EHTrack: Earphone-Based Head Tracking via Only Acoustic Signals

Linfei Ge, Qian Zhang, *Fellow, IEEE*, Jin Zhang, *Member, IEEE*, and Huangxun Chen, *Member, IEEE*

***Abstract*—Head tracking is a technique that allows for the measurement and analysis of human focus and attention, thus enhancing the experience of human–computer interaction (HCI). Nevertheless, current solutions relying on vision and motion sensors exhibit limitations in accuracy, user-friendliness, and compatibility with the majority of commercial off-the-shelf (COTS) devices. To overcome these limitations, we present EHTrack, an earphone-based system that achieves head tracking exclusively through acoustic signals. EHTrack employs acoustic sensing to measure the movement of a pair of earphones, subsequently enabling precise head tracking. In particular, a pair of speakers generates a periodically fluctuating sound field, which the user's two earphones detect. By assessing the distance and angle alterations between the earphones and speakers, we propose a model to determine the user's head movement and orientation. Our evaluation results indicate a high degree of accuracy in both head movement tracking, with an average tracking error of 2.98 cm, and head orientation tracking, with an average error of 1.83°. Furthermore, in a deployed exhibition scenario, we attained an accuracy of 89.2% in estimating the user's focus direction.**

***Index Terms*—Acoustic signal processing, human computer interaction, signal processing, systems, user interfaces.**

I. Introduction

HEAD orientation provides valuable information about people's intentions, as individuals generally turn their heads toward the desired direction to observe what interests them. Consequently, head tracking plays a crucial role in human–computer interaction (HCI) applications, such as tracking users' attention during webpage browsing for content customization or tracking user attention for exhibit introduction in museums. An accurate, user-friendly, and widely applicable head tracking solution compatible with commercial off-the-shelf (COTS) devices is desired for daily use.

Existing head tracking systems cannot fully achieve the desired goal. For vision-based solutions, prior works [1], [2], [3], [4] leveraged cameras to capture images of the human head for tracking, which requires users to sit in front of the cameras and does not work in mobile scenarios. Consequently, [5] proposed mounting cameras on headsets or smart glasses to infer head orientation from videos under a first-person perspective. However, wearing these devices in daily scenarios is inconvenient, uncomfortable, and raises potential privacy concerns. Motion sensor-based solutions [6], [7], [8] utilize inertial measurement units (IMUs), including an accelerometer, a gyroscope, and a magnetometer, to track head movement. These solutions are more cost-effective and lightweight compared to vision-based ones. However, we note that the majority of commercially available earphones, such as the Sony WF-1000XM4 [9] and Bose QuietComfort Earbuds [10], do not come equipped with IMUs. As a result, IMU-based solutions are currently incompatible with most COTS earphones.

To address this gap, we identify that the earphone-based approach holds great potential for achieving user-friendly and widely applicable head tracking. It is increasingly common to wear wireless earphones in daily scenarios, such as watching videos on a laptop or exploring an exhibition. Compared to a bulky headset-based solution, an earphone-based system is more user-friendly and versatile. The advantage of earphone-based head tracking lies in the relatively fixed location of the earphones with respect to the head, as they move and rotate in tandem with the head. As illustrated in Fig. 1, a user wearing a pair of earphones is looking at points A, B, and C. When the user turns from looking at A to B, both earphones move and rotate accordingly. Consequently, we can enable head tracking by accurately tracking the earphones. The appropriate sensing modality for the earphone-based solution still needs to be considered. Acoustic signals are suitable since the required sensors, speakers, and microphones are commonly available in earphones and relevant scenarios. As depicted in Fig. 1, with speakers deployed in the environment, earphones can utilize their microphones to receive anchor signals for continuous localization and tracking. In this way, we can transform these earphones into HCI-compatible devices.

In order to achieve accurate earphone-based head tracking, we must overcome three main challenges. First, we need a robust method capable of accurately and simultaneously tracking two earphones in motion. This can be challenging due to environmental factors that affect acoustic signals,

Manuscript received 1 June 2023; revised 27 June 2023; accepted 17 July 2023. Date of publication 25 July 2023; date of current version 24 January 2024. This work was supported in part by RGC under Contract CERG 16204820, Contract 16206122, Contract R8015, Contract FS112, and Contract 3030_006. *(Corresponding authors: Qian Zhang; Jin Zhang.)*

Linfei Ge is with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, China, and also with the Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China (e-mail: lgead@connect.ust.hk).

Qian Zhang is with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, China (e-mail: qianzh@cse.ust.hk).

Jin Zhang is with the Research Institute of Trustworthy Autonomous Systems and the Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China (e-mail: zhangj4@sustech.edu.cn).

Huangxun Chen is with the Department of Theory Lab, Huawei Hong Kong Research Center, Hong Kong, China (e-mail: chen.huangxun.amy@gmail.com).

Digital Object Identifier 10.1109/JIOT.2023.3298412

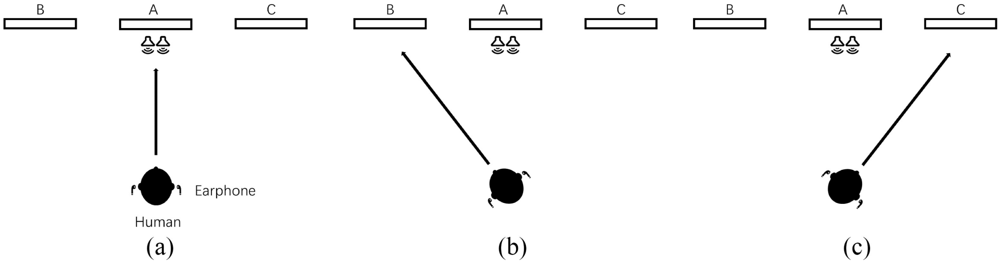


Fig. 1. Illustrated scenario of human attention. (a) Person looks at object A. (b) Person looks at object B. (c) Person looks at object C.

resulting in poor signal quality and making it difficult to accurately derive angle and distance measurements. While numerous recent works [11], [12], [13] focus on acoustic-based tracking for a single device, such as a speaker or microphone, limited research has been conducted on tracking two microphones simultaneously to enable precise head tracking. Second, achieving accurate earphone-based head tracking necessitates the establishment of an accurate relationship model between the head and the earphones. This model allows us to transform the earphones' tracking results into head movement and orientation accurately. Furthermore, since tracking the movement of the earphones is not entirely precise, a reliable head tracking model that is robust against earphone tracking errors is required. Third, we need an effective method to determine the initial location, providing a starting point for the tracking system. It is crucial that this method is not only effective but also minimally burdensome for the user.

To address the aforementioned challenges, we propose EHTrack, an earphone-based head tracking system that derives the position and orientation of the head using only acoustic signals collected by a pair of earphones. We require only that the earphones be equipped with microphones, which are compatible with most COTS devices. As a result, EHTrack significantly improves the availability of head tracking in daily scenarios. In our proposed system, we design strength-based angle tracking and phase-based distance tracking to determine the movement of the earphones. Specifically, a pair of speakers in the environment will play sine waves at different frequencies, allowing the superposition of these waves to generate a periodically changing sound field. It is worth noting that we employ an ultrasound frequency above 18 kHz, which is uncommon in everyday scenarios, to effectively mitigate the impact of external disturbances on the tracking process. We then leverage specially designed patterns of acoustic strength to enable robust inference of the angle and distance changes of the earphone (microphone) relative to the sound generators within the sound field. To accommodate multiple users simultaneously in our targeted application scenario, such as exhibitions, we have opted to use continuous wave (CW) signals instead of FMCW or Zadoff–Chu signals to alleviate the burden of synchronization. Our system, in theory, can support an unlimited number of receivers operating simultaneously. Next, we derive an analytic model to explicitly characterize the relationship between ear movement and head movement/orientation. Based on this model, we can transform the distance and angle estimations of the two earphones into head movement and orientation, thereby achieving head tracking. Additionally, we design a convenient method to obtain the initial location. This method involves simply asking the user to stand still and rotate their head by a certain angle. EHTrack will detect the rotation and assume that there is no movement but only rotation, allowing us to derive the initial location using our proposed model. This method is easy for users to perform and does not require any additional hardware for initial location estimation.

Our contributions are summarized as follows.

1) We propose EHTrack, an accurate and user-friendly earphone-based head tracking system that is widely applicable in daily scenarios and COTS devices. To the best of our knowledge, EHTrack is the first purely acoustic-based head tracking system.
2) We design a special sound field excitation scheme and integrate strength-based angle tracking and phase-based distance tracking methods. As a result, we track a pair of earphones accurately and simultaneously, overcoming the challenges posed by environmental factors that can affect acoustic signals. Additionally, we establish an analytic model that can transform the movement of the earphones into head movement/orientation accurately.
3) We have implemented EHTrack on COTS devices and conducted extensive evaluations to assess its effectiveness. Our results demonstrate that EHTrack achieves accurate head tracking with an average tracking error of 2.98 cm and an orientation tracking error of 1.83°. Furthermore, in a deployed exhibition scenario, we achieved attention direction estimation with an 89.2% accuracy rate. Furthermore, we conducted a thorough examination of the influence of various environmental factors. This analysis encompassed different touring scenarios, diverse users, environmental noise levels spanning from 40 to 80 dB, as well as spaces of varying sizes. These results demonstrate the effectiveness, robustness, and potential of EHTrack as an alternative to existing head tracking solutions.

The remainder of this article is organized as follows. We elaborate on the technical design of EHTrack in Section II. We describe the detailed implementation of EHTrack in Section III. We evaluate the system performance in Section IV. We review related works in Section VI. Finally, we discuss the limitations and conclude this article in Sections V and VII.

II. System Design of EHTrack

In this section, we first provide an overview of the system and its components. Subsequently, we describe the

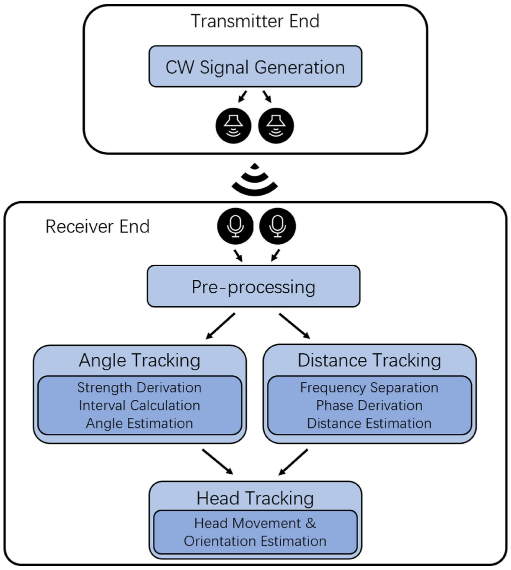


Fig. 2. System overview.

process of deriving angle and distance changes by utilizing specially designed patterns of acoustic strength. We then present our model for transforming earphone movements into head movements. Furthermore, we introduce the approach employed to determine the initial location of the user. Finally, we summarize the key aspects of our system to facilitate a comprehensive understanding of its overall functionality.

A. System Overview

Fig. 2 presents an overview of our system, which comprises two parts: 1) the "Transmitter End" and 2) the "Receiver End."

At the Transmitter End, two speakers play sine waves at distinct frequencies. Upon the superposition of these two sine waves, a periodically changing sound field is generated. The estimation of angle and distance is based on this sound field.

At the Receiver End, the strength of the sound field is captured by two microphones. Initially, preprocessing is applied to eliminate noise. Subsequently, we derive the angle and distance from each microphone. We then utilize the angle and distance tracking results to determine head movement and orientation. Ultimately, we achieve head tracking based on the aforementioned tracking results.

In the following two sections, we will introduce the method for deriving angle and distance from acoustic signals.

B. Strength-Based Angle Tracking

Angle tracking is achieved through a periodically changing sound field [14]. In our system, we employ two speakers to generate the sound field, with each speaker emitting sine waves at specific frequencies. One speaker emits a sine wave at frequency f_1, while the other emits a sine wave at frequency f_2. The superposition of these two sine waves generates the sound field. A microphone within the sound field can detect the sound strength resulting from the superposition of the two sine waves.

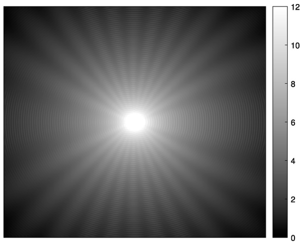


Fig. 3. Example of sound field.

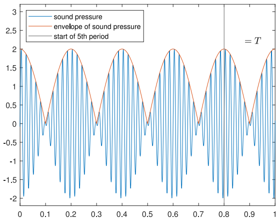


Fig. 4. Received signal with standard strength period.

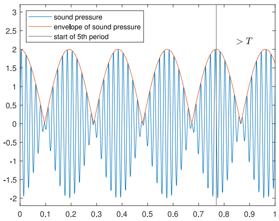


Fig. 5. Received signal with larger strength period.

If the phase difference between these sine waves is 0, they are in phase, and constructive interference occurs, causing the microphone to detect a large sound strength. Conversely, if the phase difference is π, they are out of phase, and destructive interference occurs, leading the microphone to detect a small sound strength. In essence, the strength distribution of the sound field is uneven due to the phase difference. Fig. 3 provides an example of the sound field, where bright areas indicate large sound strength and dark areas signify small sound strength.

If $f_1 = f_2$, the phase difference is solely caused by location differences, resulting in a static sound field, as depicted in Fig. 3. If $f_1 \neq f_2$, however, the phase difference is caused by both location and time, creating a dynamic sound field. In practice, the sound field appears to "rotate" around the center.

In a rotating sound field, a static microphone within the sound field detects changes in strength. The frequency of this strength is denoted by $f_0 = |f_1 - f_2|$. Fig. 4 provides an example of a received signal when $f_1 = 50$ Hz and $f_1 = 55$ Hz. The blue line represents the received signal, and the orange line illustrates the envelope. The envelope corresponds to the sound strength, which we utilize to derive the strength period. If the microphone is moving, the observed frequency of the sound field strength change, f_{obs}, will differ from f_0. Consequently, the period T_{obs} will also differ. Fig. 5 displays an example of a moving microphone, where the frequency of the signal strength is smaller than f_0, and the period T_{obs} is larger than the standard period $T_0 = (1/f_0)$. Similarly, if

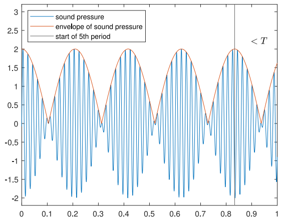


Fig. 6. Received signal with smaller strength period.

the movement direction is reversed, we will observe a larger frequency and a smaller period, as shown in Fig. 6. By calculating the period difference ΔT, we can determine the angular speed and ultimately achieve angle tracking.

C. Phase-Based Distance Tracking

Phase-based distance tracking is an efficient and accurate method, which is why we have chosen to use it for distance tracking.

Phase-based distance tracking utilizes the phase of a sound wave to track distances. Previous works [15], [16] have successfully implemented millimeter-level motion tracking based on acoustic phase. In our system, we employ a single-frequency sine wave to measure distance. If there is a speaker emitting a sine wave and a microphone receiving it, the distance between them can be calculated as $d = \lambda * (N + [\phi/2\pi])$, where λ represents the wavelength, N is an integer, and ϕ is the phase, ranging between 0 and 2π. As the distance d changes, the phase ϕ also varies. By determining the phase change, we can obtain the distance change and achieve distance tracking. Further details are provided below.

The received signal can be expressed as follows:

$$S_r = A\cos\left(2\pi f t - 2\pi f \frac{d(t)}{v_s} + \phi\right) \tag{1}$$

wherein A is the amplitude of the signal, f is the frequency of the signal, t is the time, and ϕ is a constant phase offset. In this context, v_s denotes the speed of sound, $d(t)$ represents the distance traveled by the signal at time t, and $2\pi f(d(t)/v_s)$ signifies the phase change caused by the distance traveled.

First, we derive the phase from the received signal. In our system, the received signal S_r contains two sine waves at frequencies f_1 and f_2. To derive the In-phase and Quadrature components, we multiply S_r by cosine and sine functions at either frequency f_1 or f_2. For simplicity, we use f_1 without loss of generality

$$S_{r1,\cos} = S_{r1} * \cos(2\pi f_1 t) \tag{2}$$
$$S_{r1,\sin} = S_{r1} * \sin(2\pi f_1 t). \tag{3}$$

From (2), we get the In-phase component

$$\begin{aligned} S_{r1,\cos} &= S_{r1} * \cos(2\pi f_1 t) \\ &= A\cos\left(2\pi f_1 t - 2\pi f_1 \frac{d(t)}{v_s} + \phi\right) * \cos(2\pi f_1 t) \\ &= \frac{1}{2}A\left[\cos\left(-2\pi f_1 \frac{d(t)}{v_s} + \phi\right)\right. \\ &\quad \left. + \cos\left(4\pi f_1 t - 2\pi f_1 \frac{d(t)}{v_s} + \phi\right)\right]. \end{aligned} \tag{4}$$

Next, we apply a low-pass filter to obtain the low-frequency In-phase component, $S'_{r1,\cos} = \cos(-2\pi f_1(d(t)/v_s) + \phi)$. Similarly, we can derive the Quadrature component, $S'_{r1,\sin} = -\sin(-2\pi f_1(d(t)/v_s) + \phi)$. The phase ϕ_d can be derived using the following formula:

$$\phi_d = \arctan\left(\frac{S'_{r1,\cos}}{S'_{r1,\sin}}\right). \tag{5}$$

To address the phase ambiguity problem, an "unwarp" operation is employed. Some phase points are adjusted by adding a value of $2\pi N$ to ensure signal continuity. The wavelength of the acoustic signal at frequency f_1 is given by

$$\lambda = v_s/f_1. \tag{6}$$

v_s represents the speed of sound in air. The distance tracking result is calculated using the following equation:

$$\Delta d = \frac{\phi_{d2} - \phi_{d1}}{2\pi} * \lambda. \tag{7}$$

Here, ϕ_{d1} and ϕ_{d2} represent the distance values before and after the movement, respectively. When the phase experiences a change of 2π, it corresponds to a change in distance equal to the wavelength λ. In Section II-E, we will introduce a dedicated scheme for determining the initial location. By integrating this scheme with the aforementioned derived results, our system achieves effective distance tracking.

D. Head Movement and Orientation Estimation

After angle and distance tracking of earphones, we need to know head movement and orientation from derived tracking results. In this section, we will introduce how we derive head movement and orientation from the above-mentioned angle and distance tracking results.

We use the movement of two earphones to derive head movement and orientation. When people wear two earphones, the locations of the earphone microphones are near people's ears. Thus, we can regard the movement of the earphones as the movement of the ears. In the rest of this section, we will introduce how we derive head movement and orientation from ear movement.

Fig. 7 provides an overview of the setup. $S1$ and $S2$ represent two speakers, while $M1$ and $M2$ are two microphones placed on human ears. We consider the microphones as two ends of a bar with length L, and the person's location at the center of the bar (X, Y). The bar's orientation is denoted by θ. By applying simple geometry, we can determine that the locations of the two microphones $M1$ and $M2$ follows:

$$M1\left(X + L\sin\left(\theta + \frac{\pi}{2}\right), Y + L\cos\left(\theta + \frac{\pi}{2}\right)\right)$$
$$M2\left(X + L\sin\left(\theta - \frac{\pi}{2}\right), Y + L\cos\left(\theta - \frac{\pi}{2}\right)\right).$$

Next, we consider a movement from $Location1$ to $Location2$. Assume that the angle change of $M1$ and $M2$ are $\Delta\alpha$ and $\Delta\beta$, and the distance change of $M1$ and $M2$ are

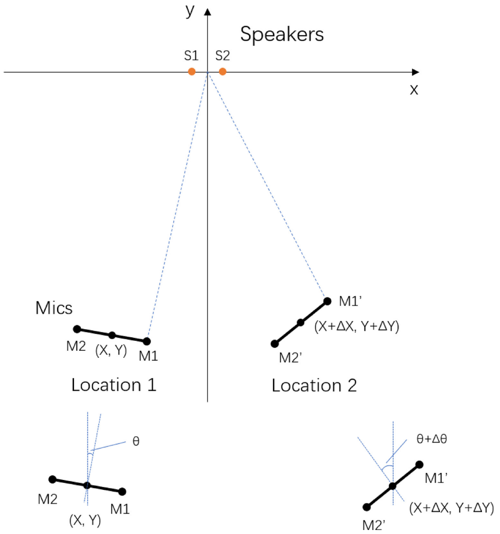


Fig. 7. Head movement and orientation.

Δd_{M1} and Δd_{M2}. Note that the angle and distance changes are relative to the origin $(0, 0)$.

During the movement, we assume that the location and orientation changes are ΔX, ΔY, and $\Delta\theta$. Consequently, the new locations of $M1'$ and $M2'$ are

$$M1'\left(X+\Delta X+L\sin\left(\theta+\frac{\pi}{2}+\Delta\theta\right), Y + \Delta Y+L\cos\left(\theta+\frac{\pi}{2}+\Delta\theta\right)\right)$$
$$M2'\left(X+\Delta X+L\sin\left(\theta-\frac{\pi}{2}+\Delta\theta\right), Y+\Delta Y + L\cos\left(\theta-\frac{\pi}{2}+\Delta\theta\right)\right).$$

To simplify the expression, we set the old locations as $M1(X_{M1}, Y_{M1})$ and $M2(X_{M2}, Y_{M2})$. Thus, the new locations are $M1'(X_{M1'}, Y_{M1'})$ and $M2'(X_{M2'}, Y_{M2'})$. The following relationships are then established:

$$X_{M1} = X + L\sin\left(\theta+\frac{\pi}{2}\right) \tag{8}$$
$$Y_{M1} = Y + L\cos\left(\theta+\frac{\pi}{2}\right) \tag{9}$$
$$X_{M2} = X + L\sin\left(\theta-\frac{\pi}{2}\right) \tag{10}$$
$$Y_{M2} = Y + L\cos\left(\theta-\frac{\pi}{2}\right) \tag{11}$$
$$X_{M1'} = X + \Delta X + L\sin\left(\theta+\frac{\pi}{2}+\Delta\theta\right) \tag{12}$$
$$Y_{M1'} = Y + \Delta Y + L\cos\left(\theta+\frac{\pi}{2}+\Delta\theta\right) \tag{13}$$
$$X_{M2'} = X + \Delta X + L\sin\left(\theta-\frac{\pi}{2}+\Delta\theta\right) \tag{14}$$
$$Y_{M2'} = Y + \Delta Y + L\cos\left(\theta-\frac{\pi}{2}+\Delta\theta\right). \tag{15}$$

We can list four equations

$$\left(-\arctan\frac{X_{M1'}}{Y_{M1'}}\right) - \left(-\arctan\frac{X_{M1}}{Y_{M1}}\right) = \Delta\alpha \tag{16}$$
$$\left(-\arctan\frac{X_{M2'}}{Y_{M2'}}\right) - \left(-\arctan\frac{X_{M2}}{Y_{M2}}\right) = \Delta\beta \tag{17}$$
$$\sqrt{X_{M1'}^2+Y_{M1'}^2} - \sqrt{X_{M1}^2+Y_{M1}^2} = \Delta d_{M1} \tag{18}$$
$$\sqrt{X_{M2'}^2+Y_{M2'}^2} - \sqrt{X_{M2}^2+Y_{M2}^2} = \Delta d_{M2}. \tag{19}$$

In these equations, we know the initial location X, Y, θ, and angle/distance changes $\Delta\alpha$, $\Delta\beta$, Δd_{M1}, Δd_{M2}. By solving (16) to (19), we can obtain the location and orientation changes ΔX, ΔY, $\Delta\theta$. These changes in location and orientation represent the movement of the head.

E. Initial Location Derivation

To initialize the head tracking system, we require the user to stand in front of the speaker and rotate their head without moving. While some motion tracking systems necessitate users to move in a specific direction or distance for initialization, we find that this approach can be challenging for users to execute accurately. Instead, we leverage our head tracking model to derive the initial location and orientation of the user through head rotation.

In the head tracking scenario, we have access to measurements of $X, Y, \theta, \Delta\alpha, \Delta\beta$, and Δd_{M1}, from which we can derive $\Delta X, \Delta Y$, and $\Delta\theta$. When determining the initial location, we capture the rotational movement of the user's head. Since there is no change in location during this type of movement, we know that ΔX and ΔY are both zero. By solving the equations derived in (16)–(19), we can derive X and Y. To improve the accuracy of the results, we use multiple rotational movements and calculate the average results of X and Y. By employing this approach, we can obtain an accurate initial location and orientation for the head tracking system without requiring users to perform specific movements or navigate to a particular location.

F. Putting It All Together

In this section, we provide a comprehensive description of our system's design. An overview of the system is depicted in Fig. 2. We will discuss the design process for both the "Transmitter End" and the "Receiver End."

1) Transmitter End: Our system is based on acoustic signals. The initial task for our system is to generate a sound field. As outlined in Sections II-B and II-C, we need to emit two sine waves at frequencies f_1 and f_2. Superposition occurs, resulting in a sound field with an uneven strength distribution. Due to the phase difference between the two sine waves, both constructive and destructive interference take place within the sound field. If $f_1 = f_2$, the sound field remains static, as illustrated in Fig. 3. If $f_1 \neq f_2$, the sound field becomes dynamic and appears to "rotate" around the center.

In our system, we opt to utilize two distinct frequencies. Consequently, even if the microphone within the sound field remains stationary, it can still detect changes in the sound field's strength.

When determining which frequencies to use, various factors must be taken into account. On the one hand, the frequency should be high enough to be classified as ultrasound and remain inaudible to humans. Generally, humans cannot hear acoustic signals with frequencies above 17 kHz. On the other hand, typical speakers exhibit poor frequency response in high-frequency bands. This implies that emitting a very high-frequency signal may result in a significantly reduced signal strength. A speaker is capable of emitting frequencies below 21 kHz. Signals at lower frequencies, such as 18 kHz, possess much greater strength than those at 21 kHz. Ultimately, we choose to set $f_1 = 18$ kHz and $f_2 = 18.1$ kHz. These frequencies are inaudible to humans and exhibit a robust strength capable of supporting large-distance sensing.

With the generated sound field, we can further derive angle and distance measurements from the acoustic strength.

2) Receiver End: The receiver end comprises several components: preprocessing, angle tracking, distance tracking, and head tracking. We will discuss each of these components in detail.

Preprocessing: Upon receiving the signal, it first undergoes preprocessing by being filtered with a high-pass filter to remove noise. Most daily-life voices have frequencies lower than 10 kHz. A high-pass filter can effectively eliminate these noises. Our system employs frequencies near 18 kHz. As a result, a high-pass filter with a cut-off frequency of 15 kHz is utilized, which removes noise and relatively enhances the signal.

Angle Tracking: As described in Section II-B, angle information is extracted from the period difference, denoted as $\Delta T = T_{\text{obs}} - T_0$. Here, T_0 represents the period when the microphone is static, which can be calculated by $T_0 = (1/|f_1 - f_2|)$. T_{obs} refers to the period we actually observe. Consequently, to obtain ΔT, the crucial aspect of angle estimation involves determining the period T_{obs}. We follow several steps to accurately determine T_{obs}, including strength derivation, period calculation, and angle estimation.

Strength Derivation: The initial step in angle estimation involves determining the received signal strength. For acoustic signals, we consider the average square sum of the signal waveform as the signal strength. Consequently, we need to calculate the average square sum of the received signal. The frequency of the received signal is approximately 18 kHz, and the sampling frequency in our system is 48 kHz. Therefore, there are $48/18 = 2.67$ points in each period. We can calculate the strength of each period; however, this may not be sufficiently accurate due to the presence of only 2.67 points in each period. Instead, we opt to use the strength of multiple periods. We calculate one strength value from 12 points by determining the average square sum of these 12 points as the strength. A group of 12 points, containing more than four periods, is accurate enough to indicate the strength.

Interval Calculation: After obtaining the strength of the received signal, we begin identifying the observed period, T_{obs}, of the signal strength.

We discover that the signal strength is the absolute value of a cosine wave. As Fig. 4 illustrates, the red line represents the strength of received signals. Consequently, the valleys are sharp and easier to identify. By determining the index of these valleys, the strength period is calculated based on the index difference, allowing us to obtain the observed strength period, T_{obs}.

Angle Estimation: Having obtained T_{obs} and the standard period $T_0 = (1/|f_1 - f_2|)$, we can calculate $\Delta T = T_{\text{obs}} - T_0$. By applying the algorithms presented in work [14], we achieve angle tracking.

Distance Tracking: We employ a phase-based method for distance estimation. A detailed mathematical description of phase-based distance estimation is provided in Section II-C.

Frequency Separation: First, we use a band-pass filter to reduce the influence of other frequencies. The received signal contains two sine waves, but distance estimation requires only one of them. Therefore, we apply a band-pass filter initially.

Phase Derivation: This step involves obtaining the phase from the received acoustic signal. According to (2) and (3), we multiply the received signal, S_r, by sine and cosine waves. The frequency of the sine and cosine waves can be either f_1 or f_2. To improve accuracy, we perform distance estimation for both f_1 and f_2. Then, we apply a low-pass filter to derive the phase. Finally, an "unwrap" function is used to make the derived phase continuous.

Distance Estimation: Using (7), we establish a relationship between phase and distance. With the phase derived in the previous step, we calculate the corresponding distance using (7).

Head Tracking: After obtaining angle and distance tracking results, we derive head tracking results from them.

In Section II-D, we provide a detailed mathematical description of how we derive head movement and orientation. Given the known angle and distance changes, $\Delta\alpha$, $\Delta\beta$, Δd_{M1}, Δd_{M2}, as presented in (16)–(19), we can derive head movement ΔX, ΔY and orientation $\Delta\theta$.

III. Implementation

In our system, the transmitter consists of a simple stereo speaker, while the receiver is implemented based on the Respeaker Kit [17]. A more detailed explanation of the system's implementation will be provided in the subsequent sections.

A. Transmitter

For the transmitter, the speaker is required to simultaneously play two sine waves. To achieve this, we utilize stereo mode to control a pair of speakers. The majority of modern audio systems support stereo mode playback, which comprises a left and a right channel. This enables the easy playback of a sine wave at frequency f_1 in one speaker and another sine wave at frequency f_2 in the other speaker. Once the speaker setup is complete, the frequencies f_1 and f_2 remain constant, allowing for the played signal to be either pregenerated or generated in real time. For convenience, we employ a pregenerated file in the form of a.wav file. Our transmitter system is compatible with most audio devices.

A laptop serves as the speaker controller, with its primary function being the playback of the pregenerated two-channel audio file. The speaker used is the "Philips SPA20," as illustrated in Fig. 8. This compact and affordable speaker operates in stereo output mode and plays sine waves at distinct frequencies. Both speakers are positioned in close proximity, with a separation distance of approximately 8 cm. This 8-cm

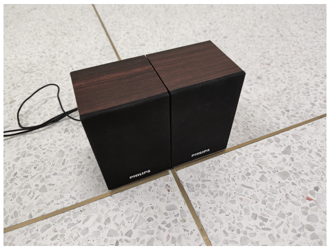


Fig. 8. Transmitter device.

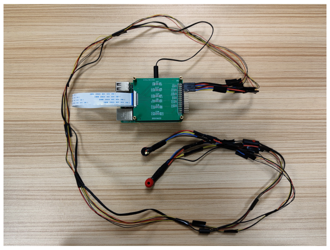


Fig. 9. Receiver kit.

separation represents the minimum distance achievable with the "Philips SPA20" speaker.

B. Receiver

We employ a Respeaker Kit [17] to facilitate two-channel recording. Currently, commercial true wireless earbuds, such as the Sony WF-1000XM3 [18], feature a microphone in each earphone. These wireless earphones, whether left or right, can operate independently. However, due to Bluetooth limitations, earphones cannot transmit dual recording signals to a smartphone simultaneously. Therefore, we modify the Respeaker Kit to achieve simultaneous two-channel recording. Our modifications involve replacing the original microphone while retaining the amplifier and ADC components of the kit, which are straightforward to implement. We anticipate that as more applications come to rely on dual earphone functionality, Bluetooth will eventually support simultaneous recording.

The implemented receiver is depicted in Fig. 9. We have designed an adapter board to connect the microphones. This design allows for the microphones to be wired out and placed freely, providing the ability to attach them to earphones. The kit can accommodate up to six microphones, although our system only requires two. In Fig. 9, two microphones (model SPU0414HR5H-SB) are connected to the board, enabling the simultaneous recording of signals from both microphones. These two microphones are attached to a pair of earphones, making them wearable for users.

IV. Evaluation

We carry out experiments to assess both movement and orientation tracking. In Section IV-A, we evaluate the accuracy of movement and orientation independently. In Section IV-B, we set up a room to deploy an exhibition scenario and assess the accuracy of attention estimation.

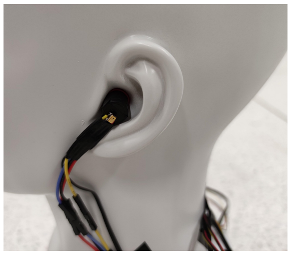


Fig. 10. Receiver attached to the head model.

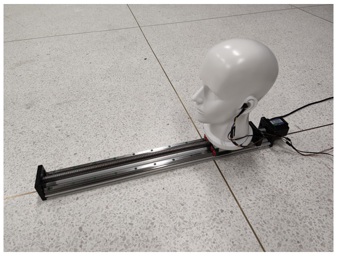


Fig. 11. Head model on the movement platform.

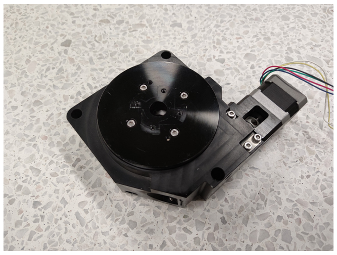


Fig. 12. Rotation platform for evaluation.

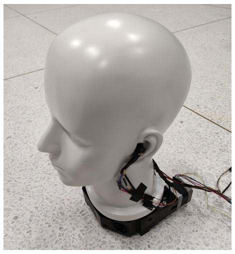


Fig. 13. Head model on the rotation platform.

A. Evaluation of Movement and Orientation Tracking

To emulate the influence of the human head, we attach the microphone to a pair of earphones and mount it on a head model, as illustrated in Fig. 10. This setup will be utilized in the evaluation of movement and orientation tracking. Additionally, to ensure an accurate evaluation, it is necessary to precisely control the movement. Consequently, we employ a linear actuator with a stepper motor for movement control, as demonstrated in Fig. 11. The platform's movement accuracy is up to 0.03 mm. For the evaluation of orientation accuracy, we use a separate rotation platform, depicted in Figs. 12 and 13, which is driven by a stepper motor. This rotation platform's accuracy is up to 0.008°.

EHTrack employs strength-based angle tracking and phase-based distance tracking to accomplish movement tracking. As

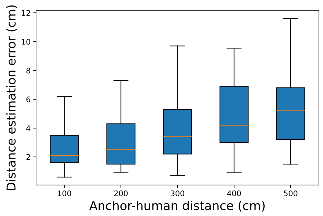


Fig. 14. Estimation error of movement parallel to the *x*-axis.

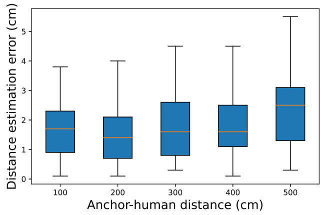


Fig. 15. Estimation error of movement parallel to the *y*-axis.

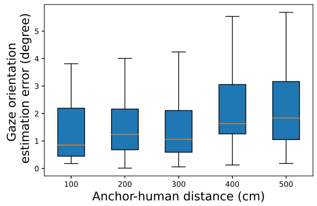


Fig. 16. Estimation error of rotation.

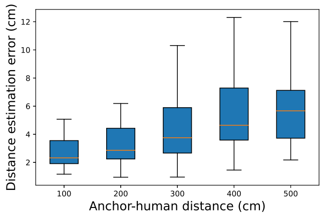


Fig. 17. Estimation error of *X*-dimension while moving.

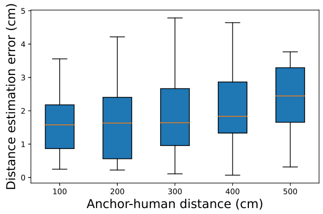


Fig. 18. Estimation error of *Y*-dimension while moving.

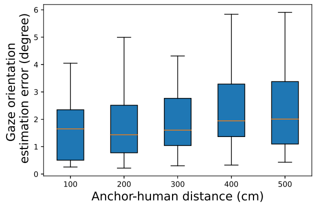


Fig. 19. Estimation error of rotation while moving.

a result, movement accuracy may vary when moving along the *X*/*Y* direction. We evaluate movement accuracy along the *x*/*y*-axis independently and subsequently assess rotation accuracy. The evaluation scenario is consistent with that of Fig. 7.

Accuracy of Movement Parallel to x-Axis: EHTrack achieves a median error of approximately 4 cm when moving parallel to the *x*-axis. We evaluate the accuracy at varying distances between the user and the speaker anchor. Fig. 14 displays the boxplot of the evaluation results. The accuracy marginally declines as the distance increases. This is because, with greater distance, the signal strength decreases due to attenuation, and the signal-to-noise ratio (SNR) also diminishes, resulting in a decrease in accuracy. However, for EHTrack, even with decreased accuracy, the estimation errors remain acceptable up to a 5-m distance.

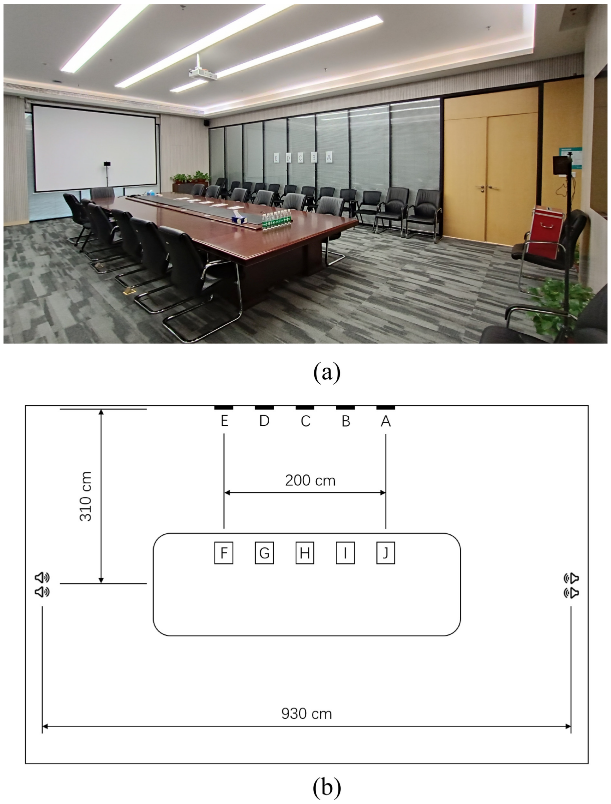
Accuracy of Movement Parallel to y-Axis: EHTrack achieves a median error of approximately 2 cm when moving parallel to the *y*-axis. Fig. 15 presents the boxplot of the evaluation results. We can also observe that movement accuracy slightly decreases as the distance increases. However, in comparison to movement parallel to the *x*-axis, the accuracy of movement parallel to the *y*-axis is higher. This is because the accuracy of movement parallel to the *y*-axis relies more on distance estimation results, while the accuracy of movement parallel to the *x*-axis depends more on angle estimation results. Generally, in terms of movement tracking, distance estimation is more accurate than angle estimation. The average movement error is 2.98 cm.

Accuracy of Orientation Tracking: EHTrack achieves a median estimation error of less than 2° in rotation, as Fig. 16 shows. The accuracy also declines as the distance increases. At farther locations, the SNR is small, causing the strength-based angle tracking to be less effective than at closer locations. The overall error of orientation tracking is 1.83°.

Sensitivity: Based on the aforementioned results in Figs. 14, 15, and 16, it is evident that our system demonstrates satisfying sensitivity within a 5-m working range. Although the tracking error does increase as the distance extends, this can be attributed to the significant attenuation of sound waves. Nonetheless, our results remain acceptable. It is worth noting that the detection range is influenced by the power of the speaker. If a larger speaker power is permitted, it would lead to an improved detection range and enhanced sensitivity.

Accuracy While Moving: In addition to evaluating the accuracy in each dimension, we assess the accuracy while the head is both moving and rotating. As Figs. 17 and 18 show, the accuracy slightly decreases compared to single-dimension evaluation, with the average accuracy being approximately 10%–20% lower due to movement interference. Also, for the rotation tracking error, as Fig. 19 shows, it suffers a decrease in accuracy, too. And due to the influence of moving, the data is not stable as tested when there is only rotation. Despite the reduced accuracy, it remains acceptable for attention estimation.

Comparison With IMU-Based Solution: We compared our system with a well-established IMU-based system, Ear-AR [8]. To ensure fair comparision, we employed a similar evaluation setup as Ear-AR, wherein a user walked within a room while their path was recorded as the ground truth, utilizing a camera for error estimation. It is worth mentioning that we chose not


(a)

(b)

Fig. 20. Evaluation scenario of attention estimation. (a) Evaluation scenario. (b) Platform layout.

to perform any halfway calibration for both EHTrack and Ear-AR in order to maintain a good user experience and enable an effective comparison between the two systems. It is worth noting that EHTrack achieves an average tracking error of 1.14 meters over a distance of 50 meters, as Fig. 21 shows, surpassing the 3-meter error exhibited by Ear-AR. However, over a distance of 100 m, EHTrack's average tracking error measures 7.56 m, slightly exceeding the 5-m error associated with Ear-AR. Nevertheless, we argue that the performance in the closer proximity range holds greater significance for the exhibition scenario we consider, as illustrated in Fig. 1, where individuals typically move around the objects on display. Additionally, EHTrack relies solely on acoustic signals, which enhances its compatibility with a wide range of COTS earphones, including Sony WF-1000XM4 and Bose QuietComfort Earbuds, as it eliminates the need for IMUs required by the Ear-AR solution.

B. Evaluation of Attention Estimation in Deployed Exhibition Scenario

To better evaluate system performance in a real head tracking scenario, we deploy an exhibition setting and assess the accuracy of attention estimation.

As depicted in Fig. 20, ten labels, ranging from “A” to “J,” were used to emulate exhibits. Five of these labels were mounted on the wall to represent paintings, while the other five were placed on a table to represent showcases.

One challenge in this setup is that a microphone may not consistently receive the sound field from a single pair of speakers. That is, non-line-of-sight (NLOS) scenes occur. Furthermore, the tracking ability of one earphone may be compromised, resulting in the system's failure to derive head movement and rotation. To address this issue, we installed two pairs of speakers in the room. Our system leverages single-frequency acoustic signals, which avoids frequency overlap and enables the deployment of multiple speakers in one room. As depicted in Fig. 20, the two pairs of speakers are positioned at opposite ends of the room to ensure comprehensive sound field coverage. Each pair operates on distinct frequencies, ensuring no interference between them. In the event of microphone occlusion, a frequency switch occurs, allowing the microphone to receive input from the alternate pair of speakers. Participants were instructed to walk around the room and view the various labels. We recorded the labels that were observed by the participants and compared these findings with EHTrack's output to determine the accuracy of attention estimation.

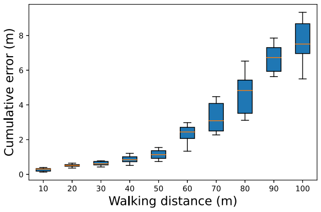


Fig. 21. Cumulative error across walking distance.

Impact of Touring Trajectory: We designed six distinct scenarios for the experiment.

1) The user remains stationary, standing without movement.
2) The user walks in a straight line, maintaining their gaze forward.
3) The user stands 2 m away from the wall-mounted labels (“A” to “E”) and focuses on them.
4) The user stands 5 m away from the wall-mounted labels (“A” to “E”) and focuses on them.
5) The user stands 2 m away from the table-mounted labels (“F” to “J”) and focuses on them.
6) The user walks to a designated location and turns to concentrate on the wall-mounted labels.

The accuracy of attention estimation for the six scenarios is presented in Fig. 22. A total of 462 attention events were recorded, yielding an overall accuracy of 89.2% In Scenario 1, all standing still events were correctly detected. In Scenario 2, 87.5% of walking events were detected. Scenarios 3 and 4 are quite similar, with the only difference being the distance from the labels. The attention estimation accuracy in Scenario 4 is lower than that in Scenario 3 because the increased distance results in lower accuracy, even when the angle accuracy remains consistent. The evaluation in Scenario 5 closely mirrors that of Scenario 3, and consequently, the accuracy is similar. Scenario 6 has the lowest accuracy at 80.6%, as both movement and rotation are present in this scenario, causing the frequency switch to generate higher tracking errors.

Impact of User: Fig. 23 displays the attention estimation accuracy for different users, ranging from 84.4% to 93.5%.

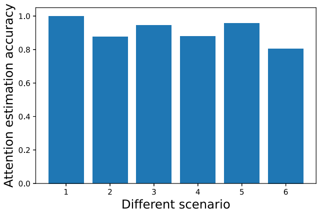


Fig. 22. Accuracy across different trajectory scenarios.

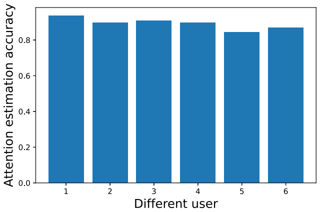


Fig. 23. Accuracy across different users.

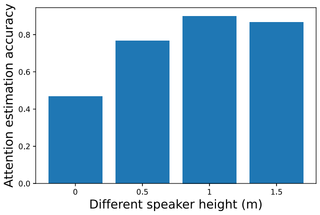


Fig. 24. Accuracy across different speaker height.

The observed variations may be attributed to users' unique walking patterns and orientation habits. Some users walk and rotate more quickly than others, leading to increased tracking errors and reduced accuracy. However, all users' accuracies exceed 80%, demonstrating the viability of this approach in an exhibition setting.

Impact of Speaker Placement: The height of the speakers may impact the sound field and, consequently, affect the system's accuracy. To evaluate the influence of speaker height on system performance, we mounted the speaker anchor at varying heights: 0, 0.5, 1, and 1.5 m. At each height, we tested 30 attention cases to determine if the system could accurately detect them. Fig. 24 presents the results for the different heights. When the speaker is at a height of 1.5 m, the system achieves an accuracy of 90%. Similar accuracy is observed when the height is 1 m. However, the accuracy decreases when the speaker is at 0.5 m. Moreover, when the speaker anchor is placed on the ground, the accuracy drops to less than 50%. In this case, the system does not perform well due to the significant height difference and interference caused by the ground. Taking into consideration that the height of a human ear (or earphone) typically falls within the range of 1.5–2 m, our results indicate that our system achieves satisfactory performance when the speaker is in close alignment with the typical height of human ears, and a significant height difference between the speaker anchor and the earphone may lead to reduced accuracy. Regarding the separation between two speakers in EHTrack, it does not exert a substantial influence on accuracy as indicated in [14]. Therefore, we believe that our system exhibits reasonable flexibility in terms of speaker placement.

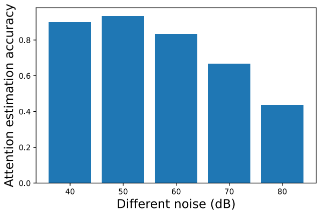


Fig. 25. Accuracy across different environment noise.

Impact of Environmental Noise Level: We also evaluated the influence of environmental noise on system performance. To simulate a noisy environment, we played loud music while the system was operating. The noise strength was measured using a decibel meter, and we maintained noise levels between 40 and 80 dB. The background noise was approximately 35 dB. Fig. 25 displays the accuracy at different environmental noise levels. As the noise level increases, the accuracy decreases.

Our system relies on acoustic signal strength, and while preprocessing is employed to remove noise, it cannot entirely eliminate the impact of environmental noise. Consequently, the system is still significantly affected by high levels of noise. However, it is important to note that noise levels exceeding 70 dB are not common in everyday activities. As such, our system can function effectively in most typical environments.

Impact of Space Size: We conducted evaluations in four spaces of varying sizes to assess the robustness of our system, taking into account the presence of different degrees of multipath effects.

Space 1: A 6 m × 10 m meeting room depicted in Fig. 26.

Space 2: A 2 m × 6 m room with an array of objects within.

Space 3: A 5 m × 5 m semi-open space.

Space 4: An outdoor space.

In Space 1, our system demonstrated impressive results, achieving an attention estimation accuracy rate of 92.1%. In Space 2, which is a more compact environment, the accuracy rate decreased to 70.3%. This reduction in accuracy can be attributed to the presence of severe multipath effects and interference, which result in significant tracking errors and degrade the accuracy of attention estimation. In Space 3, our system achieved an accuracy rate of 85.9%, surpassing the results obtained in the smaller room. During the outdoor evaluation in Space 4, our system achieved an accuracy rate of 90.6%.

EHTrack exhibits a decrease in accuracy within the smallest room, i.e., Space 2, due to the presence of clutter and pronounced multipath effects and interference in the limited space. Obstacles that completely obstruct the signals could be a nearly intractable issue for acoustic-based solutions due to the inherent limited penetrative ability of sound waves. However, we argue that our intended exhibition scenario typically encompasses relatively spacious areas, more akin to Spaces 1, 3, and 4, where EHTrack attains satisfactory performance.

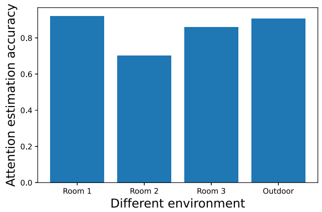


Fig. 26. Accuracy across different sized spaces.

V. Discussion and Limitation

EHTrack achieves a head tracking system based on only acoustic signals. Although EHTrack makes it possible for cheap and convenient head tracking, there are still some limitations.

EHTrack requires a map to build a relation between user heading and objects. Our system derives earphones movements relative to the speaker anchor. We should know the information of surroundings to map user's heading to a specific object.

It is important to note that EHTrack is currently only able to achieve 2-D head tracking. While our system is able to track both head movement and orientation using two earphone microphones, it is not able to track head movements in the up/down direction.

Due to limitations in Bluetooth technology, EHTrack is currently unable to be implemented on commercial wireless earphones, even though they already have the necessary hardware, including two microphones. However, we believe that as applications using both earphone microphones become more prevalent, Bluetooth will eventually support the recording of two signals.

VI. Related Works

In this section, we provide a concise review of the existing literature on head tracking and acoustic sensing. We categorize head tracking research based on the various implementation methods employed. Additionally, we discuss notable studies within the field of acoustic sensing.

A. Head Tracking

Head tracking plays a crucial role in HCI by tracking human attention and enabling computers to discern users' intentions. In this section, we discuss notable works on vision-based and IMU-based head tracking.

1) Vision-Based Head Tracking: Numerous head tracking systems are implemented using cameras [19], [20], which capture images of the human head. These images contain abundant information about the head, allowing for the derivation of head orientation and position.

Basu et al. [1] proposed a robust method for tracking rigid head motion from video. Their approach employs a 3-D ellipsoidal model of the head and interprets optical flow in terms of the possible rigid motions of the model. This method enhances robustness, even in cases of low frame rates and noisy camera images. Tan et al. [2] adopted a long short-term memory (LSTM)-based approach, wherein they utilize signals of different modalities over time to estimate continuous head orientations. HyHOPE [3] implements a head orientation and position estimation system for driver head tracking using a single video camera. The tracking module provides a fine estimate of the 3-D motion of the head. hMouse [4] further translates head tracking results into a computer mouse, enabling hands-free interaction with computers. However, camera-based head tracking methods necessitate that the user remains within the camera's view, which may lead to privacy concerns and limit the systems' deployment.

2) IMU-Based Head Tracking: With the growing prevalence of wearable devices, IMU sensors have emerged as a promising solution for head tracking. These systems determine head movement using accelerometers, gyroscopes, and magnetometers, allowing for efficient and accurate head movement derivation from a head-worn IMU.

Pedestrian dead-reckoning (PDR) has been extensively researched in recent years [21], [22], [23], [24]. These studies employ IMUs located on the foot or within a smartphone to achieve PDR, deriving body movement and orientation rather than head orientation. Some works [6], [7] utilize head-mounted IMUs for head tracking and attention detection. Ear-AR [8] further accomplishes acoustic augmented reality (AAR) through IMU-based head tracking, using IMUs in earphones and smartphones to estimate a user's location and gazing orientation. Subsequently, it delivers 3-D audio annotations to the user's ears as they move and observe AAR objects within the environment. Some eye-gaze tracking systems [25], [26] also employ IMU-based head movement compensation to enhance tracking accuracy. However, these IMU-based approaches necessitate that users wear devices containing IMUs. Given that most current earphones lack IMUs, this requirement presents a significant barrier to adoption.

B. Acoustic Sensing

Acoustic sensing involves utilizing acoustic signals for detection purposes. Many contemporary smart devices, such as smartphones and earphones, are equipped with speakers and microphones, making acoustic sensing feasible with these devices. Numerous studies have employed acoustic sensing to achieve tracking and localization.

Earphonetrack [27] incorporates earphones into the acoustic motion tracking ecosystem. EarSoundtrak [13] enables finger tracking using a speaker ring and multiple microphones. The user wears a ring equipped with a speaker that emits acoustic signals at a specific frequency. A receiver with several microphones positioned at various locations captures the signals and analyzes the phase information. The speaker's location is determined based on the phase. FingerIO [28] is another study that achieves finger tracking by transforming the speaker and microphone into an active sonar system. The speaker emits a specially designed orthogonal frequency-division multiplexing (OFDM) signal, and the microphone captures the echo produced by the human finger. By analyzing the reflected OFDM signal, 2-D finger tracking is accomplished. LLAP [15] facilitates gesture tracking through phase-based distance measurement, utilizing acoustic phase to determine movement direction and distance. For 1-D and 2-D tracking, LLAP achieves tracking accuracies of 3.5 and 4.6 mm, respectively. CAT [11] employs multiple speakers and a single microphone to enable device tracking. The speakers emit inaudible sounds at varying frequencies, which a device equipped with a microphone, such as a smartphone, receives to ascertain device movement direction and speed. CAT also incorporates an IMU to enhance performance, achieving a median error of 8 mm. Ge et al. [14] proposed a novel approach to acoustic motion tracking. Their system comprises two speakers and one microphone. The speakers emit two sine waves at distinct frequencies, creating a periodically changing sound field. A microphone within the sound field detects these changes, enabling angle tracking. Combined with phase-based distance tracking, this system successfully accomplishes motion tracking. Yang and Zheng [29] achieved head orientation estimation using two microphone arrays. As individuals speak, the sound is captured by the microphone arrays, and the energy radiation pattern is utilized to determine head orientation. Most current acoustic sensing research focuses on single-point tracking; however, this approach is insufficient for extracting head orientation. By considering two microphones together in the context of acoustic tracking, we aim to reconstruct both head movement and orientation. FaceOri [30] proposes a head orientation estimation system that measures the distance between earphone microphones and laptop speakers. Nevertheless, this method necessitates synchronization between the transmitter and receiver, making it challenging to operate on multiple devices simultaneously.

VII. Conclusion

We design and implement EHTrack, a novel head tracking system based solely on acoustic signals. By deriving the movement of the earphones, we can achieve accurate, user-friendly, and widely applicable head tracking on COST devices. Our system leverages strength-based angle tracking and phase-based distance tracking methods to obtain the movement of each earphone, and then derive a model to get accurate head movement and orientation. Our evaluation results show that EHTrack achieves an average tracking error of 2.98 cm and an orientation tracking error of 1.83°. In an exhibition scenario, we achieved an attention estimation accuracy of 89.2%. EHTrack improves the availability of head tracking, enabling its use in a wider range of scenarios. We believe that EHTrack has the potential to enable new heading-based applications in daily scenarios.

References

[1] S. Basu, I. Essa, and A. Pentland, "Motion regularization for model-based head tracking," in *Proc. 13th Int. Conf. Pattern Recognit.*, vol. 3, 1996, pp. 611–616.

[2] S. Tan, D. M. Tax, and H. Hung, "Multimodal joint head orientation estimation in interacting groups via proxemics and interaction dynamics," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 5, no. 1, pp. 1–22, 2021.

[3] E. Murphy-Chutorian and M. M. Trivedi, "Hyhope: Hybrid head orientation and position estimation for vision-based driver head tracking," in *Proc. IEEE Intell. Veh. Symp.*, 2008, pp. 512–517.

[4] Y. Fu and T. S. Huang, "hMouse: Head tracking driven virtual computer mouse," in *Proc. IEEE Workshop Appl. Comput. Vis. (WACV)*, 2007, p. 30.

[5] "Microsoft HoloLens." Accessed: Dec. 10, 2021. [Online]. Available: https://www.microsoft.com/en-us/hololens

[6] J. Windau and L. Itti, "Walking compass with head-mounted IMU sensor," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2016, pp. 5542–5547.

[7] T. Leelasawassuk, D. Damen, and W. W. Mayol-Cuevas, "Estimating visual attention from a head mounted IMU," in *Proc. ACM Int. Symp. Wearable Comput.*, 2015, pp. 147–150.

[8] Z. Yang, Y.-L. Wei, S. Shen, and R. R. Choudhury, "Ear-AR: Indoor acoustic augmented reality on earphones," in *Proc. 26th Annu. Int. Conf. Mobile Comput. Netw.*, 2020, pp. 1–14.

[9] "Sony WF-1000XM4 industry leading noise canceling truly wireless earbuds." Accessed: May 31, 2023. [Online]. Available: https://electronics.sony.com/audio/headphones/truly-wireless-earbuds/p/wf1000xm4-b

[10] "Bose QuietComfort earbuds II," Accessed: May 31, 2023. [Online]. Available: https://www.bose.com/en_us/products/headphones/earbuds/quietcomfort-earbuds-ii.html#v=qc_earbuds_ii_eclipse_grey

[11] W. Mao, J. He, and L. Qiu, "CAT: High-precision acoustic motion tracking," in *Proc. 22nd Annu. Int. Conf. Mobile Comput. Netw.*, 2016, pp. 69–81.

[12] Y. Zhang, J. Wang, W. Wang, Z. Wang, and Y. Liu, "Vernier: Accurate and fast acoustic motion tracking using mobile devices," in *Proc. Conf. Comput. Commun.*, 2018, pp. 1709–1717.

[13] C. Zhang et al., "SoundTrak: Continuous 3D tracking of a finger using active acoustics," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 2, p. 30, 2017.

[14] L. Ge, Q. Zhang, J. Zhang, and Q. Huang, "Acoustic strength-based motion tracking," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 4, no. 4, pp. 1–19, 2020.

[15] W. Wang, A. X. Liu, and K. Sun, "Device-free gesture tracking using acoustic signals," in *Proc. 22nd Annu. Int. Conf. Mobile Comput. Netw.*, 2016, pp. 82–94.

[16] Y. Liu, W. Zhang, Y. Yang, W. Fang, F. Qin, and X. Dai, "PAMT: Phase-based acoustic motion tracking in multipath fading environments," in *Proc. IEEE Conf. Comput. Commun.*, 2019, pp. 2386–2394.

[17] "Respeaker 4 MIC linear array kit." Accessed: Dec. 10, 2021. [Online]. Available: https://wiki.seeedstudio.com/ReSpeake_4-Mic_Linear_Array_Kit_for_Raspberry_Pi/

[18] "Sony WF-1000XM3 truly wireless earbuds." Accessed: Dec. 10, 2021. [Online]. Available: https://www.sony.jp/headphone/products/WF-1000XM3/

[19] S. Ohayon and E. Rivlin, "Robust 3D head tracking using camera pose estimation," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, vol. 1, 2006, pp. 1063–1066.

[20] C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau, "3D head tracking for fall detection using a single calibrated camera," *Image Vis. Comput.*, vol. 31, no. 3, pp. 246–254, 2013.

[21] Z. Xiao, H. Wen, A. Markham, and N. Trigoni, "Robust pedestrian dead reckoning (R-PDR) for arbitrary mobile device placement," in *Proc. Int. Conf. Indoor Position. Indoor Navig. (IPIN)* 2014, pp. 187–196.

[22] N. Roy, H. Wang, and R. R. Choudhury, "I am a smartphone and i can tell my user's walking direction," in *Proc. 12th Annu. Int. Conf. Mobile Syst., Appl. Services*, 2014, pp. 329–342.

[23] W. Kang and Y. Han, "SmartPDR: Smartphone-based pedestrian dead reckoning for indoor localization," *IEEE Sensors J.*, vol. 15, no. 5, pp. 2906–2916, May 2015.

[24] A. R. Jimenez, F. Seco, C. Prieto, and J. Guevara, "A comparison of pedestrian dead-reckoning algorithms using a low-cost MEMS IMU," in *Proc. IEEE Int. Symp. Intell. Signal Process.*, 2009, pp. 37–42.

[25] C.-H. Fang and C.-P. Fan, "Effective marker and IMU based calibration for head movement compensation of wearable gaze tracking," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, 2019, pp. 1–2.

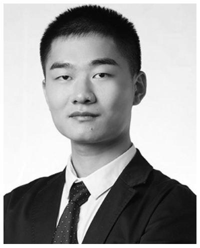
[26] T.-L. Liu and C.-P. Fan, "Visible-light wearable eye gaze tracking by gradients-based eye center location and head movement compensation with IMU," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, 2018, pp. 1–2.

[27] G. Cao et al., "EarphoneTrack: Involving earphones into the ecosystem of acoustic motion tracking," in *Proc. 18th Conf. Embedded Netw. Sens. Syst.*, 2020, pp. 95–108.

[28] R. Nandakumar, V. Iyer, D. Tan, and S. Gollakota, "FingerIO: Using active sonar for fine-grained finger tracking," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2016, pp. 1515–1525.

[29] Q. Yang and Y. Zheng, "Model-based head orientation estimation for smart devices," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 5, no. 3, pp. 1–24, 2021.

[30] Y. Wang et al., "FaceOri: Tracking head position and orientation using ultrasonic ranging on earphones," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2022, pp. 1–12.



Linfei Ge received the B.S. degree in computer science and engineering from the Southern University of Science and Technology, Shenzhen, China, in 2018. He is currently pursuing the joint Ph.D. degree in computer science and engineering with The Hong Kong University of Science and Technology, Hong Kong, and with the Southern University of Science and Technology.

His research interests include wireless sensing, acoustic sensing, and human–computer interaction.

Qian Zhang (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees in computer science from Wuhan University, Wuhan, China, in 1994, 1996, and 1999, respectively.

She is currently a Tencent Professor of Engineering and the Chair Professor with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology (HKUST), Hong Kong, where she is also serving as the Co-Director of Huawei–HKUST Innovation Lab and the Director of Digital Life Research Center. Before that, she was with Microsoft Research Asia, Beijing, China, where she was the Research Manager with the Wireless and Networking Group. She has published more than 400 refereed papers in international leading journals and key conferences in the areas of wireless/Internet multimedia networking, wireless communications and networking, wireless sensor networks, and overlay networking. She is the inventor of more than 50 granted international patents. Her current research interests include Internet of Things, smart health, mobile computing and sensing, wireless networking, as well as cybersecurity.

Jin Zhang (Member, IEEE) received the B.E. and M.E. degrees in electronic engineering from Tsinghua University, Beijing, China, in 2004 and 2006, respectively, and the Ph.D. degree in computer science from The Hong Kong University of Science and Technology, Hong Kong, in 2009.

She is currently an Associate Professor with the Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China. Her research interests are mainly in wireless sensing and mobile computing, wearable Computing and mobile healthcare, mobile crowdsensing, and blockchain.

Huangxun Chen (Member, IEEE) received the B.S. degree in computer science and technology from Shanghai Jiao Tong University, Shanghai, China, in 2015, and the Ph.D. degree in computer science and engineering from The Hong Kong University of Science and Technology, Hong Kong, in August 2020.

She has been a Researcher with the 2012 Labs, Huawei Hong Kong Research Center, Hong Kong, since October 2020.