström, "A complete characterization and solution to the microphone position self-calibration problem," in *2013 IEEE ICASSP*, 2013, pp. 3875–3879.

[9] N. Ono, H. Kohno, N. Ito, and S. Sagayama, "Blind alignment of asynchronously recorded signals for distributed microphone array," in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2009, pp. 161–164.

[10] L. Wang, T.-K. Hon, J. D. Reiss, and A. Cavallaro, "Self-localization of ad-hoc arrays using time difference of arrivals," *IEEE Transactions on Signal Processing*, vol. 64, no. 4, pp. 1018–1033, 2016.

[11] D. E. Badawy, V. Larsson, M. Pollefeys, and I. Dokmanić, "Localizing unsynchronized sensors with unknown sources," *IEEE Transactions on Signal Processing*, vol. 71, pp. 641–654, 2023.

[12] H. Miura, T. Yoshida, K. Nakamura, and K. Nakadai, "Slam-based online calibration of asynchronous microphone array for robot audition," in *2011 IEEE/RSJ IROS*, 2011, pp. 524–529.

[13] D. Su, H. Kong, S. Sukkarieh, and S. Huang, "Necessary and sufficient conditions for observability of slam-based tdoa sensor array calibration and source localization," *IEEE Transactions on Robotics*, vol. 37, no. 5, pp. 1451–1468, 2021.

[14] D. Su, T. Vidal-Calleja, and J. V. Miro, "Simultaneous asynchronous microphone array calibration and sound source localisation," in *2015 IEEE/RSJ IROS*, 2015, pp. 5561–5567.

[15] G. Grisetti, R. Kümmerle, C. Stachniss, and W. Burgard, "A tutorial on graph-based slam," *IEEE Intelligent Transportation Systems Magazine*, vol. 2, no. 4, pp. 31–43, 2010.

[16] J. Wang, Y. He, D. Su, K. Itoyama, K. Nakadai, J. Wu, S. Huang, Y. Li, and H. Kong, "Slam-based joint calibration of multiple asynchronous microphone arrays and sound source localization," *IEEE Transactions on Robotics*, DOI: 10.1109/TRO.2024.3410456, 2024.

[17] X. Li, H. Deng, J. Wang, L. Fu, and H. Kong, "Information-aware joint calibration of microphone array and sound source localization," in *International Conference on Indoor Positioning Indoor Navigation (IPIN)*, 2024, accepted and to appear.

[18] E. Robledo-Arnuncio, T. S. Wada, and B.-H. Juang, "On dealing with sampling rate mismatches in blind source separation and acoustic echo cancellation," in *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007, pp. 34–37.

[19] L. R. Rabiner, R. W. Schafer *et al.*, "Introduction to digital speech processing," *Foundations and Trends® in Signal Processing*, vol. 1, no. 1–2, pp. 1–194, 2007.

[20] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[21] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with applications to tracking and navigation: theory algorithms and software*. John Wiley & Sons, 2004.

[22] S. M. Kay, *Fundamentals of statistical signal processing: estimation theory*. USA: Prentice-Hall, Inc., 1993.

[23] T. Qin, P. Li, and S. Shen, "Vins-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 1004–1020, 2018.