HearLoc: Locating Unknown Sound Sources in 3D with a Small-Sized Microphone Array

Zhaohui Li, *Student Member, IEEE*, Yongmin Zhang, *Senior Member, IEEE*, Lin Cai, *Fellow, IEEE*, Yaoxue Zhang, *Senior Member, IEEE*

***Abstract*—Indoor Sound Source Localization (ISSL) is under growing focus with the rapid development of smart IOT intelligence. The predominant approaches typically involve constructing large microphone (Mic) array systems or extract multiple angles of arrival (AOAs). However, the performance of these solutions is often constrained by the physical size of the array. Besides, there has been limited focus on 3D localization with a single small-sized Mic array. In this paper, we propose HearLoc, an ISSL system that can locate 3D sources with a ten-cm Mic array. We demonstrate that the localization ability and dimensional capability can be significantly enhanced by incorporating the time differences of arrival (TDOAs) between the LOS and ECHO signals from nearby reflective surfaces. Our approach involves a localization method that selectively sums the correlation powers at useful TDOAs induced by each location. We also design a data processing pipeline with interpolation, normalization and pruning techniques to improve system accuracy and efficiency. To further enhance scalability, we design an iterative algorithm for the ISSL problem with multiple sources and an array location calibration scheme. Experiments demonstrate that the HearLoc can effectively locate sound sources, exhibiting $2\times$/$3.7\times$ improvements in accuracy for 2D and 3D localization, respectively, and a $4\times$ increase in efficiency compared to the existing AOA-based ISSL solutions.**

***Index Terms*—Sound source localization, angle-of-arrival, indoor acoustics, multipath**

I. Introduction

With the continuous development of AI technology, intelligent voice assistants have been widely integrated into Intelligent Agents (IA) such as smart speaker and mobile robot [1], [2]. As one of the foundations for human to interact with the IAs, voice commands usually contain abundant features such as semantics and moods, which help IAs better understand the meaning of users. In addition to these feature domains, the growth of the edge computing market has brought growing demand for sound source localization, especially in the context of IOT and HCI [3]–[5]. Various scenarios benefit from this capability, including but not limited to: (I) A smart speaker can accurately identify illegal break-ins or elder falling according to the sound type and location. (II) A sweeping robot hears a command "clean here" and can navigate to the exact location where the user (namely the sound source) stands for cleaning tasks. Compared with other localization technologies that use radio frequency or visual signals, localization directly by sound usually has advantages of less energy consumption, wider field-of-view and less privacy concerns [6]–[10].

Previous sound source localization technologies typically rely on deployments of meter-level Mic arrays [11]–[16]. Along with the miniaturization trend of IAs, obtaining the exact location of an unknown sound source with a small-sized Mic array (usually ten-cm level) is important yet. This is because when the source-array distance is several times larger than the physical size of array, the sound rays that arrive at different Mics can be considered parallel, making it only possible to determine the AOA of source, rather than its precise location. This phenomenon is also called far-field effect [17]. Due to the existence of multipath in indoor environments, the Mic array can usually receive signals from various directions. Prior studies have explored calculating multiple AOAs that sound arrives at the array. Then, the source location can be determined by the intersection of these AOA rays (i.e. triangulation method) [18]–[21]. Although significant progress has been achieved, there are still several unsolved challenges:

- **Locating Sound Sources in 3D**: Localization in 3D typically relies on constructing a 2D or 3D Mic array of meter level [22], [23]. Recent research on using Mic arrays of ten-cm for source localization is mostly limited in 2D [18]–[21]. However, in real-life situations, the source would not always be at a known 2D plane, and the deviation in height dimension can affect the localization performance with a small-sized array [21].
- **Limited Resolution**: According to the array theory [24], the resolution of AOA estimation is constrained by the physical size of array, whereas the limited acoustic sampling rate on commodity IAs also creates significant ambiguities in AOA estimation. In these cases, previous AOA-based solutions may fail to work when the obtained AOA rays are merged or approximately parallel.
- **High Computation Complexity**: To obtain more precise AOAs, the processing latency of many existing ISSL systems is usually at second-level [18]–[20]. This may not meet the requirements for many real-time applications.

To tackle these challenges, we propose HearLoc, an ISSL system that can locate 3D sound sources with only a small-sized Mic array. We simulate the collection of room ECHO signals by virtual Mic elements. As a result, multiple real

Part of this work was previously accepted by 2023 IEEE/CIC International Conference on Communications in China (ICCC).

This work has been supported in part by the National Natural Science Foundation of China under Grant No. 62172445 and 62341201, by the Young Talents Plan of Hunan Province, and Major Project of Natural Science Foundation of Hunan Province under Grant No. 2021JC0004. (Corresponding author: Yongmin Zhang.)

Z. Li and Y. Zhang are with the School of Computer Science and Engineering, Central South University, Changsha, Hunan, 410012, China. Emails:{lizhaohui,zhangyongmin}@csu.edu.cn;

L. Cai is with the Department of Electrical and Computer Engineering, University of Victoria, Victoria, V8W 3P6, Canada. E-mail: cai@ece.uvic.ca;

Y. Zhang is with the Department of Computer Science and Technology, BNRist, Tsinghua University, Beijing, 100084, China, and also with Zhongguancun Laboratory, Beijing, 100094, China. E-mail: zhangyx@tsinghua.edu.cn.

and virtual Mics can construct a cross-wall 2D Mic array, whose size is significantly larger than the physical size of the original array. This equivalent virtual array can enable 3D localization and improve the accuracy in 2D. Technically, to handle challenge 1, we build a model of multipath propagation to show the effectiveness of utilizing several useful TDOAs among multipath signals for 3D localization. As for challenge 2, we design an ISSL system that can generate localization heatmaps by searching for the location with the maximal sum of correlation powers. In order to improve the resolution, we propose a spectrum interpolation approach on general cross correlation phase transformation (GCC-PHAT) and a matrix normalization scheme that can address ambiguities in the localization heatmap. To solve challenge 3, a pruning algorithm is adopted by locating some strong correlation peaks along the strong LOS signals for accelerating. Besides, our algorithm has a low complexity, and parts of it can be designed to work offline, further reducing processing time.

Finally, to build a practical system, we first propose to solve the cases where multiple sound sources exist by an iterative algorithm. Its core is to set the correlation power at TDOAs induced by each source location to zero iteratively. Then, we propose a fine-grained array location calibration scheme by building an optimization function for the theoretical and obtained time-of-flight (TOF) information of reflection signals.

The results in datasets created by PyroomAcoustics [25] demonstrate an overall median error of $0.2m$ and $0.37m$ in 2D and 3D. The corresponding processing latency is $0.19s$ and $0.2s$ on a low-end PC. We also conduct experiments in real rooms for localization and tracking. The median error is less than $0.44m$. Our contributions can be summarized as:

- To the best of our knowledge, we make the first attempt to directly locate 3D sound sources with a small-sized microphone array, even using a simple 2-Mic system. Our source code has been released on Github[1].
- We propose a novel ISSL localization framework, mainly achieved by selectively summing the correlation powers at several useful TDOAs generated by each location. We also develop schemes to address the problems with multiple sources and array location calibration.
- Extensive experiments demonstrate a $3.7\times$ increased accuracy in 3D localization and $4\times$ improved efficiency of our scheme over the SOTA AOA-based solutions.

The rest of this paper is organized as: §II presents the related works. §III outlines the model of ISSL problem with a small-sized array. §IV describes the system architecture and §V presents the ISSL system design for a single source. §VI and §VII discuss the solutions of multiple sources localization and array location calibration. §VIII and §IX present the experimental results in simulation and real world. §X gives the discussions and §XI makes a conclusion.

II. Related Work

A. Indoor Localization

The current most widely-adopted localization system is GPS, a satellite-based navigation system that can provide precise location information. Although working excellently outdoors, its performance may significantly decrease when used indoors because of large signal attenuation and complex multipath [26]. To overcome this limitation, there are works that propose to utilize Wi-Fi, UWB and RF signals for indoor localization [6], [9], [10], achieving an accuracy of dm level. For example, it is proven feasible to construct a received-signal-strength (RSS) map collected by multiple access points to achieve localization [6]. However, these RF-based methods usually require specific equipment, large-scale deployments or extensive data collection, which may be expensive and inconvenient for daily use. The utilization of cameras for localization purposes is possible, but it may face limited acceptance at homes due to privacy concerns [7], [8].

Different from these solutions, our system can be directly deployed with a single small-sized Mic array, which can be very convenient and low-cost. Besides, our method does not need a dataset, thereby saving the effort of data collection.

B. Acoustic Source Localization

In recent years, acoustic signal has been widely used for fine-grained ranging and tracking thanks to its low propagation speed. Some works have utilized different types of signals such as sinusoidal, FMCW, OFDM [27]–[29] or pseudo white noise [30] to achieve movement tracking at cm or even mm level. Besides these works that transmit and collect known acoustic signals, unknown sounds, such as human voices, are also prevalent in our daily life. Conventional techniques for unknown sound source localization typically rely on building a large-scaled Mic array or multiple small-sized Mic arrays [11], [12]. TOF information, TDOA between Mic elements and acoustic fingerprints have also been widely utilized in 2D/3D localization [16], [22], [23]. Although effective, in order to ensure acceptable spatial resolution, the total physical array size of these systems is usually in meters. This requirement may not be feasible for space-limited deployments. Furthermore, some large-scaled Mic systems may require additional hardware for precise synchronization [15]. This may also bring additional energy consumption and high cost. To the best of our knowledge, we are the first to study source localization with a single small-sized Mic array of ten-cm level in the scene of 3D localization.

Another common solution for unknown source localization is to analyze indoor multipath propagation and extract multiple AOAs that sound arrives at a small-sized array. Several conventional AOA estimation methods, such as delay-and-sum and MUSIC, may be vulnerable to coherent signals in indoor environments [31]. To cope with these limitations, VoLoc [18] proposes an iterative-alignment-cancellation algorithm to extract AOAs for the LOS and ECHO paths of voice. Symphony [19] proposes a localization algorithm by searching for AOAs according to the linear relationship between correlation peaks. Furthermore, MAVL [20] proposes a multi-resolution algorithm that extends the sound localization model to non-line-of-sight scenarios. Although commendable, many of these localization systems rely on precise estimation of AOA, whose accuracy is mainly constrained by the physical

[1]https://github.com/Lizhaohui2000/HearLoc

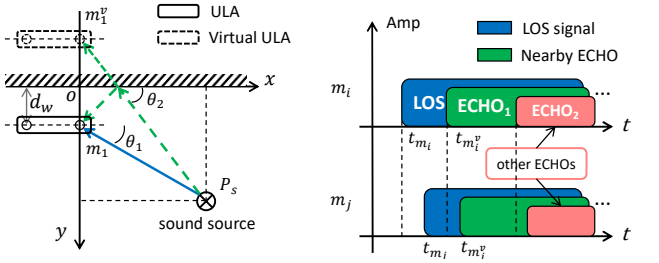


Fig. 1. (a) An illustration of the ISSL problem in 2D. The blue and green dotted line are the LOS and ECHO path, respectively. (b) The received signals in time domain from two Mics.

size of Mic array [24]. The accurate search of AOA may be extremely time-consuming. Instead, this work directly addresses the localization problem in delay domain, achieving a higher efficiency and accuracy.

III. System Model

A. Sound Propagation Indoors

We consider an indoor environment, where a small-sized Mic array with M Mics is deployed for sound collection. Let m_i, $i \in [1, M]$ denote the i-th Mic. Supposing that there are a total of K paths that sounds arrive at the array, the received signal from the i-th Mic can be described as:

$$y_i(t) = \sum_{k=1}^{K} \alpha_k a_i(\theta_k, \phi_k) s(t - t_k) + n_i(t), \tag{1}$$

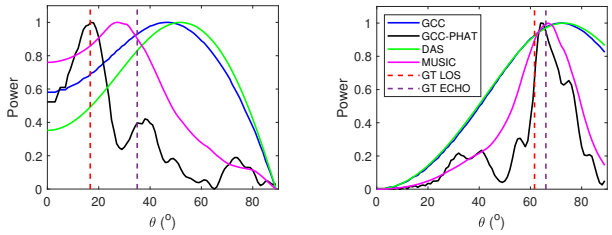
where $s(t)$ is the waveform of source signal, α_k, t_k, θ_k, ϕ_k are the signal strength, TOF, azimuth and elevation angle of the k-th ($k \in [1, K]$) path, respectively, $a_i(\cdot)$ denotes the array steering value of m_i, and $n_i(t)$ is the Gaussian white noise term. Due to the fact that the size of Mic array on most IAs is usually ten cm-level, the sound propagation follows the far-field effect assumption [17]. That is to say, the sound ray that arrives at each Mic can be viewed as parallel. We treat m_1 as the reference Mic, and the array steering value for the k-th path and the i-th Mic of a Unit Linear Array (ULA) can be presented by:

$$a_i(\theta_k, \phi_k) = e^{-j2\pi \frac{d_a(i-1)}{\lambda} \cos(\theta_k)\cos(\phi_k)}, \tag{2}$$

where d_a is the element spacing of Mics, and λ is the wavelength of sound. The objective of addressing the ISSL problem is to obtain the optimal estimation for the location of sound source according to the array received signals $y(t)$.

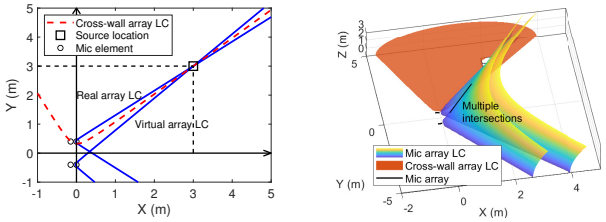
B. Localization with Nearby Reflection

To enable the function of our system, we make one core assumption: **The location of Mic array is close to at least one reflective surface**, such as a wall. For simplification, the reflective surface will be referred to as "wall" in the following text. Signals from two paths can be assumed significant: the LOS signal and ECHO reflected by the nearby wall, as shown by the blue and green dashed lines in Fig. 1(a), respectively. A virtual ULA can be generated behind the wall to simulate the receiving of ECHO signals. As shown by Fig. 1(a), we build a Cartesian coordinate system in the room space. Mic m_1 is set



(a) Large AOA difference. (b) Small AOA difference.

Fig. 2. AOA spectrum obtained through several AOA estimation algorithms in indoor environments.



(a) 2D localization. (b) 3D localization.

Fig. 3. Hyperbola and hyperboloid for 2D and 3D localization. LC is the an abbreviation for localization curve. For a better view, the hyperbola/hyperboloid induced by equation ④ in Eq. (4) is omitted.

at $P_{m_1} = (x_a, d_w, h_a)$, where x_a, h_a is the x coordinate and height of array in the room, and d_w is the wall-array distance. Correspondingly, the location of virtual Mic m_1^v can be given by $P_{m_1^v} = (x_a, -d_w, h_a)$. Let $P_s = (x_s, y_s, z_s)$ denote the coordinates of sound source.

A recently widely adopted approach to address the ISSL problem is the triangulation method [18]–[21]. It involves extracting multiple AOAs from source to the array. Then, by reversing the AOA rays and determining their intersection, the location of sound source can be obtained. Let θ_1 and θ_2 denote the AOAs of LOS and ECHO paths, respectively. The 2D coordinates (x_s and y_s) of source can be calculated as:

$$\begin{cases} x_s = \frac{2d_w}{\tan\theta_2 - \tan\theta_1}, \\ y_s = d_w + \tan\theta_1 x_s. \end{cases} \tag{3}$$

C. Issues of Existing Solutions

Utilizing the triangulation method in the ISSL problem with a small-sized Mic array presents the following two limitations:

(i) Triangulation method typically relies on accurate estimation of multiple AOAs. According to [24], the angular resolution of array is linearly related to array size as $\Delta\theta = \frac{0.89\lambda}{D}$, where $D = (N-1) \times d_a$ is the diameter of array. Given a sound frequency of $500Hz$ and $D = 0.1m$, the angular resolution is only $6.1°$. This demonstrates that the resolution of AOA estimation with a small-sized Mic array is usually error-prone. We consider two scenes in Fig. 2, i.e. when the AOA difference between LOS and ECHO paths is large and small. Given a Mic array at $(0, 0.4)m$ and a source at $(2, 1)m$ or $(2, 4)m$, the AOA difference between LOS and ECHO of a nearby wall paths in these two cases are $18.3°$ and $4.6°$, respectively. We have shown the AOA spectra calculated by several common AOA estimation algorithms, including general cross correlation (GCC), GCC-PHAT, Delay-and-Sum (DAS) and MUSIC [32], [33]. As shown by Fig. 2(b), it is hard for these algorithms to identify the AOAs of LOS and ECHO

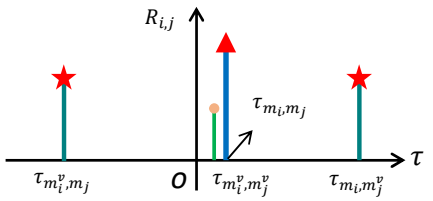


Fig. 4. Cross correlation spectrum for $y_i(t)$ and $y_j(t)$. The triangle, circle and pentagram represent the LOS-LOS, ECHO-ECHO and LOS-ECHO correlation peaks, respectively.

paths precisely when their difference is small. In this case, the AOA-based ISSL system may fail to work.

(ii) Triangulation method may not be easily extended to solve 3D locations. Although the functions constructed in Eq. (3) are sufficient for 2D localization, they become undetermined when solving 3D coordinates. As shown by the black line in Fig. 3(b), there are multiple intersections in 3D when only the first two TDOAs are utilized for localization. To address this problem, Voloc search the source in a 2D plane at known heights [18], whereas Symphony puts the source source at the same 2D plane with the Mic array [19]. These two limitations both demonstrate the weakness of source localization with only AOA information.

D. Insight of This Work

Our core observation focuses on the correlation phenomenon of indoor multipath signals. Fig. 1(b) illustrates the collected signals from m_i and m_j $(i < j)$ in time domain, which are the superposition of LOS and multiple ECHO signals. Because the source signals are usually unknown in the real life, cross correlation has been widely adopted to calculate the relative delay information in different channels [7], [15]. Based on our assumption that the LOS and one ECHO signal reflected by a nearby wall are the most typical, there are mainly four types of TDOAs between the i-th and j-th channel, i.e., path delays between LOS$_i$-LOS$_j$, ECHO$_i$-ECHO$_j$, LOS$_i$-ECHO$_j$, and ECHO$_i$-LOS$_j$ signals, respectively. Supposing that we have obtained these useful TDOA parameters, a feasible solution for localization is to build and solve an equation set as:

$$\begin{cases} |P_{m_j} - P_s| - |P_{m_i} - P_s| = \tau_{m_i,m_j} c, & ① \\ |P_{m_j^v} - P_s| - |P_{m_i^v} - P_s| = \tau_{m_i^v,m_j^v} c, & ② \\ |P_{m_j^v} - P_s| - |P_{m_i} - P_s| = \tau_{m_i,m_j^v} c, & ③ \\ |P_{m_j} - P_s| - |P_{m_i^v} - P_s| = \tau_{m_i^v,m_j} c. & ④ \end{cases} \tag{4}$$

where $\tau_{m_i,m_j} = (t_{m_j} - t_{m_i})$ denotes the TDOA from P_s to Mics m_i and m_j, $t_{m_i} = \frac{|P_s - P_{m_i}|}{c}$ is the TOF that sound arrives at m_i, and c is the sound speed, respectively. This localization model offers three key advantages:

Why is 3D localization feasible: In equation set (4), the combination of terms ① and ② corresponds to the triangulation method described in Eq. (3), which constructs two hyperbolas or hyperboloids based on focal points of $< P_{m_i}, P_{m_j} >$ and $< P_{m_i^v}, P_{m_j^v} >$. As shown by Fig. 3(b), two localization curves alone are insufficient for determining 3D coordinates because multiple intersections occur. This work introduces the use of TDOA information between real and virtual Mic elements (i.e. terms ③ and ④) for localization. They can construct two additional hyperboloids based on focal points of $< P_{m_i}, P_{m_j^v} >$ and $< P_{m_i^v}, P_{m_j} >$. Because term ③ and ④ are independent of ① and ②, it can be possible to solve 3D coordinates.

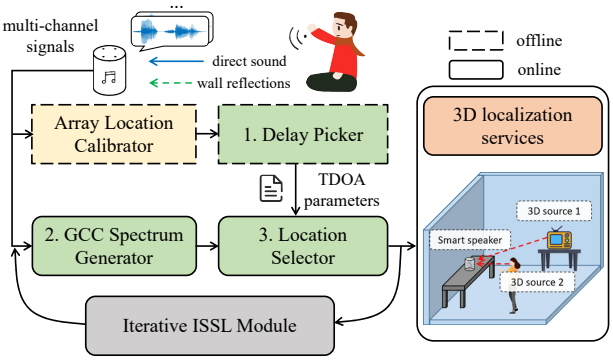


Fig. 5. System architecture of HearLoc. The ISSL system consists of: 1) delay picker, 2) GCC spectrum generator and 3) location selector. The ISSL module is used for locating multiple sources iteratively, and the array location calibrator aims to identify the location of array within a room.

Improved localization ability: Triangulation methods mainly estimate AOAs by calculating the relative delays in the LOS and ECHO paths. Their perspectives primarily focus on the original small-sized Mic array of diameter D. Instead, this work considers all useful correlation information to construct and utilize a large virtual cross-wall array with dimensions $(2d_w, D)$ for localization. According to the Fresnel range equation [24], the distance limit of near field localization is given by $d_\delta = \frac{2D^2}{\lambda}$. When the distance between source and array exceeds d_δ, only AOA of source can be obtained, not its precise location. By effectively increasing the array size from its physical diameter D to a virtual diameter $2d_w$, the localization ability can be enhanced by a factor of $(\frac{2d_w}{D})^2$.

View on correlation power: According to the cross correlation function, i.e. $Corr_{i,j}(\tau) = E[y_i(t-\tau)y_j(t)]$, the correlation powers at the four TDOAs in Eq. (4) are proportional to $\{\alpha_1^2,\ \alpha_2^2,\ \alpha_1\alpha_2,\ \alpha_1\alpha_2\}$, where α_1 and α_2 are the path signal strengths for the LOS and ECHO paths, respectively. Since the ECHO path typically experiences greater attenuation due to diffuse reflection loss on the wall, we have $\alpha_1 > \alpha_2$. Consequently, these correlation powers are ranked as $(\alpha_1)^2 > \alpha_1\alpha_2 > (\alpha_2)^2$. Under a same noise level, estimating the AOA of the ECHO path has the lowest power $(\alpha_2)^2$, thus is the most challenging. This work additionally models the correlation information between the LOS and ECHO signals, whose power is the second largest $(\alpha_1\alpha_2)$. By combining all the useful TDOA information, precise localization can be more easily achieved.

Though promising, precisely extracting and utilizing the values of useful TDOAs remains challenging due to the presence of other unpredictable multipath signals. To address this, we present our system overview and design details.

IV. System Overview

Different from previous works that estimate multiple AOAs, this work directly solves the ISSL problem in delay domain. We design the HearLoc with three core functions:

- **ISSL System (§V)**: Assuming that the location of array within a room has been calibrated by the *Array Location Calibrator* (§VII), we first generate the TDOA parameters based on the geometric relationship between sound source and array by the *Delay Picker*. Utilizing the GCC spectrum obtained by the *GCC Spectrum Generator*, we then formulate a signal processing scheme to selectively pick and sum the GCC power at TDOAs computed by each location (i.e. *Location Selector*). We design techniques of GCC spectrum interpolation, localization matrix normalization and pruning to improve system accuracy and efficiency. Finally, the optimal estimation for source location can be obtained by selecting the maximal output of Location Selector.
- **Iterative ISSL Module (§VI)**: To solve the scene of multiple sources, an iterative algorithm is designed that sets the GCC power induced by each estimated location as zero iteratively. The ISSL system will continuously updated GCC spectra for localization of other sound sources in each iteration round.
- **Array Location Calibrator (§VII)**: The built-in loudspeaker and Mic array collaborate to calibrate the location of array within a room. It is achieved by constructing an optimization function that minimizes the difference between theoretical TOF of wall ECHO and TOF obtained by signal processing.

The relationship of these modules is shown in Fig. 5.

V. ISSL System

We outline the workflow of the ISSL system for a single source in Alg. 1. It begins with the initiation of delay picker, which generates the useful TDOAs at each location (§V-A). Subsequently in the location selector (§V-C), the location support energy is computed based on the TDOA parameters obtained by the delay picker and GCC spectrum of multiple Mic pairs obtained by GCC spectrum generator (§V-B).

A. Delay Picker

We define a delay picker as the set to contain the useful TDOAs in Eq. (4). Mathematically, the delay picker for Mic pair m_i and m_j, denoted by $T_{i,j}$, can be expressed as:

$$T_{i,j} = \{\tau_{m_i,m_j},\ \tau_{m_i^v,m_j^v},\ \tau_{m_i,m_j^v},\ \tau_{m_i^v,m_j}\}. \tag{5}$$

The TDOA parameters in Eq. (5) can be precisely calculated by Eq. (4). We should note that the delay picker has infinite resolution. But in signal level, the delay picker should be rounded at the sampling point level as $T_{i,j} = \lceil T_{i,j} \times Fs \rfloor / Fs$, where Fs is the sampling rate and $\lceil \cdot \rfloor$ is the rounding process.

Modeling more than one ECHO: It should be noted that, this paper is mainly under the assumption that the LOS and one strong ECHO signal are present. However, it is also promising to model more ECHOs and help localization. Supposing that the array is at a corner, there can exist two significant ECHO signals. Two virtual arrays can be generated correspondingly, denoted by $v_1 = \{m_1^{v_1}, \cdots, m_M^{v_1}\}$ and $v_2 = \{m_1^{v_2}, \cdots, m_M^{v_2}\}$.

Algorithm 1 Workflow of ISSL system for a single source

Input: Array received signal $y(t)$
Output: Estimated location of sound source P_s^*

```
Generate a location grid in the room space as Ω_room;
for each P_s^u ∈ Ω_room do
    Generate the delay picker T_{i,j} at P_s^u for each Mic pair
    < m_i, m_i > by Eq. (5) and Eq. (8), i, j ∈ [1, M], i < j;
end for
Compute the GCC spectrum of each Mic pair by Eq. (7)
as {C_{i,j}};
Extract the top W peaks in C_{1,M} as C_W;
Initialize localization array E_arr as a whole zero array;
for each P_s^u ∈ Ω_room do
    if R(τ_{m_1,m_M}(P_s^u)) ∈ C_W then
        Compute E_vec at P_s^u by Eq. (9);
        Update the d-th row of E_arr as E_vec(P_s^u);
    end if
end for
Normalize each column in E_arr in range 0−1 to get Ê_arr;
Sum Ê_arr along each row to get the array of location
support energy E;
According to Eq. (10), select the location that produces
the maximal E as the estimated source location P_s^*.
```

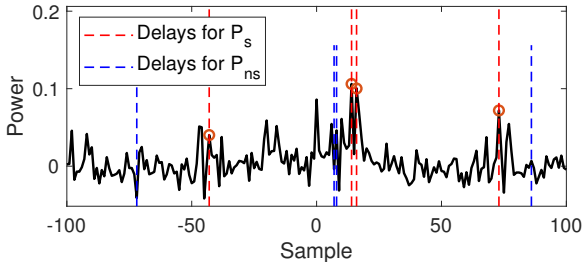


Fig. 6. Relative delays in delay picker generated by ground truth (GT) and non-source locations on the GCC-PHAT spectrum.

The multipath signals correlate with each other, and the delay picker in this case can be given by:

$$T_{i,j} = \{\underline{\tau_{m_i,m_j}},\ \underline{\tau_{m_i^{v_1},m_j^{v_1}}, \tau_{m_i,m_j^{v_1}}, \tau_{m_i^{v_1},m_j}},\ \underline{\tau_{m_i^{v_2},m_j^{v_2}}, \tau_{m_i,m_j^{v_2}}, \tau_{m_i^{v_2},m_j}},\ \underline{\tau_{m_i^{v_1},m_j^{v_2}}, \tau_{m_i^{v_2},m_j^{v_1}}}\}. \tag{6}$$

It is consisted of four parts: the relative delay between Mics in (I) real array, (II) real and array v_1, (III) real and virtual array v_2, and (IV) array v_1 and v_2, respectively. Supposing that signals from K paths are modeled, the number of TDOAs in the delay picker will be K^2. We evaluate the effects of modeling more ECHOs in §VIII-A5.

B. GCC Spectrum Generator

After generating the delay picker according to the locations of source and array within the room, we utilize General Cross Correlation-Phase Transform (GCC-PHAT) for its robustness in multipath environments [32], which can also be seen in Fig. 2. In the following, GCC-PHAT is abbreviated as GCC for simplification. The GCC power at delay τ for the i-th and j-th channels can be computed in frequency domain as:

$$C_{i,j}(\tau) = \sum_{n=0}^{N-1} Re\left(R_{i,j}^{phat} e^{-j2\pi f_n \tau}\right), \tag{7}$$

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2024.3507035

Algorithm 2 Process of FFT interpolation

Input: Conjugate multiplication term $R_{i,j}^{phat}$, interpolation factor I

Output: GCC spectrum with sub-sampled delay bins $\hat{C}_{i,j}$

1: Perform FFT shift in $R_{i,j}^{phat}$ to obtain $R_{i,j}^{phat,shift}$;
2: Compute $N_{pad} = N \times (I-1)/2$;
3: $R_{i,j}^{phat,shift,pad}$=[zeros(N_{pad}); $R_{i,j}^{phat,shift}$; zeros(N_{pad})];
4: Perform inverse FFT shift on $R_{i,j}^{phat,shift,pad}$ to obtain $R_{i,j}^{phat,pad}$;
5: Compute $\hat{C}_{i,j} = Re(IFFT(R_{i,j}^{phat,pad}))$.

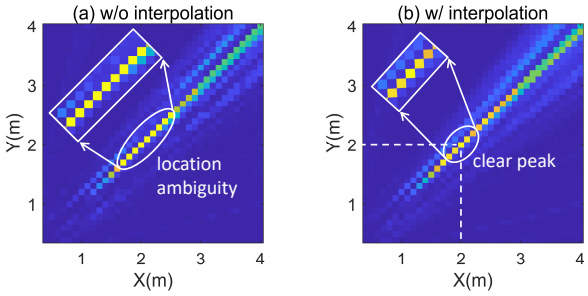


Fig. 7. Localization heatmap (a) without (b) with interpolation. The Mic array is at $(0, 0.3)m$ and $0.3m$ away from a wall. For a better view, the LSEs in the heatmap are normalized and computed to their fourth power.

where $R_{i,j}^{phat} = \frac{Y_i(f_n)Y_j^*(f_n)}{|Y_i(f_n)Y_j^*(f_n)|}$ is the conjugate multiplication term with phase weighting, $Y_i(f)$ denotes the frequency representation of $y_i(t)$, N is the number of sampling points, f_n is the n-th ($n \in [1, N]$) frequency component, $|\cdot|$, $Re(\cdot)$ and $(\cdot)^*$ denote taking amplitude, real part, and complex conjugation, respectively. Instead of computing the correlation power for all delays, it is feasible to utilize inverse FFT (IFFT) on $R_{i,j}^{phat}$ to obtain the GCC spectrum more efficiently.

We set a 4-Mic ULA whose $d_a = 0.05m$ and $0.3m$ away from a wall. Unless specified, the array in the following text is set to this configuration. Let the 2D location of Mic be $(0, 0.3)m$. According to Eq. (5), we generate the delay pickers for the source and non-source locations, which are at $(2, 2)m$ and $(1, 2)m$, respectively. Fig. 6 shows the obtained spectrum with GCC-PHAT for Mic 1 and 4. As indicated by the red and blue dashed lines, it can be observed that the correlation powers are strong at TDOAs generated by the source, where weak at TDOAs associated with non-source locations.

Interpolation to address localization ambiguity: Since the sampling rate on most off-the-shelf acoustic devices is $48kHz$ at most [16], we observe a serious location ambiguity phenomenon when determining adjacent points. According to the location selector (§V-C), we set a sound source at $(2, 2)m$ and the search step to $0.1m$. As shown by Fig. 7(a), it is observed that there are locations with the same location support energy. The problem occurs in the rounding process of $T_{i,j}$, which rounds it to a closest sampling point with a step of $1/Fs$. Adjacent locations may have the same delay picker because of the rounding process. To address this problem, we perform FFT interpolation on $R_{i,j}^{phat}$ to improve the resolution of GCC spectrum. As described in Alg. 2, this method involves in padding zeros on the conjugate multiplication term in frequency domain and then perform IFFT. The delay picker

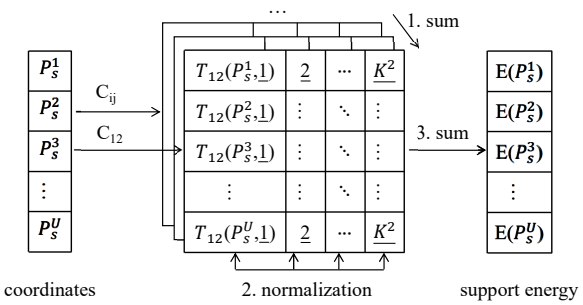


Fig. 8. Scheme of GCC normalization. The GCC powers are normalized by columns to balance their ratio differences.

with an interpolation factor $I(\geq 1)$ can be rewritten as:

$$\hat{T}_{i,j} = \frac{\lceil T_{i,j} \times Fs \times I \rfloor}{Fs \times I}. \tag{8}$$

In this way, the TODA in delay picker can fall into sub delay bins with a resolution of $\frac{1}{Fs \times I}$. This helps improve both the temporal and spatial resolution in localization.

Fig. 7(b) shows the LSE heatmap with an interpolation factor of 9. It can be observed that the interpolation process can enhance the distinctiveness of LSEs for adjacent points.

C. Location Selector

1) Location Support Vector: An intuitive idea for localization is to extract multiple local GCC peaks and then utilize their corresponding TDOAs to solve Eq. (4). However, we discover that the GCC peaks induced by the GT location are usually hard to identify accurately in such multipath-rich environment, as shown by Fig. 6. If the wrong TDOAs are picked, there would produce a significant localization error in solving the equation set. Instead, we define the existence likelihood of source location as the power sum of the GCC power induced by each location. Mathematically, the location support vector at P_s^u, where $u \in [1, U]$ and U is the number of generated points on grid, is defined by the sum of GCC powers at each type of TDOA as:

$$E_{vec}(P_s^u) = \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} C_{i,j}(\hat{T}_{i,j}(P_s^u)). \tag{9}$$

The GCC powers are summed across all Mic pairs to improve robustness. The size of $E_{vec}(P_s^u)$ is equal to the delay picker, i.e., $1 \times K^2$ (1×2^2 in the main part of this work).

2) GCC Power Normalization: We have attempted to directly sum all the GCC powers in the location support vector for localization. However, a suboptimal localization performance is observed with this method. It is primarily due to the magnitude difference of GCC peaks at these four types of TDOA. As depicted in Fig. 4, their ratio rank of GCC power satisfies $(\alpha_1)^2 > \alpha_1\alpha_2 > (\alpha_2)^2$ in theory. Directly summing the GCC powers may produce increased attention towards the direction with a higher GCC power.

To address this problem, we adopt a min-max normalization approach on the localization array, which aims to balance the difference power of GCC peaks. The scheme can be illustrated in Fig. 8. The location support vectors computed by multiple locations can construct an array, i.e. the localization array

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2024.3507035

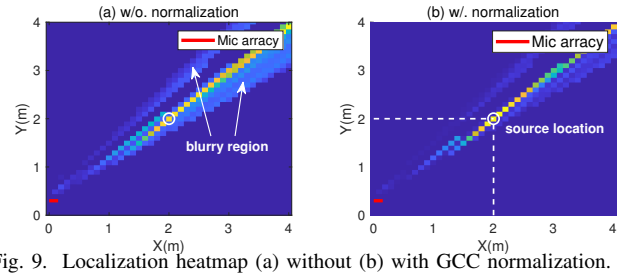


Fig. 9. Localization heatmap (a) without (b) with GCC normalization.

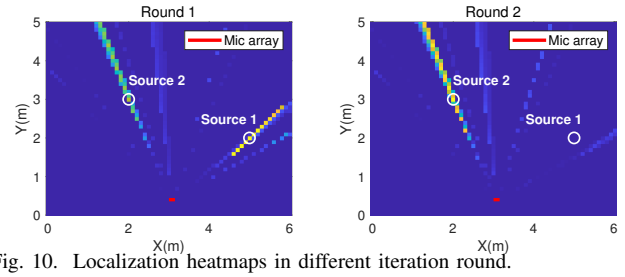


Fig. 10. Localization heatmaps in different iteration round.

E_{arr}, and its size is $U \times K^2$. Normalization is applied on each column of the localization array. Let $\hat{E}_{arr}$ denote the GCC power array after normalization. We define the location support energy (LSE) at P_s^u by the sum of normalized GCC power in each row as $E(P_s^u) = \sum_{i=1}^{K^2} \hat{E}_{arr,i}(P_s^u)$.

After the normalization process, the optimal estimation of sound source location, denoted by P_s^*, can be obtained by extracting the maximal LSE in a room space $\Omega_{room} = \{P_s^1, \cdots, P_s^u, \cdots, P_s^U\}$ as:

$$\textbf{P1:} \quad P_s^* = \arg\max_{P_s^u}(E(P_s^u)), \ \forall P_s^u \in \Omega_{room}. \tag{10}$$

Different from the triangulation method that has infinite solution space, we solve **P1** by grid search in limited room spaces. The search space is set in 2D or 3D to accommodate the respective localization requirements. Fig. 9 shows the localization heatmaps before and after adopting GCC normalization. It can be observed that the blurry regions near the true AOA direction can be clearly reduced with the proposed normalization process.

3) Reduction of Searching Space: Directly searching for all the locations in a room space may be time-consuming and redundant for many real-time applications. We thus propose to reduce the search space by locating the LOS-LOS correlation peak first. The idea comes from the fact that the sound source has a higher possibility that comes from the direction inferred by some strong LOS-LOS correlation peaks. To improve fault tolerance, we select the top W largest peak candidates in $R_{1,M}$ to be the constrained search range. We only compute the LSE at delay picker that contains these . In this work, we define the amount of W as:

$$W = \left\lceil \beta I \times \frac{DFs}{c} \right\rceil, \tag{11}$$

where β is the percentage of pruning, I is the aforementioned interpolation factor, $\frac{DFs}{c}$ is the largest sample shift between Mic pair m_1 and m_M, and $\lceil \cdot \rceil$ represents rounding up. We only compute the support energy at locations that can produce the W peaks at $R_{1,M}$ for saving time. Along with the GCC normalization process, our method can place greater emphasis on the local peaks on the GCC spectrum.

VI. Iterative ISSL Module to Detect Multiple Sources

In practical scenes, there may exist multiple sound sources simultaneously. The correlation peaks produced by different sources will occur on each correlation spectrum, and in this case, it could be challenging to identify the attribution of each correlation peak. However, because the ECHO path mainly experiences the reflection and path attenuation losses [20], the strength of the ECHO path can be written as $\omega l_a \alpha$, where ω is the material reflection coefficient on wall, l_a is the path attenuation in air, and α is the source strength. Furthermore, the path attenuation of the ECHO path is related to its additional traveling distance, which is $2d_w$ maximal. Given $d_w = 0.4m$ and the attenuation factor in air $\gamma = 5 \times 10^{-3} dB/m^2$, l_a satisfies $e^{-\gamma \times 2d_w} = 0.996 \approx 1$, which follows the exponential decay law. This shows, the additional attenuation of the ECHO path in air is negligible, and the ratio of useful correlation powers can be simplified as $\{\alpha^2, \omega^2\alpha^2, \omega\alpha^2, \omega\alpha^2\}$. For a fixed room and wall material, ω can also be viewed as a constant. Thus, all the correlation powers produced by each source are only determined by its source strength α. We also note that, according to the cross correlation properties, the collected correlation spectrum is the linear superposition of multiple correlation phenomena produced by different sources.

Based on these observations, we propose an iterative ISSL algorithm. Our insight is to extract the current most significant combination of correlation peaks on the collected cross correlation spectrum iteratively. Specifically, For each iteration, we can obtain one optimal estimation of sound source by running Alg. 1. Then, we set the GCC powers at TDOAs induced by this location to all zeros. In the next iteration, a new localization heatmap can be generated according to the updated GCC spectrum. This process can eliminate the impact of each source iteratively because the combination of useful correlation peaks of one source can be viewed as a whole. Note that, the process of setting zero is applied on the original GCC spectrum without interpolation because of the physical sampling limits. The interpolation process is then performed again for localization before the next iteration. The technical details are presented in Alg. 3.

We examine the localization performance in different iteration rounds by setting two sources at $(2, 3)m$ and $(5, 2)m$. The Mic array is at $(3, 0.3)m$ and $0.3m$ away from a wall. We illustrate the localization heatmaps for two rounds in Fig. 10. It can be observed that the LSEs of the two sources are different in Fig. 10(a) because of their distinct path signal strength. Furthermore, as shown by Fig. 10(b), the impact of source 1 can be well eliminated by our setting zero process.

VII. Array Location Calibrator

The location selector assumes that the array's location is known in advance. However, in practical scenarios, determining the initial location of the array is often challenging or prone to fluctuations because of external interference. As

[2]The values of attenuation factor γ in air are from [34].

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2024.3507035

Algorithm 3 Iterative ISSL algorithm for multiple sources

Input: Array received signal $y(t)$, number of sources F
Output: Location set of F sources $\{P_{s,1}, \cdots, P_{s,F}\}$

Compute the original conjugate multiplication term as $R = \{R_{i,j}^{phat}\}$, where $i, j \in [1, M], i < j$;
Compute GCC spectrum with interpolation factor I as $\hat{C} = \{\hat{C}_{i,j}\}$;
for each $f \in [1, F]$ **do**
  Run Alg. 1 on $\hat{C}$ to obtain the optimal estimation of source location $P_{s,f}$;
  for each $i \in [1, M]$ **do**
    for each $j \in [i+1, M]$ **do**
      Perform IFFT on $R_{i,j}^{phat}$ to obtain $C_{i,j}$ without interpolation;
      Set $C_{i,j}(\hat{T}_{i,j}(P_{s,f}))) = 0$;
      Perform FFT on $C_{i,j}$ to obtain $R_{i,j}^{f}$;
      Input $R_{i,j}^{f}$ and interpolation factor I to Alg. 2 and obtain the interpolated $\hat{C}_{i,j}^{f}$;
      Update $\hat{C}_{i,j}$ to $\hat{C}_{i,j}^{f}$.
    end for
  end for
end for

a result, there is a need for an effective and accurate array location calibration scheme to ensure system usability.

Besides the Mic array, most IAs such as smart speakers and sweeping robots, are usually equipped with a built-in loudspeaker, which can be controlled to actively produce known sounds. This makes it possible for us to calculate the actual TOFs of sound from the built-in loudspeaker to the Mic array. The geometric relationship between the array and room can be inferred by analyzing the propagation delays. Fig. 11 illustrates the sound propagation when performing array location calibration. Let d_w and ψ denote the wall-array distance and Mic orientation, respectively. The location of the i-th Mic in a ULA can be expressed as:

$$P_{m_i} = ((i-1)d_a \cos(\psi), d_w + (i-1)d_a \sin(\psi)). \quad (12)$$

As mentioned before, the location of the i-th virtual Mic $P_{m_i^v}$ has the same x but inverse y coordinates as P_{m_i}. Supposing the location of loudspeaker P_s is set in the middle of array, its position can be given as $(\frac{D}{2}\cos(\psi), d_w + \frac{D}{2}\sin(\psi))$.

In order to obtain the geometric relationship between room and array, we build an optimization function for a fine-grained result. This is achieved by minimizing the delay difference between the real measurements and model based results. The optimization parameters are wall-array distance d_w and Mic orientation ψ. The function can be constructed as:

$$\textbf{P2:} \quad \min_{d_w, \psi} \sum_{i=1}^{M} \left(\frac{|P_s - P_{m_i^v}| \times Fs}{c} - t_{m_i^v} \right)^2, \quad (13)$$

where $t_{m_i^v}$ is the estimated TOF from loudspeaker to Mic m_i^v. Note that, both P_s and $P_{m_i^v}$ are functions of d_w and ψ. The combination of d_w and ψ that can produce the minimal value in **P2** is selected to be the optimal estimation result of the array location.

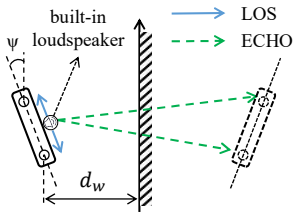


Fig. 11. Array location calibration.

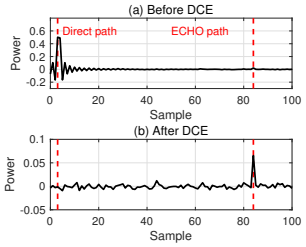


Fig. 12. Channel response spectrum before and after direct components elimination (DCE).

To solve **P2**, the accurate extraction of $t_{m_i^v}$ is still a problem. We design to transmit Zadoff-Chu (ZC) sequences due to their property of strong autocorrelation [35]. Because audible sound for array location calibration can interfere with daily human activities, we design the ZC sequences in $18k$-$22kHz$, which is inaudible to most people. Then we can obtain the channel information by performing cross correlation on the received signal $y(t)$ and $s(t)$. As shown by the results in Fig. 12(a), it can be observed that there is an extremely high peak caused by the direct path due to its strong power, whereas the GCC peak caused by the ECHO path is less apparent. To address this problem, we eliminate the direct sound components by subtracting the signal that only contains the sound from loudspeaker to Mic array, which can be collected in advance. Fig. 12(b) shows the result after direct components elimination. It can be observed that the GCC peak induced by the nearby wall reflection becomes clear and easy for identification. Meanwhile, an iterative algorithm similar to Alg. 3 can be developed to calculate the geometric relationships between the array and multiple walls.

VIII. Simulation

Settings: We generate the simulation datasets with Pyroomacoustics [25], which is a widely-adopted room acoustic simulator [13], [16], [36], [37]. A 4-Mic ULA of $0.15m$ (spacing is $0.05m$) is placed close to a wall at $0.4m$ by default. In fact, other distances are also acceptable and will be evaluated next. We choose three types of rooms in our experiment, i.e., room 1: $4 \times 4 \times 2.8m^3$, 2: $4 \times 6 \times 2.8m^3$ and 3: $8 \times 5 \times 3.2m^3$. Their reverberation time levels (RT_{60}) are set to $0.5s$, $0.4s$ and $0.3s$, respectively. The sound speed is set to $c = 343m/s$. In 2D experiments, the source is set at the same plane with the Mic array and the step size of $0.4m$. In 3D experiments, the height of source is additionally set to 0-$0.8m$ higher over the array with an interval of $0.2m$. Trigger words, including "OK Google", "Hi siri" and "Alexa" are randomly selected to be the source signal, and their duration is $1s$. During data collection, the audio sampling rate is set to $48kHz$. The search step along all directions is set to $0.1m$. The search space is set in the 2D plane of room and $1m$ higher over the array. With regard to parameter configuration, the factors of FFT interpolation I and pruning β are set to 9 and 0.2, respectively. The evaluation metric is defined by the euclidean distance between the GT and estimated location. If not specifically mentioned, the experiments are in room 3 with 2D simulation by default.

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2024.3507035

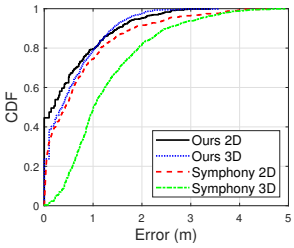


Fig. 13. Overall error CDF in 2D and 3D.

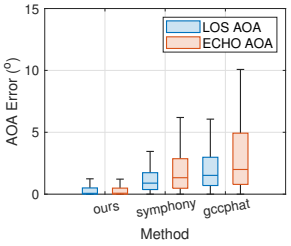


Fig. 14. AOA Accuracy of three algorithms.

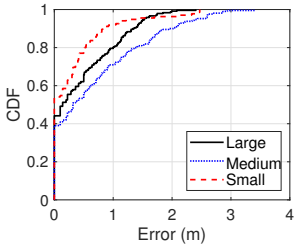


Fig. 15. Error CDF over room types.

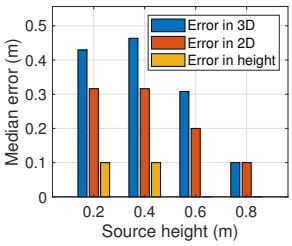


Fig. 16. Results over heights.

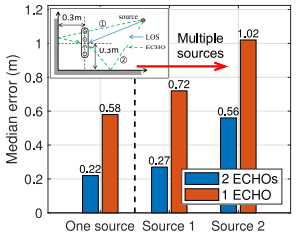


Fig. 17. Error CDF by modeling different number of ECHOs.

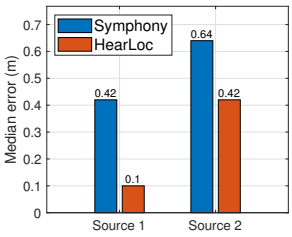


Fig. 18. Error CDF with multiple sources.

Implement Details: Because the parameters in the delay picker and array location calibrator are independent of source signals, we optimize the system efficiency by dividing it into offline and online parts. The offline part involves generating the delay picker and running the array location calibrator, which is a one-time effort. The online part involves the processes of GCC spectrum generator, location selector and iterative ISSL module for signal processing.

Baselines: We choose a recent relevant work Symphony [19] as the baseline. Symphony proposes a different algorithm that extracts the linear relationship of LOS-LOS and ECHO-ECHO correlation peaks between ULA elements for AOA estimation. Then, triangulation method is used for localization. Due to the 2D assumption of Symphony, the solution space is set at the 2D plane that the Mic array locates.

A. Marco Benchmarks

1) Overall performance: Fig. 13 shows the cumulative distribution function (CDF) of localization errors in different dimensions. We have achieved a median error of $0.2m$ and $0.37m$ in 2D and 3D across different rooms, respectively. The results demonstrate the ability of our system to locate sound source accurately in both 2D and 3D. The median localization error of Symphony is $0.4m$ in 2D, which is slightly weaker than that of our system. This is mainly because we have additionally considered 2 more LOS-ECHO correlation peaks for localization. In the context of 3D localization, the median error of Symphony is $1m$, which is significantly affected by the height estimation error.

2) Error in AOA estimation: Fig. 14 illustrates the accuracy of AOA estimation with three methods, i.e. ours, Symphony and GCC-PHAT [32]. Different from our method, Symphony considers the TDOAs of LOS and ECHO path, while GCC-PHAT only models the LOS-LOS correlation peak. Their median AOA errors of LOS path are $0.07°$, $0.9°$ and $1.7°$, respectively. We have achieved the highest accuracy of AOA estimation over the baseline methods. This is owing to our consideration of more TDOA information for localization.

3) Room type: Fig. 15 shows the localization performance in different rooms (i.e., room 3). The median errors in the large, medium and small rooms are $0.14m$, $0.3m$, $0m$, respectively. The result demonstrates the effectiveness of our system for deployments in different types of room. We have achieved the best localization performance in the small-sized room, even though its reverberation level is the strongest. Although there are more strong unmodeled ECHO signals in this room, the correlation powers at interested TDOAs are also more significant in this case.

4) Source height: Fig. 16 examines the influence of source height on the localization error. We vary the source from $0.2m$ to $0.8m$ higher than the array in increments of $0.2m$. The blue, red and yellow bars in the figure represent the median localization error in 3D, 2D and height, respectively. The results demonstrate an error of $< 0.5m$ in 3D localization and $\leq 0.1m$ in z-axis at all heights. Meanwhile, it is worth noting that introducing the 3D localization model can effectively ensure the error of 2D localization at a low level.

5) Modeling Multiple ECHOs: The main part of our model relies on the assumption that the LOS and one strong ECHO signals are present. In cases at a corner, there may be other strong ECHO signals that are not modeled. To investigate this problem, we set the Mic array at a corner, with its distance to two near walls of both $0.3m$. As shown by 17, the median error when modeling one or two ECHO paths is $0.58m$ and $0.22m$, respectively. Although the localization performance decreases in the corner, the system accuracy can be significantly improved by modeling more ECHOs, showing an improvement of $2.6\times$ in localization accuracy. Additionally, in scenes with multiple sources, modeling more ECHOs can also help improve the localization performance for the first and second sources by 2.7 and 1.8 times, respectively.

6) Multiple sources: To deal with scenes with multiple sources, we place 2 sources randomly in room 2. Then, we compute the CDF of Euclidean distance error for these 2 sources. As shown by Fig. 18, we achieve a median localization error of $0.1m$ and $0.42m$ for the first and second strong sources (namely source 1 and 2). The localization performance of the second source decreases slightly. This is mainly because this work only sets zero at 4 TDOAs on the GCC spectrum, while some other interference correlation peaks induced by multipath may affect the localization accuracy in subsequent iterations. Compared with the baseline method Symphony [19], we have achieved a $4\times$ /$1.5\times$ improved localization accuracy for multiple sources.

7) Array location calibration: Fig. 19 shows the effects of array location calibration. The Mic array is set close to a wall

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2024.3507035

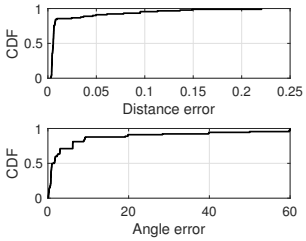


Fig. 19. Distance and angle error of array location calibration.

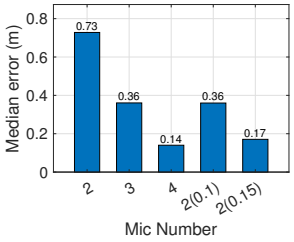


Fig. 20. Impact of Mic number and spacing.

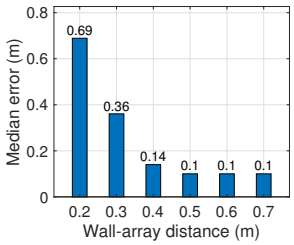


Fig. 22. Impact of wall-array distance.

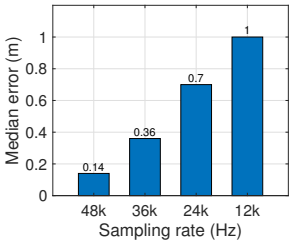


Fig. 23. Impact of sampling rate.

Fig. 21. Ablation study.

Name	Median error (m)		Latency (s)	
	2D	3D	2D	3D
Symphony [19]	0.4	1	0.82	
Ours	**0.2**	**0.37**	**0.19**	**0.2**
w/o. norm	0.36	0.51	0.19	0.2
w/o. interp	0.36	0.58	0.03	0.05
w/o. pruning	0.20	0.46	0.2	0.21
w/o. all	0.50	0.63	0.04	0.06
First 2 TODAs	1.14	1.3	0.19	0.2
First 3 TODAs	0.32	0.68	0.19	0.2

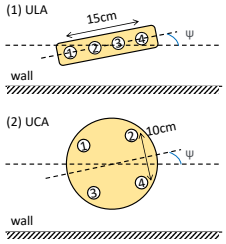
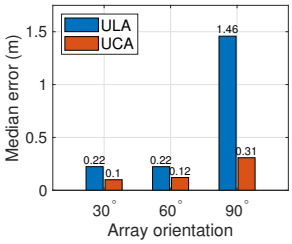


Fig. 24. Impact of Mic orientation.



at a distance of 0.2-0.8m with a random inclination angle. The calibration algorithm is then run to infer the wall-array distance and angle. The results show that we have achieved small median errors of $0.006m$ and $1.32°$ in distance and angle, respectively. This demonstrates the effectiveness of the proposed location calibration scheme, which is the foundation of the entire localization system.

8) Efficiency: We evaluate the processing latency of the four modules of our system. The algorithm is run by Matlab on PC of Intel i5-11400, 2.6GHz. For the online part, the GCC spectrum generator costs $0.193s$ for a one second speech, and the location selector for 2D and 3D localization costs $0.014s$ and $0.022s$, respectively. For the offline part, the delay picker and array location calibrator cost $0.23s$ and $0.03s$ to complete. Because Symphony performs signal correlation for multiple times to obtain more precise AOAs, its latency is about 4 times larger than our system.

9) Ablation study: We conduct ablation studies to verify the effectiveness of our approach. Table. 21 shows the median localization error and latency in 2D and 3D under different settings. Firstly, we observe a large localization error increase when not adopting the normalization and interpolation schemes. This is because normalization eliminates the power differences between different cross correlation types. Besides, interpolation can improve the resolution of GCC spectrum. We also observe that pruning can improve the efficiency but ensure the accuracy. The reason is that, although pruning reduces search space, it forces the localization along AOA directions with high correlation powers. Furthermore, in order to examine our design of delay picker, we selectively only utilize the first two and three TDOAs in Eq. (5) for localization. We have observed a significant increase in accuracy when additionally considering the third and fourth TDOAs. This is because they also provide useful TDOA information that can be utilized in localization.

B. Micro Benchmarks

1) Mic number and spacing: Fig. 20 shows the system performance with changes of Mic number and spacing. We first only use 2 and 3 adjacent Mics in the 4-Mic ULA for localization. It can be found that the localization error increases with the decrease of Mic number. This is because more Mics allow for a greater number of combinations for cross correlation, which can lead to a more robust estimation for the source location. We also investigate the impact of Mic spacing on the localization performance. We use a two-Mic pair with spacing of $0.1m$ and $0.15m$ in the 4-Mic ULA. Compared with the benchmark, their localization accuracy decreases by $0.22m$ and $0.03m$, respectively. The reason is that the increase of array size can improve the spatial resolution. It is worth noting that the localization error of 2-Mic with a spacing of $0.15m$ is comparable to that of the original 4-Mic ULA. This observation highlights the potential of our system for utilization on other IAs with a small-sized two-Mic array, such as smartphones and mobile robots.

2) Wall-array distance: Fig. 22 shows the results under different wall-array (WA) distances, which are set at $\{0.1, 0.2, ..., 0.8\}m$. The results reveal that, as the WA distance increases, the localization error gradually decreases to a small value. This because a larger WA distance can construct a cross-wall array of larger size, which improves the spatial resolution. In real-life scenario, the WA distance is not supposed to be too large because the power of ECHO signal may decrease significantly due to a long attenuation path in air.

3) Sampling rate: Fig. 23 demonstrates the impact of sampling rate. We use downsampling on the original dataset of $48kHz$ to construct the downsampled datasets. As shown by the figure, it can be found that the localization error gradually increases with the decrease of sampling rate. The reason is that, a higher sampling rate can provide delay bins with a finer-grained resolution, which helps improve the discriminability between close source locations.

4) Array orientation: There can be multiple array orientations in real life. To investigate their impact, we vary the orientation angle ψ of a ULA and 4-Mic UCA from 0-

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2024.3507035

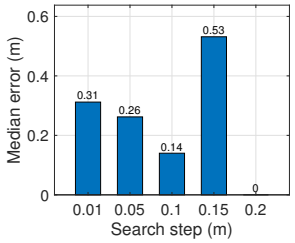


Fig. 25. Impact of search step.

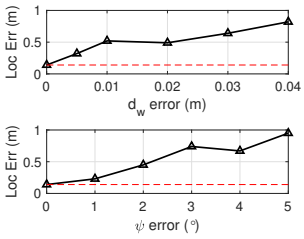


Fig. 26. Error propagation with errors in d_w and ψ.

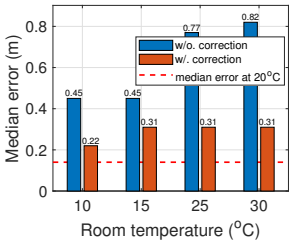


Fig. 27. Impact of room temperature and sound speed correction.

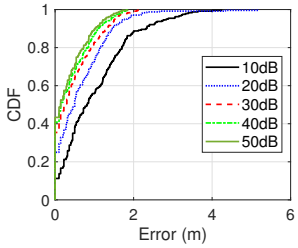


Fig. 28. Error CDF at differnt SNRs of white noise.

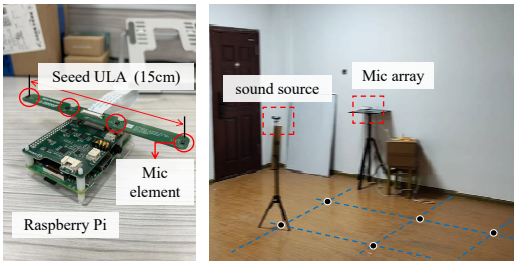
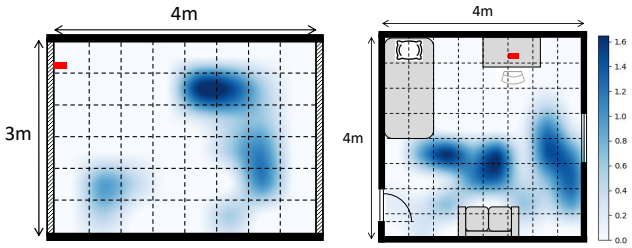


Fig. 29. (a) 4-Mic ULA and Raspberry Pi (b) Experimental scene.



(a) Hall (b) Bedroom

Fig. 30. Error distribution in a (a) $3 \times 4m^2$ hall and (b) $4 \times 4m^2$ room. The red line represents the location of Mic array.

$90°$. The spacing of the UCA is $10cm$. Fig. 24(a) is the illustration of array orientation changes. As shown by Fig. 24(b), orientation has a more significant impact on the ULA. This occurs because, when the orientation of the ULA is $90°$, the constructed 2D cross-wall array becomes a 1D linear array. In this case, the array can not distinguish between left and right positions. In contrast, the UCA maintains localization errors below 0.31m across all orientations, as its virtual cross-wall array remains a 2D array, regardless of orientation changes.

5) Search step: Fig. 25 shows the results under different search steps. It can be found that reducing the search step can not improve the localization performance. This outcome can be attributed to the limited temporal and spatial resolution of off-the-shelf Mic arrays. With the decrease of search step, more location ambiguity may incur, thereby affecting the localization performance. Note that, the localization error is large when the search step is $0.15m$. This is because in this case, the search grid does not correspond to the grids of placing sources, which is set to $0.4m$.

6) Error propagation: Fig. 26 illustrates the changes of localization error with different estimation errors in wall-array distance and orientation (i.e., d_w and ψ). The red dashed line represents the result with no estimation error in the stage of array location calibration. The results reveal that the localization accuracy significantly reduces with the increase of error in d_w and ψ. This is mainly because the delay picker is dependent of array location, and its error will accumulate in the localization stage. However, as reported in §VIII-A7, the results of the proposed array location calibration scheme can ensure that the localization error remains below $0.5m$.

7) Room temperature: In the above experiments, the sound speed is set to $343m/s$ under room temperature of $20°C$ by default. To investigate the impact of temperature changes, we set the room temperature at $10°C - 30°C$ with a step of $5°C$ but the sound speed is still set to $343m/s$. The blue bars in Fig. 27 demonstrates an obvious increase of localization error when temperature deviation exists, and the maximal error even reaches $0.82m$. Fortunately, this issue can be alleviated by measuring the ambient temperature with temperature sensors. We correct the sound speed by $c = 331.4 + 0.6 \times T$, where T is temperature in degree Celsius [27]. As shown by the results in red bars of Fig. 27, it is observed that the localization error can be significantly reduced by the correction process.

8) Noise level: Fig. 28 evaluates the system performance change under SNRs of environmental noise from $50dB$ to $10dB$. We achieve this by adding Gaussian white noise on the original recordings. It can be observed that high level of noises can significantly affect the localization accuracy ($0.76m$ at $10dB$). This is mainly because strong white noises can induce numerous fake peaks on the GCC spectrum.

IX. Real-world Experiments

For validations in the real world, we collect data with Seeed ReSpeaker 4-Mic ULA (spacing $d_a = 5cm$) [38], which is mounted to a Raspberry Pi 3B+ [39]. The source is a smartphone that plays trigger words. The hardware and experimental scene are shown in Fig. 29(a) and Fig. 29(b). We collect data in a $3 \times 4m^2$ hall, bedroom $4 \times 4m^2$ and office $8 \times 8m^2$. Other settings are the same as the simulation section.

Localization in different rooms: Fig. 30 illustrates the heatmap of localization error in two rooms. We have achieved an average error of $0.22m$ and $0.21m$ in hall and bedroom, respectively. Compared with our simulations, the performance only decreases slightly. The main reason is that the simulation room is empty, with only walls and no other objects. Instead, the existence of windows, doors and other reflective objects in real environments can make the collected signals more complex. Because the reflection coefficients of these non-wall materials are usually smaller than the wall [34], they only have a limited impact on the localization performance.

Source trajectory tracking: We investigate the system performance of tracking mobile sources in office. The frame length is set to $0.5s$. Fig. 31(a) and Fig. 31(b) show 2D trajectories with line and circle (radius is $3m$), respectively. The results reveal an average error between the estimated

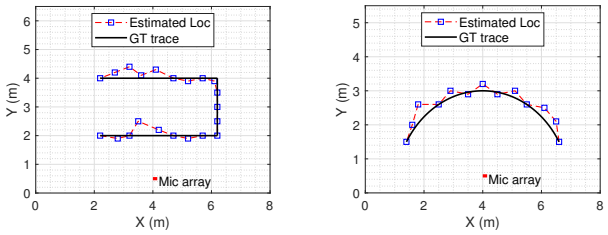


Fig. 31. Source tracking in 2D by (a) line and (b) circle.

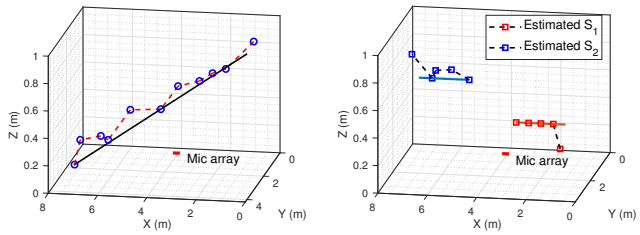


Fig. 32. Source tracking in 3D: (a) a single source and (b) multiple sources.

locations and ground truth trace of $0.12m$. Fig. 32(a) describes a source movement by line from point $(1, 1, 0.8)m$ to $(7, 4, 0.2)$m, and Fig. 32(b) illustrates the tracking results in a scene of multiple sources. Their average errors in the two cases are $0.44m$ and $0.16m$, respectively.

Moving interference: To verify our robustness against interference, we collect data of a sound source with the interference from a walking human, either in silence or talking. The settings are shown in Fig. 33(a). The results in Fig. 33(b) show that, interference on either side of the source only slightly affects the accuracy. However, when the interference is in front of the source, it can obstruct the line-of-sight sound propagation to the Mic array. In this case, the accuracy decreases significantly. Additionally, when the interference is talking, the accuracy also decreases due to the presence of additional unmodeled cross correlation peaks.

Efficiency on Raspberry Pi: The algorithm is locally deployed on a Raspberry Pi 3B+ with Python 3. The average latency of signal processing with a multi-channel speech of size $(24k, 4)$ for localization in 2D and 3D is $1.58s$ and $3.44s$, respectively. We believe that the latency can be further reduced by implementing the algorithm in C++.

X. DISCUSSIONS

How large can the wall-array distance be: We observe the limit of wall-array distance is constrained by three aspects. The first one is that, the SNR ratio between LOS-LOS and ECHO-ECHO correlation should not be smaller than a threshold. This is because the ECHO-ECHO correlation peak can be hard to identify in this case. The second and third are that the sound pressure levels of ECHO speech and ultrasound signals for array location calibration should not be lower than a noise level. We define three signal strengths, including the LOS speech: $A_{LOS} = A_0 e^{-\gamma D}$, ECHO speech $A_{ECHO} = \omega A_0 e^{-\gamma(D+2\times d_w)}$ and ECHO ultrasound $A^u_{ECHO} = \omega A_0 e^{-\gamma_u(2\times d_w)}$, where $A_0 = 10^{\frac{P_s}{20}}$ is the amplitude of source, P_s is the sound pressure level of source, γ is the sound attenuation factor in air, D is the distance between source and array, γ and γ_u are the attenuation

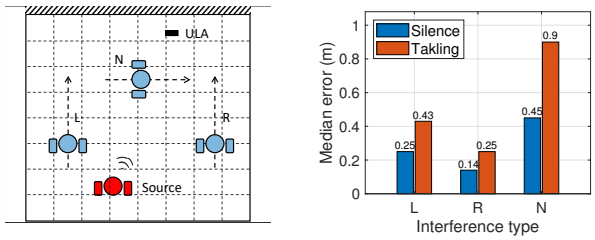


Fig. 33. Moving interference. L and R: walking forward in left and right, and N: walking in front of source.

factor of speech and ultrasound in air, and ω is the reflection coefficient. Mathematically, the constraints can be given by:

$$\begin{cases} \left(\frac{A_{ECHO}}{A_{LOS}}\right)^2 = (\omega e^{-2\gamma d_w})^2 > \text{SNRR}_\delta & ① \\ 20\log_{10}(A_{ECHO}) > P_\delta & ② \\ 20\log_{10}(A^u_{ECHO}) > P^u_\delta & ③ \end{cases} \tag{14}$$

In ①, given $\gamma = 5 \times 10^{-3} dB/m$, $\omega = 0.75$ and $\text{SNRR}_\delta = 1/2$, the wall-array distance should be less than $d^s_w = \frac{1}{4\gamma}\ln(\omega^2/\text{SNRR}_\delta) \approx 3.7m$. In ② and ③, because ultrasound experiences $10\times$ larger attenuation in air than speech in low frequency range, constraint ③ is stricter. Given $P_s = 60dB$, $\gamma_u = 0.39dB/m$ and $P_\delta = 20dB$, d_w should be less than $d^u_w = \frac{\log_{10}\beta + (P_s - P^u_\delta)/20}{2\gamma_u} \approx 2.4m$. The actual limit of wall-array distance of our system should be $\min\{d^s_w, d^u_w\} = 2.4m$.

Why not consider the TDOA between m_i and m^v_i: Note that, we do not consider the localization hyperbola on focal points of $< P_{m_i}, P_{m^v_i} >$. This is because the relative delay that signal reaches m_i and m^v_i can only be obtained by autocorrelation of the collected signal at m_i itself. However, autocorrelation can not incorporate phase weighting because the term $\frac{Y_i(f)Y_i(f)}{|Y_i(f)||Y_i(f)|}$ equals to a constant 1. Additionally, Fig. 2 shows that autocorrelation without phase weighting has poor accuracy for speech signals. A further advantage of considering the relative delays of $< m_i, m^v_j >$ and $< m^v_i, m_j >$ is that we can calculate them simultaneously while performing cross correlation on two channels to obtain relative delays for $< m_i, m_j >$ and $< m^v_i, m^v_j >$.

Scenes with array mobility: This work assumes that the Mic array is in a static state, while the source can be mobile. However, if the array itself is with mobility, the collected signals can become more complex due to significant changes in reflective surfaces and potential device vibrations. We leave this intriguing problem for future work.

XI. CONCLUSION

This paper presents HearLoc, an ISSL system for unknown sources in 3D with a ten-cm Mic array. By effectively utilizing the correlation among multipath signals, the localization ability and dimension can be significantly improved from the original small-sized array to a large cross-wall virtual array. Technically, we design algorithms to localize single and multiple sources, and a calibration scheme for array location within a room. Experiments in various settings show the effectiveness of the proposed scheme, outperforming the existing AOA-based ISSL solutions in both accuracy and efficiency. In our future endeavors, we aim to investigate source localization algorithms for Mic arrays with mobility.

REFERENCES

[1] "Apple homepod," https://www.apple.com/homepod/, 2024.

[2] "Figure 01," https://www.figure.ai/, 2024.

[3] X. Pang, Z. Wang, D. Liu, J. C. Lui, Q. Wang, and J. Ren, "Towards personalized privacy-preserving truth discovery over crowdsourced data streams," *IEEE/ACM Transactions on Networking*, vol. 30, no. 1, pp. 327–340, 2021.

[4] Z. Wang, K. Liu, J. Hu, J. Ren, H. Guo, and W. Yuan, "Attrleaks on the edge: Exploiting information leakage from privacy-preserving co-inference," *Chinese Journal of Electronics*, vol. 32, no. 1, pp. 1–12, 2023.

[5] Y. Zhang, W. Wang, J. Ren, J. Huang, S. He, and Y. Zhang, "Efficient revenue-based mec server deployment and management in mobile edge-cloud computing," *IEEE/ACM Transactions on Networking*, vol. 31, no. 4, pp. 1449–1462, 2023.

[6] J. Gao, C. Zhang, Q. Kong, F. Yin, L. Xu, and K. Niu, "Metaloc: Learning to learn indoor rss fingerprinting localization over multiple scenarios," in *IEEE International Conference on Communications*. IEEE, 2022, pp. 3232–3237.

[7] D. Berghi and P. J. Jackson, "Leveraging visual supervision for array-based active speaker detection and localization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[8] R. Qian, D. Hu, H. Dinkel, M. Wu, N. Xu, and W. Lin, "Multiple sound sources localization from coarse to fine," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*. Springer, 2020, pp. 292–308.

[9] S. Yang, D. Zhang, R. Song, P. Yin, and Y. Chen, "Multiple wifi access points co-localization through joint aoa estimation," *IEEE Transactions on Mobile Computing*, vol. 23, no. 2, pp. 1488–1502, 2024.

[10] C. Wang, J. Liu, Y. Chen, H. Liu, L. Xie, W. Wang, B. He, and S. Lu, "Multi-touch in the air: Device-free finger tracking and gesture recognition via cots rfid," in *IEEE Conference on Computer Communications*. IEEE, 2018, pp. 1691–1699.

[11] S. Woźniak and K. Kowalczyk, "Passive joint localization and synchronization of distributed microphone arrays," *IEEE Signal Processing Letters*, vol. 26, no. 2, pp. 292–296, 2018.

[12] Y. Sun, W. Wang, L. Mottola, J. Zhang, R. Wang, and Y. He, "Indoor drone localization and tracking based on acoustic inertial measurement," *IEEE Transactions on Mobile Computing*, pp. 1–15, 2023.

[13] Y. Wu, R. Ayyalasomayajula, M. J. Bianco, D. Bharadia, and P. Gerstoft, "Sslide: Sound source localization for indoors based on deep learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 4680–4684.

[14] Q. Yang and Y. Zheng, "Deepear: Sound localization with binaural microphones," *IEEE Transactions on Mobile Computing*, vol. 23, no. 1, pp. 359–375, 2024.

[15] A. Canclini, F. Antonacci, A. Sarti, and S. Tubaro, "Acoustic source localization with distributed asynchronous microphone networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 439–443, 2012.

[16] W. Luo, Q. Song, Z. Yan, R. Tan, and G. Lin, "Indoor smartphone slam with acoustic echoes," *IEEE Transactions on Mobile Computing*, pp. 1–15, 2023.

[17] J. C. Curlander and R. N. McDonough, *Synthetic aperture radar*. Wiley, New York, 1991.

[18] S. Shen, D. Chen, Y.-L. Wei, Z. Yang, and R. R. Choudhury, "Voice localization using nearby wall reflections," in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, 2020, pp. 82–95.

[19] W. Wang, J. Li, Y. He, and Y. Liu, "Localizing multiple acoustic sources with a single microphone array," *IEEE Transactions on Mobile Computing*, pp. 1–15, 2022.

[20] M. Wang, W. Sun, and L. Qiu, "Mavl: Multiresolution analysis of voice localization," in *18th USENIX Symposium on Networked Systems Design and Implementation*, 2021, pp. 845–858.

[21] S. Chen, R. Tan, Z. Wang, X. Tong, and K. Li, "Voicemap: Autonomous mapping of microphone array for voice localization," *IEEE Internet of Things Journal*, vol. 11, no. 2, pp. 2909–2923, 2024.

[22] L. Kraljević, M. Russo, M. Stella, and M. Sikora, "Free-field tdoa-aoa sound source localization using three soundfield microphones," *IEEE Access*, vol. 8, pp. 87 749–87 761, 2020.

[23] N. Nasri, M. Rached, S. Chenini, and A. Kachouri, "3d indoor localization through a wireless acoustic sensor networks," *Progress In Electromagnetics Research B*, vol. 81, pp. 123–139, 2018.

[24] S. J. Orfanidis, "Electromagnetic waves and antennas," 2002.

[25] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 351–355.

[26] N. M. Drawil, H. M. Amar, and O. A. Basir, "Gps localization accuracy classification: A context-based approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 1, pp. 262–273, 2012.

[27] H. Wan, L. Wang, T. Zhao, K. Sun, S. Shi, H. Dai, G. Chen, H. Liu, and W. Wang, "Vector: Velocity based temperature-field monitoring with distributed acoustic devices," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 3, pp. 1–28, 2022.

[28] L. Wang, W. Wang, H. Dai, and S. Liu, "Magsound: Magnetic field assisted wireless earphone tracking," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 1, pp. 1–32, 2023.

[29] H. Wan, S. Shi, W. Cao, W. Wang, and G. Chen, "Resptracker: Multi-user room-scale respiration tracking with commercial acoustic devices," in *IEEE Conference on Computer Communications*, 2021, pp. 1–10.

[30] C. Cai, H. Pu, M. Hu, R. Zheng, and J. Luo, "Sst: Software sonic thermometer on acoustic-enabled iot devices," *IEEE Transactions on Mobile Computing*, vol. 20, no. 5, pp. 2067–2079, 2020.

[31] S. Ma, G. Wang, R. Fan, and C. Tellambura, "Blind channel estimation for ambient backscatter communication systems," *IEEE Communications Letters*, vol. 22, no. 6, pp. 1296–1299, 2018.

[32] 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.

[33] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[34] H. E. Bass, L. C. Sutherland, A. J. Zuckerwar, D. T. Blackstock, and D. Hester, "Atmospheric absorption of sound: Further developments," *The Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 680–683, 1995.

[35] J. Tao, L. Yang, and X. Han, "Enhanced carrier frequency offset estimation based on zadoff–chu sequences," *IEEE Communications Letters*, vol. 23, no. 10, pp. 1862–1865, 2019.

[36] J. M. Vera-Diaz, D. Pizarro, and J. Macias-Guarasa, "Acoustic source localization with deep generalized cross correlations," *Signal Processing*, vol. 187, p. 108169, 2021.

[37] T. Zhang, H. Phan, Z. Tang, C. Shi, Y. Wang, B. Yuan, and Y. Chen, "Inaudible backdoor attack via stealthy frequency trigger injection in audio spectrogram," in *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, 2024, pp. 31–45.

[38] "Respeaker 4-mic linear array kit for raspberry pi - seeed wiki," https://wiki.seeedstudio.com/, 2024.

[39] "Raspberry pi 3 model b," https://www.raspberrypi.com/products/raspberry-pi-3-model-b/, 2024.