MOM: Microphone based 3D Orientation Measurement

Zhihui Gao
Duke University
Durham, North Carolina, USA
zhihui.gao@duke.edu

Ang Li
Duke University
Durham, North Carolina, USA
ang.li630@duke.edu

Dong Li
University of Massachusetts Amherst
Amherst, Massachusetts, USA
dli@cs.umass.edu

Jialin Liu
University of Massachusetts Amherst
Amherst, Massachusetts, USA
jialinliu@umass.edu

Jie Xiong
University of Massachusetts Amherst
Amherst, Massachusetts, USA
jxiong@cs.umass.edu

Yu Wang
Tsinghua University
Beijing, China
yu-wang@tsinghua.edu.cn

Bing Li
Capital Normal University
Beijing, China
bing.li@cnu.edu.cn

Yiran Chen
Duke University
Durham, North Carolina, USA
yiran.chen@duke.edu

ABSTRACT

While a tremendous amount of effort has been devoted to localization, the orientation of a device, especially in 3D space, is seldom explored. Although many sensor-based methods utilizing gyroscope, accelerometer, and magnetometer have been proposed to measure 3D orientation, these methods generally suffer from high cumulative errors and performance degradation when the device is moving. In this paper, we present MOM, the first microphone-based system that estimates the 3D orientation of a device. The key idea of MOM is to employ free sound sources in our surrounding environment as anchors. The prior knowledge of these sound sources, including the signal waveform and the locations of the sound sources, is not required to be known. In particular, we propose an angle-of-arrival (AoA) extraction algorithm that compares fine-grained time delays over microphones at a low computational cost. We implement our system on three platforms including a 6-microphone array Seeed Studio ReSpeaker, a commodity earphone Sennheiser AMBEO smart headset and a commodity smartphone Google Pixel 4. Extensive experiments show that MOM can achieve significantly higher accuracy compared with status quo approaches and is robust against cumulative errors. We apply MOM to two real-life applications, i.e., head tracking and 3D reconstruction, to demonstrate the applicability and generality of MOM in practice.

1 INTRODUCTION

Real-time knowledge of location of people or objects has become essential for services in many fields including navigation [48], logistics [47], assembly line [17, 42], virtual reality (VR) [20] and health care [16]. A tremendous amount of effort has been devoted to improving the accuracy and enhancing the robustness of localization [6, 38, 45, 46]. However, much less attention has been paid to obtaining the orientation information of a device which is equally important in many real-life applications. For example, the orientation of the Xbox game controller plays a crucial role in providing users a rich experience in gameplay. In 3D reconstruction, multiple photos are taken at different locations and orientations. The orientation diversity presents critical information for 3D reconstruction.

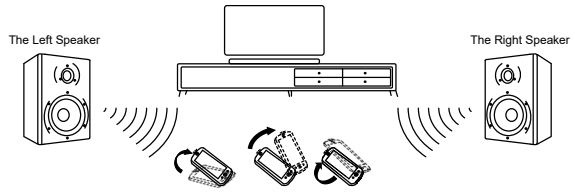


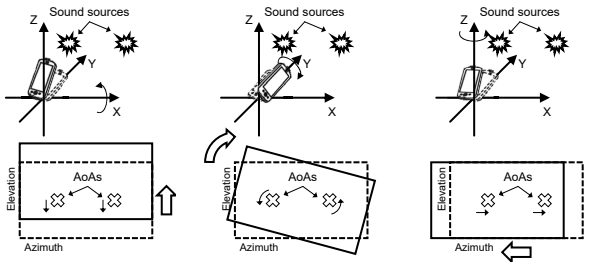
Figure 1: An example of MOM's application in the game room, where the device utilizes the sound from the stereo speakers for 3D orientation measurement.

Though closely related, orientation tracking is very different from localization in multiple aspects: (i) Location and orientation information are orthogonal to each other. When a user is using a smartphone, the location of the smartphone may remain unchanged while the orientation can vary significantly. (ii) A device's orientation is usually described in the 3D space while most localization systems only care about the device's spatial information in the azimuth 2D space. (iii) For a device with a small size, orientation can vary dramatically with a subtle movement. Therefore, the decimeter-granularity which is fine enough for most localization systems is too coarse for orientation estimation.

Traditionally, highly accurate 3D orientation information can be obtained from multiple high-speed cameras carefully deployed. Although the achieved accuracy is high, the extremely high cost is a major barrier hindering its wide adoption. In this work, we adopt such a camera-based system Qualisys [1] for ground-truth measurement. The cost of this Qualisys system is around $40,000 and it needs to be well calibrated so it cannot be easily moved to a new environment.

Using an inertial gyroscope is another way to measure the 3D orientation of a device. A gyroscope can measure the instant angular velocity of a device and the device's orientation can be obtained by taking the integral of the angular velocity. However, due to hardware noise, the orientation estimated by the gyroscope inside commodity devices is usually coarse and suffers from cumulative errors [36, 51]. The cumulative error can quickly reach 30° within a few minutes on a commodity smartphone [51].

978-1-6654-9624-7/22/$31.00 ©2022 IEEE
DOI 10.1109/IPSN54338.2022.00018



(a) Rotate by X axis **(b) Rotate by Y axis** **(c) Rotate by Z axis**

Figure 2: The AoA variations, including the elevation angles and the azimuth angles, of two sound sources when the device rotates around X (a), Y (b) and Z (c) axes.

To address this cumulative error, recent works utilize other sensors to either calibrate the readings from the gyroscope or measure the device's orientation directly [13, 14, 36, 42, 44, 51]. The most widely used sensors are accelerometer [14, 51] and magnetometer [13, 36, 51]. For the accelerometer, when the device's orientation varies, the gravitational acceleration at the X, Y and Z axes change accordingly. By taking the direction of the gravitational acceleration which is a constant as a reference, the orientation of the device can be estimated. However, this method only works when the device is static. When the device moves with non-zero acceleration, the measured acceleration is a summation of the gravitational acceleration and the device's moving acceleration. In this case, the gravitational acceleration cannot be extracted from the measured acceleration and used as a reference for orientation estimation. Similarly, a magnetometer can also be used to measure the geomagnetic field which can serve as a reference to infer the orientation of the device [36, 51]. A magnetometer can usually achieve a relatively good performance in an environment with no ferromagnetic material. However, a lot of daily-used electronic devices such as headphones and smartphones contain ferromagnetic material whose magnetic field can pollute the orientation estimates.

In this paper, we present MOM, a system that utilizes microphones widely available inside smart devices to measure their orientation. A lot of sound sources available in our surrounding environment can be used as anchors for device orientation estimation. Our proposed system has two unique advantages:

- We do not need to deploy a dedicated sound source to transmit controlled signals for orientation tracking.
- There is no need to know the location of the sound source. This makes our system flexible for real-world adoption.

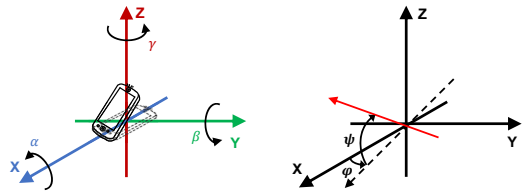
The basic idea of MOM is that when a device rotates, the angle-of-arrivals (AoAs) of signals from fixed sound sources change accordingly as illustrated in Fig. 2 and therefore the device orientation information can be obtained. As demonstrated in our experiments (Sec. 4), with just two microphones, the AoAs of the sound source can be extracted and the orientation of the device in 2D can then be obtained. With a third microphone, MOM can obtain the 3D orientation information. One observation favoring our design is that many existing smartphones are equipped with two microphones and some of them such as iPhone XR even have three built-in microphones. In addition, most smart speakers have 3-7 microphones [4]. We demonstrate that, without adding any extra hardware, the proposed microphone-based approach outperforms traditional inertial sensor-based orientation tracking approaches.

To extract the AoA of a sound source, we cannot directly adopt traditional AoA extraction schemes [24, 41]. There are two main categories of methods that can obtain the AoA of acoustic signals: phase-based and distance-based. Phase-based methods can be applied to not just acoustic signals but also RF signals. On the other hand, distance-based methods are only applicable to acoustic signals owing to the extremely low propagation speed in the air (340 m/s). In this work, we do not use any dedicated sound source to emit high-frequency ultrasound but utilize free acoustic sources in our environment and hence, we have no control of the sound sources. In this case, the frequency of the acoustic signal is usually low and varies frequently. Therefore, no stable phase readings can be obtained for phase-based AoA estimation and we adopt the distance-based method in our design. For distance-based methods, AoA estimation accuracy depends on the microphone sampling rate and the spacing between adjacent microphones. Specifically, a higher sampling rate presents more fine-grained distance measurements. Larger spacing between the microphones means a longer extra distance for the signal to reach the second microphone and a more accurate estimation of the extra distance. The higher accuracy of distance estimation implies a more accurate AoA estimation.

Hereby, we design an AoA extraction algorithm in Sec. 3.2 by measuring the distance difference of two microphones from the sound source, which is equivalent to measuring the time delay of the signals arriving at the two microphones. To capture this delay, we adopt the trial and error method. We try a possible AoA and calculate the corresponding delay. We then shift one signal by this corresponding delay and compare the similarity between the shifted signal and the other signal using correlation. When the correct AoA is tried, the correlation generates a high peak.

Note that we are not able to shift a signal by an arbitrarily small amount in the time domain due to the sampling rate limit. We address this issue based on one key observation: the signal frequency of sound sources in the surrounding environment has a much lower frequency (below 8 kHz) compared to those ultrasound signals (18-21 kHz) adopted in existing works. Therefore, the Nyquist frequency is much lower than the sampling rate supported by commodity smartphones (48 kHz). Thus, we can adopt the well-known Whittaker-Shannon interpolation formula (WS formula) [43] to interpolate signal sample points for more fine-grained time shifts and more accurate AoA estimation accordingly. However, the WS formula is a signal processing scheme that incurs a high computational cost. To address this issue, we only calculate the correlations for a few AoAs and estimate the correlations for other AoAs based on the observation that the correlations change smoothly. In this way, the computation cost is greatly reduced and the high accuracy of AoA estimation is still retained.

We also identify a practical issue associated with the proposed system. As we do not use any dedicated sound source, the uncontrolled signals can vary dramatically in reality. Fortunately, in a typical environment, usually more than five sound sources can be identified. For 3D orientation tracking, two sound sources are required and for 2D tracking, one sound source suffices. We therefore



(a) Euler angle (b) Spatial angle

Figure 3: The spatial illustrations for Euler angle (a) and spatial angle (b).

can select the most reliable sound sources among many available ones in the surrounding environment for orientation tracking.

We summarize our contributions as follows:

- We involve orientation tracking into the ecosystem of acoustic sensing. We exploit microphones that are widely available in commodity devices and free sound sources in our surrounding environment to track the orientation of a device.
- We adopt a distance-based method for AoA estimation and effectively utilize the signal's lower-frequency property to achieve accurate AoA estimation without incurring a high computational cost.
- The proposed system is flexible to be deployed in real-world settings. It works with various sound sources in the environment such as a vibrating razor and it does not need to know the locations of the sound sources.
- We implement MOM on three platforms, a 6-microphone array Seeed Studio ReSpeaker, a commodity earphone Sennheiser AMBEO headset and a commodity smartphone Google Pixel 4. Comprehensive experiments demonstrate the robustness of the proposed schemes under various conditions.

2 PRELIMINARIES

In this section, we present the preliminary knowledge of 3D orientation related to our design.

2.1 3D Orientation and Rotation

A device's orientation is usually measured in 3D space, and the orientation of an object has 3 degree-of-freedoms (DoFs). This means the orientation of a target can be fully characterized by a minimum of 3 variables. Similarly, the rotation of an object, i.e., the difference between two orientations, also has 3 DoFs. The rest of this section briefly introduces two different representation schemes which are widely used to describe 3D orientation and rotation: *Euler angle* and *rotation matrix*.

Euler Angle. As Fig. 3(a) illustrates, the Euler angle uses a vector with 3 angles to describe the 3D orientation of a device:

$$[\alpha, \beta, \gamma], \tag{1}$$

where α, β and γ are the angles that the object rotates around its X, Y and Z axis respectively. Note that the order matters and different axis orders can bring different results. In this paper, we use the intrinsic X-Y-Z sequence for Euler angle representation, where the object rotates around its X, Y and Z axis successively. Euler angle is a visually intuitive representation to describe the orientation.

Rotation Matrix. A rotation matrix $\mathbf{R}$ is a 3 by 3 matrix that satisfies the following equations:

$$\mathbf{R}\mathbf{R}^T = \mathbf{I}, |\mathbf{R}| = 1. \tag{2}$$

One vector point in the 3D space can be represented in different coordinates (e.g., the world reference coordinate C^W and the device reference coordinate C^D in our system). The 3D representations of the vector point in C^W and C^D are denoted as $\mathbf{v}^W$ and $\mathbf{v}^D$ respectively. We can then use a rotation matrix from the world reference coordinate C^W to the device reference coordinate C^D, $\mathbf{R}_{WD}$, to characterize the relationship between the two coordinates and transfer a vector point from one coordinate to the other.

$$\mathbf{v}^D = \mathbf{R}_{WD}\mathbf{v}^W. \tag{3}$$

The rotation matrix is a convenient representation for calculation and we use this representation in formulas.

2.2 3D Direction

The 3D direction describes the direction of the signal source (transmitter) to the receiver. 3D Direction is usually represented by the spatial angle or the direction vector.

Spatial Angle. Different from 3D orientation, 3D direction can be fully characterized by two parameters and has only 2 DoFs. In this paper, we define spatial angle as a combination of an elevation angle ψ and an azimuth angle φ, shown in Fig. 3(b). For our system, the upper semi-space is symmetrical to the lower semi-space and we only present the information of upper semi-space. Thus, the range of the elevation angle ψ goes from 0° to 90° and the range of the azimuth angle φ goes from 0° to 360°.

Direction Vector. A direction vector $\mathbf{n}$ is a 3D vector whose norm is 1. The transformation from the spatial angle (ψ, φ) to the direction vector $\mathbf{n}$ is as below:

$$\mathbf{n} = [cos\psi cos\varphi, cos\psi sin\varphi, sin\psi]^T. \tag{4}$$

Similar to the rotation matrix, the direction vector is convenient to be used in formulas.

3 SYSTEM DESIGN

In this section, we first present the system overview following by the details of each design component.

3.1 System Overview

There are three modules in the proposed system, as shown in Fig. 4: AoA extraction, sound source selection and orientation estimation.

AoA Extraction. AoA extraction module extracts the AoAs of the sound sources in the environment. The input of this module is the acoustic signals $s_1(t), s_2(t), \ldots, s_N(t)$ recorded by N microphones over time. The received acoustic signal is a mixture of multiple uncontrolled sound sources in the environment. The AoA extraction algorithm divides recorded acoustic signals into time windows $\mathcal{T}_i$ and examines each time window to obtain the AoA information. The instant AoA can be assumed as unchanged within a small time window (e.g., 10 ms). In each time window, the AoA extraction algorithm extracts AoAs $(\psi_1, \varphi_1), (\psi_2, \varphi_2), \ldots, (\psi_M, \varphi_M)$ of the M signal sources in the environment. Each AoA of a sound

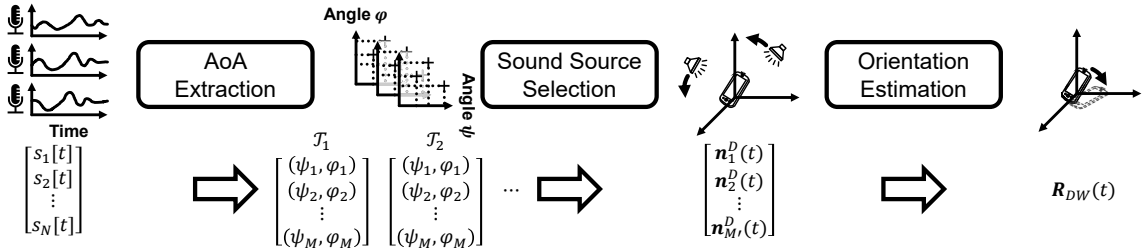
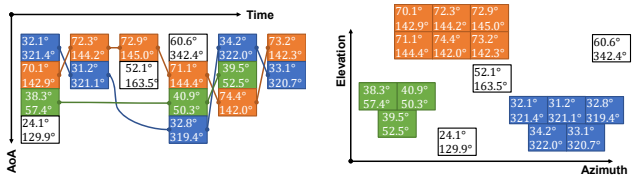


Figure 4: The system overview of MOM.

source is a spatial angle that contains the elevation angle ψ and the azimuth angle φ.

Sound Source Selection. Note that we do not use any dedicated



(a) 2-Dimension clustering (b) Clustering results

Figure 5: The 2-dimension clustering algorithm that recognizes two reliable sound sources (blue and orange), one unreliable sound source (green) and other outliers (white).

sound sources and utilize existing sound sources already in the environment for AoA estimation. The sound sources in the real world are not under our control and can vary dramatically. Some of the sound sources appear to be reliable and emit out stable acoustic signals whose AoAs can be extracted over multiple time windows, while signals from some other sound sources can be intermittent. In this module, we select M' reliable AoAs among the M AoAs extracted in each time window. For the M' reliable sound sources, their AoA over time t can be represented as norm vectors: $\mathbf{n}_1^D(t), \mathbf{n}_2^D(t), \ldots, \mathbf{n}_{M'}^D(t)$.

Sound source selection is achieved through a 2-dimensional clustering scheme as shown in Fig. 5(a). We measure the AoAs of detectable sound sources in multiple consecutive time windows. These AoAs are then clustered based on their spatial proximity (AoA similarity). Each cluster represents a sound source. Then we use the size of the cluster, i.e., the number of AoAs within the cluster, to determine if the sound source is reliable as shown in Fig. 5(b). The cluster with a small size, such as the green cluster, means that it emits out intermittent signals and should be excluded from our choice. The two clusters (blue and orange) are chosen as the reliable sound sources to estimate the target orientation.

Orientation Estimation With the selected M' reliable sound sources $\mathbf{n}_i^D(t)$, this module jointly predicts the final 3D orientation $\mathbf{R}_{DW}(t)$ of the device. Specifically, we develop a gradient descent algorithm to make the estimated rotation matrix $\mathbf{R}_{DW}(t)$ fit the observed $\mathbf{n}_i^D(t)$ best. This estimated 3D orientation $\mathbf{R}_{DW}(t)$ is the final output of MOM.

3.2 AoA Extraction

We design a novel AoA extraction algorithm in MOM. This algorithm contains three steps: signal similarity measuring, optimal AoA searching and multiple AoA detaching. In the signal similarity measuring step, we introduce how we construct the optimal function $F(\psi, \varphi)$ to measure signals' similarity at a given AoA (ψ, φ). The optimal AoA searching step efficiently finds the optimal AoA that maximizes $F(\psi, \varphi)$. Note that in the first two steps, we only consider one AoA. In reality, there can be multiple sound sources. In the last step, we explain how to extract the AoAs of multiple sources when signals are mixed at the receiver. Specifically, we employ the first two steps to identify the first AoA. We prove in Sec. 3.2.3 the AoA extraction scheme for a single-source is still effective in the multiple-source scenario. Then, we remove the component of the first extracted signal in the mixed signal. This removal does not affect the AoA estimation of the remaining sound sources. In this way, the AoAs of multiple sources can be extracted one by one.

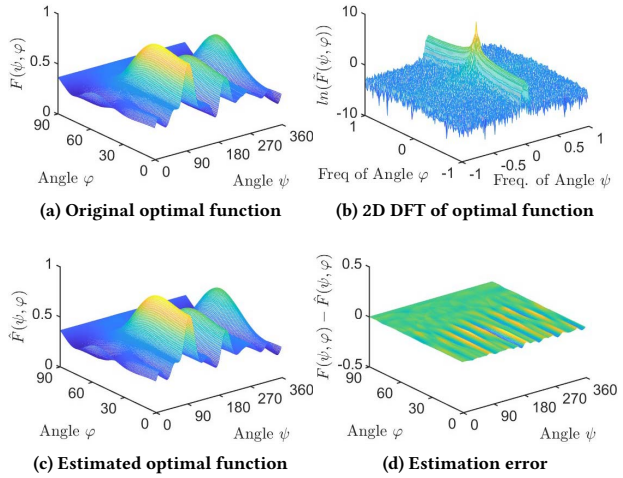
3.2.1 Signal Similarity Measuring. In signal similarity measuring, we employ an optimal function $F(\psi, \varphi)$ to evaluate the likelihood of an assumed AoA (ψ, φ). Given an assumed AoA (ψ, φ), we take the origin of the device's reference coordinate as the reference and calculate the time delays at all the microphones. Note that during the rotation process, the delays can be positive or negative. Then, we shift the discrete acoustic signals in a fine-grained manner with the help of the WS formula. Given an AoA (ψ, φ), the direction vector of the sound source $\mathbf{n}^D$ in the device reference coordinate can be expressed as:

$$\mathbf{n}^D = [cos\psi cos\varphi, cos\psi sin\varphi, sin\psi]^T. \tag{5}$$

Then, the delay τ_i of the ith microphone located at (x_i, y_i, z_i) on the device can be expressed as:

$$\tau_i = -\frac{1}{v_s}\mathbf{n}^D \cdot [x_i, y_i, z_i]. \tag{6}$$

where v_s is the velocity of sound. Since the amount of delay may not exactly equal to an integer number of sampling intervals, we adopt the WS formula [43] to estimate the shifted signal in a finer-grained manner without being constrained by the sampling interval. Specifically, we denote the original discrete signal at the ith microphone as $s_i[j]$. The amount of shift is $\tau_i \cdot f_s$ and the shifted signal at sample j can be written as $\hat{s}_i[j + \tau_i \cdot f_s]$, where f_s is the sampling frequency. According to WS formula, its signal value at $j + \tau_i \cdot f_s$ is calculated as the sum of infinite number of weighted terms of $s_i[\cdot]$. In our implementation, we approximate the calculation by including the



(a) Original optimal function (b) 2D DFT of optimal function

(c) Estimated optimal function (d) Estimation error

Figure 6: The optimal function $F(\psi, \varphi)$ is smooth (a) and its energy of 2D DFT is mainly located at low frequency (b). Thus, we can selectively calculate the points in the optimal function $F(\psi, \varphi)$ and estimate other points (c) with low errors (d).

nearest $2N_{WS}+1$ discrete samples around sample j. The choice of N_{WS} is discussed in Sec. 4.4. The mathematical representation of the above procedure is presented as:

$$\begin{aligned}\hat{s}_i[j+\tau_i \cdot f_s] &= \sum_{k=-\infty}^{+\infty} s_i[k] \cdot sinc(j+\tau \cdot f_s - k) \\ &\approx \sum_{k=j-N_{WS}}^{j+N_{WS}} s_i[k] \cdot sinc(j+\tau \cdot f_s - k).\end{aligned} \tag{7}$$

By now, the shifted signals can be obtained and we can calculate the similarity between the shifted signal and the other original signal. If the assumed AoA (ψ, φ) is correct, the two signals should be highly correlated and a peak can be generated with a simple correlation operation. To make the signals equally weighted across multiple microphones, we calculate the overall correlation between each pair of the N signals. It can be proved that such operation is equivalent to calculating the sum of the correlations between each signal and the mean signal $\bar{\hat{s}}(t)$:

$$F(\psi, \varphi) = \sum_{i=1}^{N} \mathrm{Corr}_{\mathcal{T}}\left[\hat{s}_i[j+\tau_i \cdot f_s], \bar{\hat{s}}[j]\right], \tag{8}$$

where the mean signal $\bar{\hat{s}}$ is the mean of shifted signals over N microphones:

$$\bar{\hat{s}}[j] = \frac{1}{N}\sum_{i=1}^{N} \hat{s}_i[j+\tau_i \cdot f_s]. \tag{9}$$

3.2.2 Optimal AoA Searching. Optimal AoA searching searches the AoA which maximize $F(\psi, \varphi)$ at a low computational cost. As defined in Sec. 2.2, the search range is 90° and 360° in the elevation angle space and azimuth angle space respectively.

The challenge here is that the WS formula for estimating the shifted signal is computationally heavy. The key observation enabling us to reduce the computational cost is that the original optimal function $F(\psi, \varphi)$ is smooth over the solution space as shown in Fig. 6(a). We further observe that the energy of this $F(\psi, \varphi)$ is concentrated in the low frequency range. As shown in Fig. 6(b), after applying a 2D discrete Fourier transform (2D DFT) on $F(\psi, \varphi)$, 99.9% of the energy is concentrated on the 0.6% of the low frequency part. This means we can consider the optimal function $F(\psi, \varphi)$ as a 2D low frequency signal and employ very sparse sampling to fully recover the optimal function. Therefore, we do not need to calculate the optimal function $F(\psi, \varphi)$ by WS formula for all the ψ, φ values. Instead, we can just calculate a few and estimate the other values based on the sparse sampling property, significantly reducing the computational cost.

Specifically, we calculate a total of $N_\psi \times N_\varphi$ points based on the WS formula. We denote the calculated points as $F[i, j]$, where i ranges from 1 to N_ψ and j ranges from 1 to N_φ. We can then deduce the other values of the optimal function as $\hat{F}(\psi, \varphi)$ with the $N_\psi \times N_\varphi$ points obtained above. When deducing the value of $\hat{F}(\psi, \varphi)$, we should first localize the nearest calculated point, denoted as $F[n_{\psi 0}, n_{\varphi 0}]$. Then we can estimate the value with the nearest $(2N_{near}+1) \times (2N_{near}+1)$ sampling points as:

$$\begin{aligned}\hat{F}(\psi, \varphi) &= \sum_{i=n_{\psi 0}-\infty}^{n_{\psi 0}+\infty} \sum_{j=n_{\varphi 0}-\infty}^{n_{\varphi 0}+\infty} F[i, j] \cdot sinc(\psi - i) \cdot sinc(\varphi - j) \\ &\approx \sum_{i=n_{\psi 0}-N_{near}}^{n_{\psi 0}+N_{near}} \sum_{j=n_{\varphi 0}-N_{near}}^{n_{\varphi 0}+N_{near}} F[i, j] \cdot sinc(\psi - i) \cdot sinc(\varphi - j).\end{aligned} \tag{10}$$

The empirical value for N_{near} is 20 which ensures small errors in the optimal AoA searching. Fig. 6(c) shows the estimated optimal function $\hat{F}(\psi, \varphi)$ by only calculating 10×20 points in the $90° \times 360°$ solution space. Its overall error is only 1.14%, as shown in Fig. 6(d). A more detailed evaluation of selected point number on system performance is presented in Sec. 4.4

The computational cost of estimating the optimal function by Eq. 10 is much lower than calculating the true values by Eq. 8. When calculating the true value of the optimal function, we need to repeat the WS formula for all the samples of the acoustic signals within the time window $\mathcal{T}$ for all the N microphones, whose complexity is $O(N\mathcal{T}f_s \cdot N_{WS})$. For the estimation algorithm here, we only need to perform Eq. 10 once with a much smaller complexity of $O(N_{near}^2)$.

To further reduce the computational costs, we use gradient descent instead of searching the whole solution space of $\hat{F}(\psi, \varphi)$. This is because those points $F[n_\psi, n_\varphi]$ provide an overview of $F(\psi, \varphi)$, which helps avoid the local maximums. We initialize the gradient descent at the AoA with the maximum value of $F[n_\psi, n_p]$. Then, we estimate the gradient of the estimated optimal function $\hat{F}(\psi, \varphi)$ according to Eq. 10. After taking gradient descent for several steps, we output the optimal AoA.

3.2.3 Multiple AoA Detaching. So far, we discuss how our AoA extraction algorithm deals with a single sound source. Next, we

illustrate how our algorithm detaches multiple sound sources by iteration. We prove the optimal function is still effective when there are M sound sources. In another word, the AoA with the maximum value of $F(\psi, \varphi)$, in this case, is still the AoA of one of the M sound sources. To simplify the notation, we use the symbol $s_i(t)$ to represent the signal, where t is the time, different from $s_i[j]$ in Sec. 3.2.1, where j is the sample index. We assume the received acoustic signal at the ith microphone $s_i(t)$ is the superposition of M sound sources $s^j(t)(j = 1..M)$ coming from different AoAs:

$$s_i(t) = \sum_{j=1}^{M} s^j(t + \tau_i^j) + n_i(t), \tag{11}$$

where $n_i(t)$ is the additive white Gaussian noise. When we correctly predict the AoA for the 1st sound source with (ψ^1, φ^1), i.e., we correctly predict its delay $\tau_i = \tau_i^1$, we can write $F(\psi, \varphi)$ as:

$$F(\psi,\varphi) = \sum_{i=1}^{N} \Bigg[\sum_{t\in\mathcal{T}} \Big(s^1(t) + \sum_{j=2}^{M} s^j(t + \tau_i^j - \tau_i^1) + n_i(t)\Big) \cdot \Big(s^1(t) + \sum_{j=2}^{M} \frac{1}{N} \sum_{k=1}^{N} s^j(t + \tau_k^i - \tau_k^1) + \bar{n}(t)\Big) \Big/ \sqrt{\sum_{t\in\mathcal{T}} s_i(t)^2 \cdot \sum_{t\in\mathcal{T}} \bar{s}(t)^2} \Bigg] \tag{12}$$

The sound sources and noise are independent. Besides, a signal's auto-correlation at 0 is larger than other locations. We then have:

$$F(\psi,\varphi) \approx \sum_{i=1}^{N} \frac{\sum_{t\in\mathcal{T}} s^1(t) \cdot s^1(t)}{\sqrt{\sum_{t\in\mathcal{T}} s_i(t)^2 \cdot \sum_{t\in\mathcal{T}} \bar{s}(t)^2}}. \tag{13}$$

That is to say, only the sound source whose AoA is correctly predicted can make the $F(\psi, \varphi)$ explicitly greater than 0. Therefore, we can still use this optimal function to find the AoA in the multiple sound sources scenario. Thus, we can detach multiple AoAs by iteration. We first find the AoA (ψ_1, φ_1). Then we remove this AoA's effect on $s_i(t)$ by:

$$s_i^{\text{new}}(t) = s_i^{\text{old}}(t) - \text{Corr}_{\mathcal{T}}\big[s_i(t + \tau_i^1), \bar{s}(t)\big] \cdot \bar{s}(t). \tag{14}$$

After removing its effect, we find the next optimal AoA (ψ_2, φ_2) and repeat the process. The removal of a sound source decreases the total energy in $s_i(t)$. We repeat the extraction process until there is no explicit energy left in $s_i(t)$.

3.3 Sound Source Selection

In sound source selection, we select reliable sound sources by checking the stability of the AoAs extracted over time. This process is critical because unreliable sound sources can cause severe performance degradation in the final orientation estimation. Existing works [18, 30] adopt bipartite matching to identify stable AoAs over time. In these works, a bipartite matching algorithm is executed to match similar AoAs over different time windows. However, these algorithms do not work well on matching AoAs in our work, in which dedicated sound sources are not used. In reality, burst noise can appear suddenly and frequently, inducing unexpected temporary AoAs. Besides, sound sources can have random pauses over time, such as human voice. Hence, instead of using all sound sources, we select reliable ones as anchors for orientation estimation.

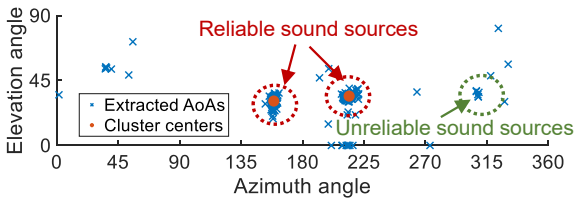


Figure 7: The clustering results of AoAs, where there are 2 reliable sound sources marked as red circles and several unreliable sound sources.

3.3.1 2-Dimensional Clustering. We propose a 2-dimensional clustering scheme to deal with the above issues, as shown in Fig. 7. To extract reliable AoAs, we cluster all AoAs extracted in the consecutive windows based on AoA proximity. The AoAs of a reliable sound source appear in most of the time windows and form a cluster of a larger size. On the other hand, those AoAs generated by sudden noise or unreliable sound sources usually cannot or just form a smaller cluster. Thus, we can perform clustering to select reliable sound sources which have a large size. Specifically, the density-based spatial clustering of applications with noise (DBSCAN) algorithm [11] is adopted for our clustering.

While the principle is simple, there are two trade-offs to be considered. First, how many sound sources M' should be chosen for orientation estimation? Generally, more sound sources provide richer information for orientation estimation. However, too many sound sources can also degrade the performance. This is because the more sound sources we use, the more likely we include an unreliable sound source in our estimation. One single unreliable source can lead to large errors, overshadowing the benefit of including more sound sources. Based on our experiments, $M' = 2$ or 3 is good for accurate orientation estimation. When the number of sound sources M' is increased to four, the average orientation accuracy starts decreasing. Second, how many time windows should be utilized to perform the clustering operation? We can be more confident to judge if a sound source is reliable or not with more time windows. Also, it is less likely to discard the sound source when it is justly occasionally muted. However, when the device being tracked is rotating quickly, the AoAs from a sound source vary in a large range. The larger number of time windows can further increase the AoA variation range. This large AoA variation can confuse the clustering algorithm when two sound sources are close to each other and two sources can be wrongly clustered as one single source.

3.4 Orientation Estimation

In orientation estimation, we jointly estimate the final 3D orientation over time using the M' reliable sound sources selected. Now the direction vectors of reliable sound sources in the device reference coordinate $\mathbf{n}_i^D$ is calculated, we can estimate the device's 3D orientation $\mathbf{R}_{DW}$ in the world reference coordinate. According to Eq. 3, we can transform each $\mathbf{n}_i^D$ to the corresponding $\mathbf{n}_i^W$ in the world reference coordinate as below:

$$\mathbf{n}^W = \mathbf{R}_{DW}\mathbf{n}^D. \tag{15}$$

The M' sound sources can form M' equations above, which exceeds the DoF of the unknown variable $\mathbf{R}_{DW}$. There is no such $\mathbf{R}_{DW}$ that satisfies all these equations. Thus, we consider the loss function $L(\mathbf{R}_{DW})$ that minimizes the overall squared errors in these equations:

$$L(\mathbf{R}_{DW}) = \sum_i ||\mathbf{R}_{DW}\hat{\mathbf{n}}_i^D - \hat{\mathbf{n}}_i^W||^2. \tag{16}$$

To minimize $L(\mathbf{R}_{DW})$, the usual way is to calculate its derivative and take gradient descent. However, rotation matrices cannot be directly differentiated. Hereby, we conduct a transform from the Lie Group to Lie Algebra [7], where the rotation matrix $\mathbf{R}_{DW}$ in Lie Group becomes a corresponding vector ϕ in Lie algebra as:

$$L(\phi) = \sum_i ||exp(\phi^+)\hat{\mathbf{n}}_i^D - \hat{\mathbf{n}}_i^W||^2 \tag{17}$$

where ϕ^+ is the screw matrix of a 3D vector ϕ. Here in Lie algebra, ϕ can be differentiated by perturbation model. The differentiated loss function is:

$$\begin{aligned}\frac{\partial L}{\partial \varphi} &= \sum_i 2\big[exp(\phi^+)\hat{\mathbf{n}}_i^D - \hat{\mathbf{n}}_i^W)\big] \cdot \frac{\partial (exp(\phi^+)\hat{\mathbf{n}}_i^D)}{\partial \varphi} \\ &= \sum_i 2\big[exp(\phi^+)\hat{\mathbf{n}}_i^D - \hat{\mathbf{n}}_i^W)\big] \cdot \big[-(exp(\phi^+)\hat{\mathbf{n}}_i^D)^+\big].\end{aligned} \tag{18}$$

Now, we can apply gradient descent on it. That is, we can update ϕ at the learning rate of r as:

$$\begin{aligned}\phi^{\text{new}} &= \phi^{\text{old}} - r \cdot \frac{\partial L(\phi)}{\partial \varphi} \\ &= \phi^{\text{old}} - \sum_i 2r\big[exp(\phi^+)\hat{\mathbf{n}}_i^D - \hat{\mathbf{n}}_i^W)\big] \cdot \big[-(exp(\phi^+)\hat{\mathbf{n}}_i^D)^+\big].\end{aligned} \tag{19}$$

Then, we conduct an inverse transform from Lie Algebra to Lie Group. The new equation that updates the rotation matrix $\mathbf{R}_{DW}$ is:

$$\mathbf{R}_{DW}^{\text{new}} = exp\Big[(\sum_i 2r\big[\mathbf{R}_{DW}^{\text{old}}\hat{\mathbf{n}}_i^D - \hat{\mathbf{n}}_i^W)\big] \cdot \big[-\mathbf{R}_{DW}^{\text{old}}\hat{\mathbf{n}}_i^D)^+\big])^+\Big] \cdot \mathbf{R}_{DW}^{\text{old}}. \tag{20}$$

Hereby, we can estimate the device's 3D orientation $\mathbf{R}_{DW}(t)$ over time. Specifically, we conduct a gradient descent operation to estimate its instant 3D orientation $\mathbf{R}_{DW}$ using Eq. 20, initializing $\mathbf{R}_{DW}^{\text{old}}$ with the 3D orientation we previously estimated.

4 EVALUATION

We implement our system design on three devices, including a 6-microphone array, a commodity earphone and a commodity smartphone. Our exhaustive evaluation reveals MOM's advantages over state-of-the-art systems. Moreover, we showcase the usage of the developed system using two applications, earphone-based head tracking and smartphone-based 3D reconstruction.

4.1 Implementation

We implement MOM on three platforms: Seeed Studio ReSpeaker, Sennheiser AMBEO smart headset, and Google Pixel 4 smartphone. For performance comparison, we also implement the start-of-the-art inertial sensor-based schemes [36, 51] on an inertial measurement unit (MPU 6050 [5]). The ground truth is measured by the camera-based motion capture system, Qualisys [1]. We briefly introduce each platform below.

Seeed Studio ReSpeaker. Seeed Studio ReSpeaker [2] is a 6-microphone circular array on Raspberry Pi 3b+ as shown in Fig. 8(a). The layout of ReSpeaker's microphones is the same as that of Amazon Echo [4]. We can capture raw acoustic signals at the six microphones through the Raspberry Pi's Linux terminal. Its sampling rate is 48 kHz on each microphone.

Sennheiser AMBEO Smart Headset. Sennheiser AMBEO smart headset [3] is a commodity earphone with one microphone in each earbud as shown in Fig. 8(b). These two microphones support in-ear binaural audio recording at a sampling rate of 48 kHz. We implement 2D MOM for head tracking on AMBEO and the tracking performance outperforms the state-of-the-art AirPods Pro's head tracking using the inertial sensor (gyroscope).

Google Pixel 4. Google Pixel 4 is a commodity smartphone with two built-in microphones as shown in Fig. 8(c). We use Pixel 4 to demonstrate 3D image reconstruction.

MPU 6050. We use the gyroscope, accelerometer and magnetometer within the MPU 6050 to implement the state-of-the-art inertial sensor-based systems for comparison. MPU series are widely used nowadays in smartphones and wearable devices [36]. It provides 9-axis angular velocity, acceleration and magnetic readings at a sampling rate of 10 Hz.

Motion Capture System Qualisys. Qualisys [1] is a camera-based motion capture system as shown in Fig. 8(d). This system captures an object's location and orientation using 10 cooperated high-speed cameras. Qualisys is able to achieve a sub-millimeter displacement measurement accuracy and an orientation error below 0.6 °. We use the measurements from Qualisys as ground truths.

4.2 State-of-the-art Sensor Based Systems

We compare the performance of MOM with three start-of-the-art inertial sensor based baselines: Gyro, A3 [51] and MUSE [36].

Gyro. This baseline only utilizes the gyroscope that measures angular velocity, which is the most common method for orientation measurement on commodity devices. Given an initial orientation, it estimates the device's 3D orientation by accumulating angular change over time.

A3. A3 estimates 3D orientation using the accelerometer and the magnetometer. A3 first estimates the device's Euler angle on the X axis and Y axis with the accelerometer. Then the orientation on the Z axis is obtained from the magnetometer.

MUSE. MUSE measures 3D orientation in two phases: a static phase and a dynamic phase. When the device is static and the accelerometer can be trusted, MUSE is in the static phase and estimates the 3D orientation using the accelerometer and the magnetometer. In the dynamic phase, MUSE estimates 3D orientation using the gyroscope and calibrates the cumulative errors using the magnetometer. The principle in the static phase is similar to A3, so we only consider MUSE's dynamic phase in our evaluation.

4.3 Evaluation Metrics

We utilize axis-angle error to measure the orientation discrepancy between the estimated orientation and the ground truth. Specifically, we consider static errors and dynamic errors respectively to evaluate the orientation estimation performance of each method.

(a) ReSpeaker & MPU 6050 (b) AMBEO (c) Pixel 4 (d) Qualisys

Figure 8: The three platforms where MOM is implemented: the 6-microphone array ReSpeaker (a), the commodity earphone AMBEO (b) and the commodity smartphone Pixel 4 (c). The motion capture system Qualisys (d) provides the ground truth.

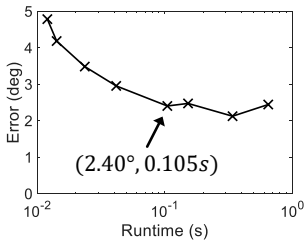


Figure 9: MOM's trade-off between AoA errors and the computational costs under different settings.

Axis-Angle Error. We use the axis-angle discrepancy between the estimated 3D orientation and the ground truth to represent the error. This error is the minimal angle needed to rotate the estimated orientation to the ground truth. This error ranges from 0° to 180°, where 0° refers to the same orientation and 180° refers to the opposite orientation.

Static Error. The static error refers to the orientation error when the device is static. Its application includes estimating the orientation of a camera in 3D reconstruction.

Dynamic Error. The dynamic error is measured as the device is moving and rotating. In this work, the performance of the head tracking application is evaluated using this dynamic error metric.

4.4 Comparison with the State-of-the-Arts

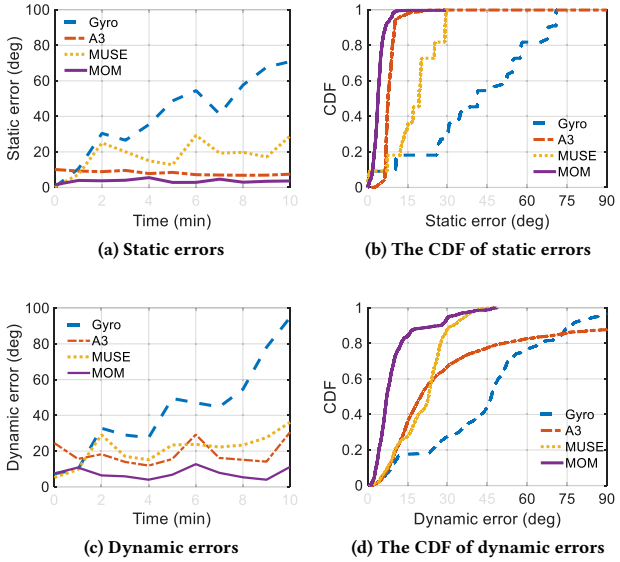
We first evaluate the performance of MOM's AoA extraction algorithm in terms of accuracy and computational cost against a state-of-the-art system Symphony [41]. Then, we compare MOM's overall performance on 3D orientation estimation with three baselines, Gyro, A3 and MUSE introduced in Sec. 4.2.

4.4.1 AoA Extraction. AoA accuracy is measured by the axis-angle error and the computational cost is measured as the averaging runtime on a laptop per AoA extraction. As mentioned in Sec. 3.2.2, there is a trade-off between the AoA accuracy and the computational cost in MOM. To evaluate this trade-off, we evaluate the performance of MOM under eight different settings. As shown in Fig. 9, we can see that when the computational cost is around 0.105 s, we maintain a good balance between accuracy and computational cost. When we further increase the number of samples, the computational cost is increased with a marginal accuracy increase.

We implement a state-of-the-art AoA extraction system, Symphony [41] for performance comparison. Symphony localizes a sound source using a single microphone array by leveraging the wall's reflection to create a virtual array. In doing so, the AoA extraction algorithm in Symphony can distinguish the AoAs of the light-of-sight path and the wall-reflection path. Its 2D AoA error is 5.09° on elevation (ψ) and 7.80° on azimuth (φ), respectively. These results are consistent with the result (4.2° on elevation) reported [41]. The detailed comparison between Symphony and MOM is shown in Table 1, where MOM outperforms Symphony on both AoA estimation accuracy and computational cost.

Table 1: AoA extraction algorithm comparison.

Algorithm	ψ Error	φ Error	3D AoA Error	Runtime
Symphony	5.09°	7.80°	9.31°	0.610s
MOM	1.19°	2.09°	2.40°	0.105s



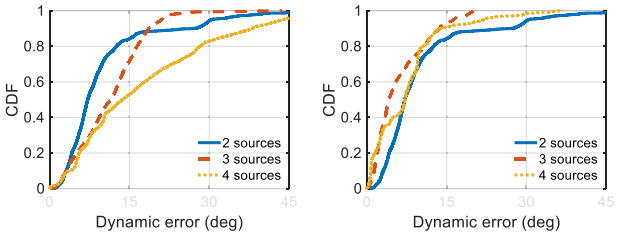
(a) Static errors (b) The CDF of static errors

(c) Dynamic errors (d) The CDF of dynamic errors

Figure 10: The overall performance comparison among different approaches within 10 minutes.

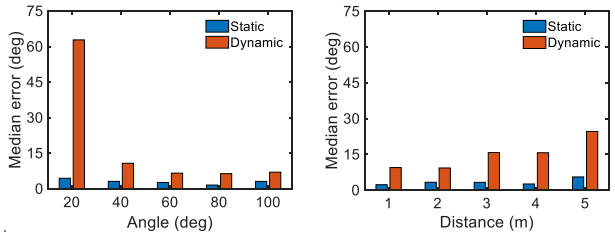
4.4.2 Overall Performance. We implement MOM on ReSpeaker and the three baselines on MPU 6050. We use a pair of stereo speakers as two sound sources. The two speakers are located around one meter away from the target device and the received volume is 63.5 *dB*. In this experiment environment, the volume of the background noise is around 43 dB. The two speakers' elevation angle with respect to the target device is around 30° and their relative azimuth angle is around 60°. We also conduct experiments with other sound sources commonly found in real life such as razors and human voices. The results are presented and discussed in Sec. 4.6.

We compare the performance of MOM with three baselines for 10 minutes. For the static case, ReSpeaker stays static in the first five seconds of each minute. For the rest of the time, the device rotates



(a) Close to each other (b) Far away from each other

Figure 11: The cumulative density functions of dynamic errors when utilizing 2, 3 and 4 sound sources.



(a) Impact of angles (b) Impact of distances

Figure 12: The median static and dynamic errors of MOM over different angles (a) and distances (b) of sound sources.

following a pre-defined pattern. We measure the errors for 5 seconds at the beginning of each minute. For the dynamic case, the device keeps rotating all the time for 10 minutes. The static errors over time and the cumulative density function (CDF) of the four systems are shown in Fig. 10(a)-(b). Over the 10 minutes, Gyro suffers from a severe cumulative error. The error increases at a rate of around 7° per minute. After 10 minutes, Gyro's static errors increase to 71°. Compared to Gyro, MUSE calibrates out the cumulative errors with the magnetometer. The errors fluctuate between 20° to 30° after 2 minutes. Note that when the device is static, the accelerometer can be trusted. However, A3 still has a constant static error around 8.12°. The reason is that in an indoor environment the geomagnetic field is polluted, leading to a constant angle measurement error on A3's Z axis. MOM achieves a significantly smaller error over the 10 minutes, i.e., a mean static error of 2.44°.

As for the dynamic case, the results are shown in Fig. 10(c)-(d). The performance of Gyro and MUSE is slightly worse than that in the static case. By the end of 10 minutes, the error of Gyro reaches 94.75° while the error of MUSE is 36.10°. When the device is dynamic, in addition to the constant error from the magnetometer, the accelerometer can not provide the correct direction of the gravitational acceleration and brings in extra errors to A3. The median error of MOM in the dynamic phase is 5.26°, slightly higher than that in the static phase. The possible reason for this error increase is that the AoAs vary in a larger range and bring errors.

Overall, MOM achieves higher accuracy than the state-of-the-art systems and it does not suffer from cumulative errors as Gyro and MUSE do. This is an important property for real-world deployment as cumulative error requires frequent calibration.

4.5 Impact of Factors

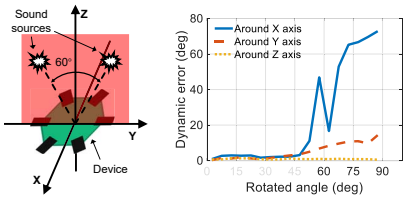
4.5.1 Number of Sound Sources. So far we evaluate MOM with only two sound sources for orientation estimation. There are often more than two sound sources in practice. Thus, we conduct experiments to evaluate the impact of the number of sound sources on system performance. We vary the number of sound sources from 2 to 4 in this experiment. We place two more sound sources near the first two sound sources at a distance of 0.5 m. Fig. 11(a) shows the CDF of the dynamic error with 2, 3 and 4 sound sources. Note that we use dynamic error in this experiment, which is more representative than the static error. Interestingly, adding two close-by extra sound sources does not improve the performance of MOM. On the contrary, the more sound sources we include, the larger error we obtain. We believe this is because the two additional sound sources are too close to the original ones, and thus the clustering algorithm mistakenly clusters the sound sources, leading to larger errors. We further place the four sounds sources far away from each other and the results in Fig. 11(b) shows performance improvement. Therefore, we can conclude that blindly increasing the number of sound sources without careful placement cannot guarantee performance improvement. As such, we need to make sure the sound sources are sparsely located which can bring in diversity and do not confuse the clustering algorithm. In the rest of this section, we use two sound sources by default.

4.5.2 Relative Angle and Source-microphone Distance. The relative angle refers to the angle between the connections from the two sound sources to the device. The static and dynamic errors with different relative angles are presented in Fig. 12(a). In this experiment, unless specifically mentioned, the setup is the same as that in the overall performance evaluation in Sec. 4.4.2. The results show that a low static error can be achieved at all relative angles. However, when the relative angle is small (e.g., 20°), the dynamic error is large. This is because when the device is rotating rapidly, the clusters of the two sound sources can be entangled, and hence become difficult to distinguish between each other. In this case, MOM cannot effectively identify the two sound sources, leading to larger errors.

We also evaluate the impact of distance between the sound sources and the device. As shown in Fig. 12(b), the error does not increase rapidly with longer distances. Such results show that signal attenuation over distances does not significantly degrade the AoA extraction accuracy, which ensures MOM's robustness in the scenario where the sound sources are far away or with small volumes.

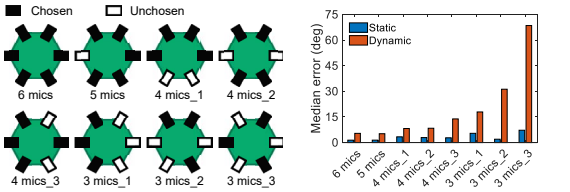
4.5.3 Source-Device Orientation. The source-device orientation refers to the device's orientation with respect to the sound sources as the transmissions from sound sources are usually directional. Generally, with different source-device orientations, the AoA accuracy varies [38]. We evaluate the robustness of MOM against different source-device orientations. The initial source-device orientation is illustrated in Fig. 13(a), i.e., two sound sources reside on the Y-Z plane with a relative angle of 60°.

The orientation accuracy over X, Y and Z axes are presented in Fig. 13 (b). Here we conduct the experiment by rotating the



(a) Initial orientation (b) Dynamic errors for different orientations

Figure 13: The initial orientation of the device (a) and the median dynamic errors at different rotated angles around the X, Y and Z axes (b).



(a) Different microphone layouts (b) Impact of microphone layout

Figure 14: The eight different microphone layouts (a) and their corresponding median static and dynamic errors (b).

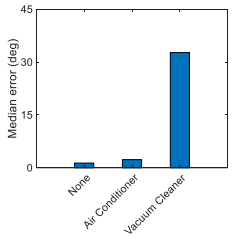


Figure 15: The static errors under different background noises.

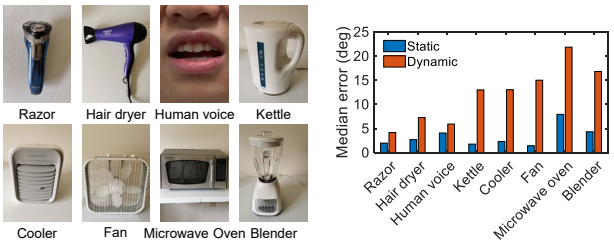
device. Hence, only the dynamic errors are reported. In general, when the device is rotating around X axis, the errors stay small. However, when the rotation angle is larger than 60°, the orientation estimation error increases dramatically. The reason is when the rotation angle is greater than 60°, one of the sound sources is behind the device and therefore the sound signal from this sound source cannot be received anymore. In this case, with only one available sound source, the performance of MOM unavoidably degrades. In addition, the error increases gradually with increasing rotation angle around the Y axis. This process is equivalent to gradually decreasing the elevation angle of the sound sources in the device's reference coordinate. It shows that the error gradually increases as the elevation angle decreases, which matches the observation reported in [38]. As for Z axis, the error keeps small due to the fact that the circular microphone array is symmetric around the Z axis.

4.5.4 Number of microphones and layout. In practice, not all the devices are equipped with six microphones. In this experiment, we evaluate 8 layouts with different combinations of 3-6 microphones on ReSpeaker. Fig. 14(a) shows the details of these layouts. The median static errors and dynamic errors are presented in Fig. 14(b). The results show that the estimation errors increase when fewer microphones are used. In particular, when there are three microphones, which is the minimal number of microphones required for 3D orientation detection, the estimation errors increase dramatically. In this case, the layout plays an important role in MOM's performance, especially in the dynamic errors. The best layout in which the microphones distribute unevenly achieves a dynamic error of 17.93° while the dynamic error for the worst layout is 68.54°.

4.6 Deployment in Real World

We further deploy MOM in real-world settings, where different sound sources and background noises exist.

4.6.1 Sound Source Diversity. We evaluate the performance of MOM with different types of real-world sound sources, including hair dryer, blender, human voice, razor, microwave, fan, cooler and kettle, as shown in Fig. 16(a). As Fig. 16(b) illustrates, the performance varies across different sound sources. We observe that the smaller the size of the sound source, the more accurate the orientation estimates. The reason is that a smaller sound source is more like a point source, whose AoAs are more focused. Interestingly, the sound source volume does not have a significant impact on MOM's



(a) Different daily sound sources (b) Impact of sound source type

Figure 16: The eight daily sound sources (a) and their corresponding median static and dynamic errors (b).

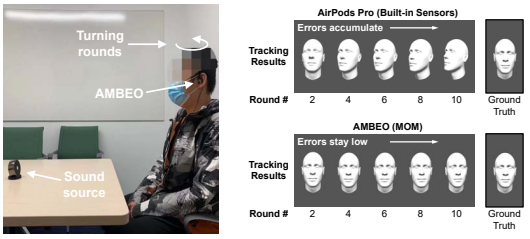
performance. In our evaluation, the razor is with the smallest size and small volume. We achieve both the smallest static errors (2.0°) and dynamic errors (4.2°) with the razor as the sound source.

4.6.2 Background noises. We now evaluate the system performance in the presence of background noise. We consider two commonly-seen background noises, i.e., noise from the air conditioner and noise from the vacuum cleaner. When there is no explicit background noise, the sound volume at the receiver is around 44 dB. We employ the same setup in Sec. 4.4.2 to evaluate the static error of orientation estimates. The results are shown in Fig. 15. We can see that the errors stay low when the air conditioner is working. However, a much larger error can be observed when the vacuum cleaner is working. This is because the noise level caused by the vacuum cleaner is much higher (66.0 dB) and the noise source also keeps moving during the vacuum cleaner's operation.

4.7 Applications

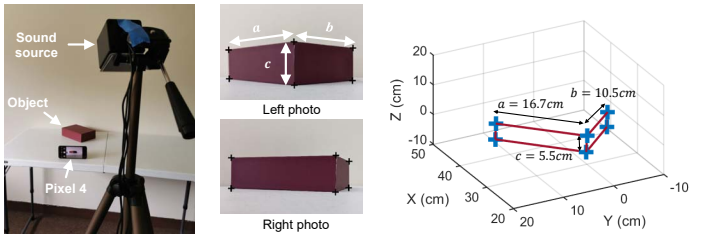
To further demonstrate MOM's applicability in real life, we develop two applications of MOM on commodity devices: earphone-based head tracking and smartphone-based 3D reconstruction.

4.7.1 Earphone-based Head Tracking. Recent commodity earphones, such as AirPods Pro, begin to integrate built-in gyroscopes, which enable earphones to track users' head motions. Such head tracking facilitates multiple on-ear applications: head gesture recognition [21, 39] and acoustic augmented reality (AAR) [48]. For these on-ear applications, the orientation around the Z axis (i.e., γ in Euler angle) is the most critical information. This is because the head's γ alone is enough for most applications, such as AAR. In addition,



(a) Setup (b) Head tracking result

Figure 17: In head tracking, the experiment setup (a) and the APP's screenshots of the head tracking results (b) after even rounds of turning.



(a) Setup (b) Images (c) 3D reconstruction result

Figure 18: In 3D reconstruction, the experiment setup (a), the example of two images of an object (b), and the reconstructed object in 3D space (c).

the range of γ varies from 0° to 360° when a user is walking around, while that of the other two axes is limited in a small range.

However, measuring a head's γ remains challenging on earphones for two reasons. First, due to the very limited size, the integrated mini-version gyroscope [22] performs not as well as those in smartphones, suffering from serious cumulative errors [14]. Besides, the magnetometer that is commonly used for calibration γ can be easily polluted by the ferromagnetic material in the earphones [13]. Fortunately, most commodity earphones are equipped with one microphone in each earbud, which ensures the feasibility of MOM's deployment on earphones. In our evaluation, a user wearing earphones turns 10 rounds[1] in a chair as shown in Fig. 17(b). We display the APP's screenshots of the head orientation results after even rounds of turning. As Fig. 17(b) shows, the measurement error of the baseline based on the gyroscope sensor inside AirPods Pro ($200) accumulates during the process of user's turning, reaching 60° after 10 rounds. Meanwhile, we implement MOM on a commodity earphone, Sennheiser AMBEO smart headset ($60), to measure the head's orientation (γ). We can observe a much smaller error, merely $2-3°$ after 10 rounds of head turning in Fig. 17(b).

4.7.2 Smartphone-based 3D Reconstruction. 3D reconstruction is widely used in diverse real-life applications such as virtual reality [20] and autonomous driving [8]. In autonomous driving [8], multiple cameras are deployed to detect the vehicles' sizes using 3D reconstruction. For these applications, multiple images are taken at different orientations. A small orientation deviation can lead to significant 3D reconstruction errors and therefore 3D reconstruction requires high accuracy in orientation measurements.

As Fig. 18(a) shows, this application is implemented on a Pixel 4 smartphone by using the top and bottom microphones of the phone. With accurate camera orientations captured by MOM, the two images, shown in Fig. 18(b), can be used to reconstruct the shape of the target red box in 3D space, shown in Fig. 18(c). We employ the method introduced in [26, 32] for 3D reconstruction. We first calculate the key points' spatial angle received at the camera based on the key points' pixel positions on the images. Such spatial angle can be transformed to the world reference coordinate by applying the camera's orientation information obtained by MOM. Then, we draw two rays from the camera locations, along the transformed spatial angles. Finally, the points in the 3D space with the minimum distances to the two rays serve as the reconstructed key points. With the orientation information obtained from our system, the dimensions of the target (i.e., a, b, c in Fig. 18(c)) can be estimated accurately with a mean error less than 5%.

[1]One round indicates an orientation change of 360° back to the initial orientation.

5 RELATED WORK

In this section, we discuss the related work in three broad categories, i.e., acoustic sensing, AoA extraction and orientation estimation.

Acoustic Sensing. Owing to the low propagation speed in the air (340 m/s), acoustic signals can be utilized to achieve fine-grained sensing. Inaudible ultrasound is usually adopted for a more conformable user experience. Ultrasound-based object tracking has been successfully implemented on commodity smartphones [49] and earphones [10]. In the field of smart health, ultrasound has been exploited to monitor human lung functions [37], measure heartbeat [50] and detect eye blink [25], which push the sensing granularity to sub-millimeter level. Besides ultrasound, a recent study [40] adopts white noise to monitor an infant for respiration, motion and cry sensing. Furthermore, environment temperature can also be estimated by measuring the signal speed in the air as sound speed is closely related to the air temperature [9].

AoA Extraction. When a signal is received by a device, the AoA information can be exploited in numerous applications, such as localization and navigation. Traditional methods for AoA extraction include MUSIC [33] and ESPIRIT [31]. ArrayTrack [46] first utilizes an antenna array on a WiFi access point to obtain the AoA information for indoor localization. MD-Track[45] further exploits information from multiple dimensions (angle, time and Doppler shift) rather than just angle information to further distinguish signals mixed together for multi-target localization.

As for acoustic signals, there are two types of AoA extraction: *active* and *passive*. For active AoA extractions, a device transmits dedicated acoustic signals and extracts AoAs from the reflected signals. Active AoA extraction has been widely used in device tracking [27] and object imaging [28]. To improve the tracking accuracy and sensing range, a recurrent neural network (RNN) is applied [29]. In FM-Track [24], signal parameters from the time domain, frequency domain and spatial domain are jointly estimated to achieve fine-grained multi-target sensing. In passive AoA extraction, unknown ambient acoustic signals are utilized. GCC-PHAT [19] extracts AoAs

based on delays. To extract multiple AoAs, VoLoc [35] proposes an algorithm that extracts AoAs iteratively. Symphony [41] improves GCC-PHAT by considering delays at multiple microphones. The AoA accuracy of these works is usually limited by the sampling rate and the spacing of the microphones.

Orientation Estimation. Compared to location tracking, orientation estimation is an equally important task in many real-life applications. The traditional way to estimate orientation is using the gyroscope with Kalman filter [23] and complementary filter [12]. To overcome the cumulative error generated by the gyroscope, other sensors are also exploited. The most commonly used sensors are accelerometers and magnetometers. Their usage can be found on various smart devices, including smartphones [36, 51] and earphones [13, 14].

Without a need for inertial sensors, some other methods are also proposed to track object orientation. RFID-based solutions [17, 42] required deploying multiple RFID tags to track the orientation of a device. A recent work [34] also exploited the polarization information to track the orientation of RFID tags. Although RFID tags are cheap, the RFID reader equipped with polarized antennas is expensive. It is also difficult to deploy multiple RFID tags on small-sized electronic devices such as a smartphone in our system. In a recent work [44], an antenna array was utilized to track the orientation of a device. However, a dedicated Intel 5300 WiFi card needs to be used which is not available in consumer-level devices such as smartphones. The GPS based solutions [15] are mainly used on drones. Due to the coarse distance accuracy of the GPS module and the requirement of installing multiple GPS modules at the target device, this method is feasible to track the orientation of a larger device such as a drone but is not suitable in tracking small devices such as smartphones and earphones. Also GPS-based solutions only work in the outdoor environment, while our system focuses on the scenarios in an indoor environment where the GPS signals are too weak to be utilized for tracking. The induction coil method measures the time-varying magnetic flux density to track the orientation of a device. Based on the fact that the magnetic flux density changes as a magnetic object rotates, MET [16] utilized multiple induction coils to track the orientation of an electric toothbrush by measuring its motor's magnetic fields.

6 DISCUSSION

In this section, we briefly discuss the limitations of the proposed system and potential future work.

Ultrasound Implementation. So far, we utilize free audible sound sources in the environment for orientation tracking. Besides audible signals, dedicated ultrasound sources can also be employed in special scenarios, such as in a virtual reality room. We believe with a controlled chirp signal in the inaudible band, even higher accuracy can be achieved.

Far field requirement. MOM estimates a device's orientation based on the AoAs of signals from the sound sources to the device. In this work, we make an assumption that the sound sources are in the far field. In other words, sound sources are sufficiently far away (e.g., 1-2 m) from the device. If the device is very close to the sound sources (e.g., less than 20 cm), this assumption does not hold and the AoA estimation can be inaccurate, degrading the system performance.

Privacy Concern. To obtain the device orientation information, MoM continually captures acoustic signals in the environment including human voice. These collected signals may leak private information. A possible solution is to collect acoustic signals only in some time windows instead of all of them. That is to say, we collect signals in one time window, and drop signals in the following several windows. In this way, while AoA information can still be successfully extracted, voice privacy is protected.

Moving Sound Source. The sound sources are assumed to be at fixed locations in MOM. In reality, some of the sound sources may move over time. These moving sound sources can fail the proposed system and therefore should not be taken as anchors. Fortunately, when a sound source moves, the signal frequency varies due to the Doppler effect. Hence, those sound sources with obvious Doppler frequency shifts can be detected and excluded from being taken into consideration as anchors.

7 CONCLUSION

In this paper, we present MOM, a microphone-based orientation estimation system. MOM does not need to deploy dedicated sound sources and achieves highly accurate tracking performance, outperforming the state-of-the-arts. The proposed system can be easily integrated into widely available commodity hardware such as smartphones and earphones. The proposed system involves orientation tracking into the ecosystem of acoustic sensing, moving one step closer to real-life adoption of acoustic sensing.

REFERENCES

[1] The world's motion capture technology partner. https://www.qualisys.com/.
[2] ReSpeaker 6-Mic Circular Array Pi HAT. https://wiki.seeedstudio.com/cn/ReSpeaker_6-Mic_Circular_Array_kit_for_Raspberry_Pi/, 2018.
[3] AMBEO Smart Headset. https://en-us.sennheiser.com/in-ear-headphones-3d-audio-ambeo-smart-headset, 2021.
[4] Audio Hardware Configurations. https://developer.amazon.com/en-US/docs/alexa/alexa-voice-service/audio-hardware-configurations.html, 2021.
[5] MPU-6050 Six-Axis (Gyro + Accelerometer) MEMS MotionTracking™ Devices. https://invensense.tdk.com/products/motion-tracking/6-axis/mpu-6050/, 2021.
[6] P. Bahl and V. N. Padmanabhan. Radar: An in-building rf-based user location and tracking system. In *Proceedings IEEE INFOCOM 2000.*, 2000.
[7] M. Bloesch, H. Sommer, T. Laidlow, M. Burri, G. Nuetzi, P. Fankhauser, D. Bellicoso, C. Gehring, S. Leutenegger, M. Hutter, et al. A primer on the differential calculus of 3d orientations. *arXiv preprint arXiv:1606.05285*, 2016.
[8] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019.
[9] C. Cai, Z. Chen, H. Pu, L. Ye, M. Hu, and J. Luo. Acute: acoustic thermometer empowered by a single smartphone. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, 2020.
[10] G. Cao, K. Yuan, J. Xiong, P. Yang, Y. Yan, H. Zhou, and X.-Y. Li. Earphonetrack: involving earphones into the ecosystem of acoustic motion tracking. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, 2020.
[11] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, 1996.
[12] M. Euston, P. Coote, R. Mahony, J. Kim, and T. Hamel. A complementary filter for attitude estimation of a fixed-wing uav. In *2008 IEEE/RSJ international conference on intelligent robots and systems*, 2008.
[13] A. Ferlini, A. Montanari, A. Grammenos, R. Harle, and C. Mascolo. Enabling in-ear magnetic sensing: Automatic and user transparent magnetometer calibration. In *2021 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 2021.
[14] A. Ferlini, A. Montanari, C. Mascolo, and R. Harle. Head motion tracking through in-ear wearables. In *Proceedings of the 1st International Workshop on Earable Computing*, 2019.

[15] M. Gowda, J. Manweiler, A. Dhekne, R. R. Choudhury, and J. D. Weisz. Tracking drone orientation with multiple gps receivers. In *Proceedings of the 22nd annual international conference on mobile computing and networking*, 2016.

[16] H. Huang and S. Lin. Met: a magneto-inductive sensing based electric toothbrushing monitoring system. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, 2020.

[17] C. Jiang, Y. He, X. Zheng, and Y. Liu. Orientation-aware rfid tracking with centimeter-level accuracy. In *2018 17th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 2018.

[18] K. Joshi, D. Bharadia, M. Kotaru, and S. Katti. Wideo: Fine-grained device-free motion tracing using {RF} backscatter. In *12th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 15)*, pages 189–204, 2015.

[19] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE transactions on acoustics, speech, and signal processing*, 1976.

[20] Z. Lai, Y. C. Hu, Y. Cui, L. Sun, N. Dai, and H.-S. Lee. Furion: Engineering high-quality immersive virtual reality on today's mobile devices. *IEEE Transactions on Mobile Computing*, 2019.

[21] M. Laporte, P. Baglat, S. Gashi, M. Gjoreski, S. Santini, and M. Langheinrich. Detecting verbal and non-verbal gestures using earables. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*, 2021.

[22] J. H. Lau. System-in-package (sip). In *Semiconductor Advanced Packaging*, 2021.

[23] H.-J. Lee and S. Jung. Gyro sensor drift compensation by kalman filter to control a mobile inverted pendulum robot system. In *2009 IEEE International Conference on Industrial Technology*, 2009.

[24] D. Li, J. Liu, S. I. Lee, and J. Xiong. Fm-track: pushing the limits of contactless multi-target tracking using acoustic signals. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, 2020.

[25] J. Liu, D. Li, L. Wang, and J. Xiong. Blinklistener: " listen" to your eye blink using your smartphone. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(2):1–27, 2021.

[26] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, 1999.

[27] W. Mao, J. He, and L. Qiu. Cat: high-precision acoustic motion tracking. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, pages 69–81, 2016.

[28] W. Mao, M. Wang, and L. Qiu. Aim: Acoustic imaging on a mobile. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*, 2018.

[29] W. Mao, M. Wang, W. Sun, L. Qiu, S. Pradhan, and Y.-C. Chen. Rnn-based room scale hand motion tracking. In *The 25th Annual International Conference on Mobile Computing and Networking*, 2019.

[30] K. Qian, C. Wu, Y. Zhang, G. Zhang, Z. Yang, and Y. Liu. Widar2. 0: Passive human tracking with a single wi-fi link. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*, pages 350–361, 2018.

[31] R. Roy and T. Kailath. Esprit-estimation of signal parameters via rotational invariance techniques. *IEEE Transactions on acoustics, speech, and signal processing*, 1989.

[32] O. Saurer, M. Pollefeys, and G. H. Lee. A minimal solution to the rolling shutter pose estimation problem. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015.

[33] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE transactions on antennas and propagation*, 1986.

[34] L. Shangguan and K. Jamieson. Leveraging electromagnetic polarization in a two-antenna whiteboard in the air. In *Proceedings of the 12th International on Conference on emerging Networking EXperiments and Technologies*, 2016.

[35] 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.

[36] S. Shen, M. Gowda, and R. Roy Choudhury. Closing the gaps in inertial motion tracking. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, 2018.

[37] X. Song, B. Yang, G. Yang, R. Chen, E. Forno, W. Chen, and W. Gao. Spirosonic: monitoring human lung function via acoustic sensing on commodity smartphones. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, 2020.

[38] T.-C. Tai, K. C.-J. Lin, and Y.-C. Tseng. Toward reliable localization by unequal aoa tracking. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, 2019.

[39] S. Voelker, S. Hueber, C. Corsten, and C. Remy. Headreach: Using head tracking to increase reachability on mobile touch devices. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.

[40] A. Wang, J. E. Sunshine, and S. Gollakota. Contactless infant monitoring using white noise. In *The 25th Annual International Conference on Mobile Computing and Networking*, 2019.

[41] W. Wang, J. Li, Y. He, and Y. Liu. Symphony: localizing multiple acoustic sources with a single microphone array. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, 2020.

[42] T. Wei and X. Zhang. Gyro in the air: tracking 3d orientation of batteryless internet-of-things. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, 2016.

[43] E. Whitaker. On the functions which are represented by the expansion of interpolating theory. In *Proc. Roy. Soc. Edinburgh*, 1915.

[44] C. Wu, F. Zhang, Y. Fan, and K. R. Liu. Rf-based inertial measurement. In *Proceedings of the ACM Special Interest Group on Data Communication*, 2019.

[45] Y. Xie, J. Xiong, M. Li, and K. Jamieson. md-track: Leveraging multi-dimensionality for passive indoor wi-fi tracking. In *The 25th Annual International Conference on Mobile Computing and Networking*, 2019.

[46] J. Xiong and K. Jamieson. Arraytrack: A fine-grained indoor location system. In *10th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 13)*, 2013.

[47] Y. Yang, Y. Ding, D. Yuan, G. Wang, X. Xie, Y. Liu, T. He, and D. Zhang. Transloc: transparent indoor localization with uncertain human participation for instant delivery. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, 2020.

[48] Z. Yang, Y.-L. Wei, S. Shen, and R. R. Choudhury. Ear-ar: indoor acoustic augmented reality on earphones. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, 2020.

[49] S. Yun, Y.-C. Chen, H. Zheng, L. Qiu, and W. Mao. Strata: Fine-grained acoustic-based device-free tracking. In *Proceedings of the 15th annual international conference on mobile systems, applications, and services*, 2017.

[50] F. Zhang, Z. Wang, B. Jin, J. Xiong, and D. Zhang. Your smart speaker can" hear" your heartbeat! *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2020.

[51] P. Zhou, M. Li, and G. Shen. Use it free: Instantly knowing your phone attitude. In *Proceedings of the 20th annual international conference on Mobile computing and networking*, 2014.