Ear-AR: Indoor Acoustic Augmented Reality on Earphones

Zhijian Yang
University of Illinois at Urbana Champaign
zhijian7@illinois.edu

Yu-Lin Wei
University of Illinois at Urbana Champaign
yulinlw2@illinois.edu

Sheng Shen
University of Illinois at Urbana Champaign
sshen19@illinois.edu

Romit Roy Choudhury
University of Illinois at Urbana Champaign
croy@illinois.edu

ABSTRACT

This paper aims to use modern earphones as a platform for acoustic augmented reality (AAR). We intend to play 3D audio-annotations in the user's ears as she moves and looks at AAR objects in the environment. While companies like Bose and Microsoft are beginning to release such capabilities, they are intended for outdoor environments. Our system aims to explore the challenges indoors, without requiring any infrastructure deployment. Our core idea is two-fold. (1) We jointly use the inertial sensors (IMUs) in earphones and smartphones to estimate a user's indoor location and gazing orientation. (2) We play 3D sounds in the earphones and exploit the human's responses to (re)calibrate errors in location and orientation. We believe this fusion of IMU and acoustics is novel, and could be an important step towards indoor AAR. Our system, *Ear-AR*, is tested on 7 volunteers invited to an AAR exhibition – like a museum – that we set up in our building's lobby and lab. Across 60 different test sessions, the volunteers browsed different subsets of 24 annotated objects as they walked around. Results show that *Ear-AR* plays the correct audio-annotations with good accuracy. The user-feedback is encouraging and points to further areas of research and applications.

CCS CONCEPTS

• Human-centered computing → Ubiquitous and mobile computing systems and tools; **• Computer systems organization → Embedded and cyber-physical systems**; **• Information systems → Multimedia information systems**; **Mobile information processing systems**.

KEYWORDS

Augmented Reality, Acoustics, Smart Earphones, Wearable Computing, Indoor Localization, Inertial Measurement Unit (IMU), Dead Reckoning, Motion Tracking, Sensor Fusion, Head Related Transfer Function (HRTF), Spatial Audio

ACM Reference Format:
Zhijian Yang, Yu-Lin Wei, Sheng Shen, and Romit Roy Choudhury. 2020. Ear-AR: Indoor Acoustic Augmented Reality on Earphones. In *MobiCom 2020 (MobiCom '20), September 21–25, 2020, London, United Kingdom.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3372224.3419213

1 INTRODUCTION

Acoustic Augmented Reality (AAR) is the ability to overlay acoustic information on physical reality. Imagine a scenario as follows. When visiting a museum, Alice's earphone narrates the history of paintings as she pauses to look at them. The paintings do not have any form of beacons or codes attached; instead, the earphone tracks Alice's indoor location and gazing orientation, and using a known map of museum exhibits, infers what Alice is looking at.

Later, when Alice asks her earphone to guide her to another gallery, the earphone plays a 3D voice that says "follow me". The voice signal is carefully designed and played across the two earphones so that it appears to come from the direction in which Alice should walk. Alice simply follows the *perceived direction* of the voice and reaches the gallery; she does not pull out her phone, nor checks for maps or signposts in the museum. This interplay of (spatial) movement with (spatial) sounds allows for such an AAR experience. One could imagine other applications as well, such as playing AAR games in a school building, finding people in a crowded place, or even military scenarios where ground troops coordinate via 3D acoustics. This paper uses the theme of a museum throughout the paper for ease of explanation; the core technical problems generalize across scenarios.

Companies like Bose and Microsoft are actively engaging in AAR and beginning to roll out products [20, 55, 66]. Apple also announced spatial audio [9] in its latest release, the Airpods Pro [8]. However, the current services are in their early stages and intended for simple applications where indoor localization is not needed. Bose regards indoor AAR as a high priority research area [34], and is pursuing infrastructure-based indoor AAR solutions [35]. This paper considers an *infrastructure-free* approach, building only on earphones and smartphones that are commonly carried by users.

From a research perspective, an AAR system requires 3 pieces to come together, namely (1) tracking Alice's indoor location, (2) tracking Alice's head orientation, and (3) designing 3D sounds that appear to come from a desired direction. Figure 1 illustrates an example. To navigate Alice to her requested gallery, the AAR system uses Alice's location and head orientation to infer that the 3D sound should arrive from an angle θ from her gazing direction.

MobiCom '20, September 21–25, 2020, London, United Kingdom

ACM ISBN 978-1-4503-7085-1/20/09…$15.00
https://doi.org/10.1145/3372224.3419213

The "follow me" voice signal is then designed as a function of this θ.

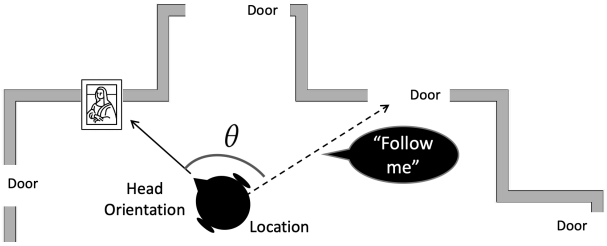

Figure 1: Example museum scenario for indoor AAR.

Of the 3 problems above, head orientation and 3D sound design have largely been solved in literature [47, 64]. Bose has rolled out products [21] that play annotations about landmarks, say the Eiffel Tower, when a user turns her head towards it. Oculus and Dolby are also using head orientation and 3D sounds to enable new immersive experiences in virtual reality (VR).

On the other hand, the problem of indoor localization remains elusive. While the topic has also been studied widely (including IMU-based tracking [16, 19, 24, 26, 30, 37, 42, 67, 72]), it is still an open research problem. An infrastructure free system which is easily deployable is especially difficult because IMU dead reckoning is known to be inaccurate and there are few chances to calibrate. This paper sees a new opportunity through earphones. The opportunity has 2 parts:

1. The placement of the IMU near the ear presents a crucial advantage in tracking human location. This advantage was largely missing when the IMU was on smartphones and wrist-worn devices.
2. The human brain's ability to sense the direction of sound (played by the earphone) offers a new form of opportunity to correct IMU estimation errors.

We expand on each of these ideas next.

(1) Observe that IMU-based localization is the problem of estimating the displacement of the body/torso while a user is walking. It involves estimating 3 components: number of steps, step length, and walking direction. With the current phone or watch IMU, none of the problems are easy to solve reliably because these IMUs are severely affected by activities of the limbs, such as motions of the arm, legs, typing, talking, etc. The IMU senses the "sum" of all these motions and accurately separating the body's displacement from the limb's interferences has almost been impossible.

This paper finds that earphone IMUs, by virtue of being on the upper extreme of the body, receive a surprisingly clean signal. This signal is free from limb interference, and yet, preserves crucial features of body motion. In some sense, *the human skeleton/muscles serve as a low pass filter for its own lower-body motion; only macro motions – like the torso's up/down bounce – propagate to the ear.* This clean upper-body signal, in collaboration with the smartphone's IMU, reveals geometric parameters of the human body skeleton, ultimately translating to better localization. *Ear-AR* exploits these geometric relationships between the motions of the leg and the head and there is no reliance on training or user-data. Hence, the solutions are expected to scale to different users and walking patterns.

(2) Unfortunately, any IMU based tracking is bound to diverge over time [19] – a better design only *slows* down the divergence rate. This is fundamental because IMU measurements are in a local reference frame, hence there is no way to correct the true trajectories. To understand this through an analogy, consider a blind-folded person trying to walk in a straight line. She has no way of telling whether, and how much, she has drifted; hence, she can never correct for it. Her only hope is to hear or touch something in the environment that reveals her drift in the global reference frame, offering an opportunity to reset. Similarly, in this paper, *Ear-AR* utilizes the interplay of directional sounds and true reality as a reset opportunity. The idea is as follows.

Assume *Ear-AR*'s location estimate has drifted. Now, assume the user is near an audio-annotated object and begins to hear a directional sound from the earphone. Due to the location drift, this direction would be incorrect, i.e., if the user looks exactly in the sound's direction, the object would not be present in the line of sight. However, if the drift is not large (or if the audio describes the object) the user should still be able to correctly identify the object and look at it. This offers the core opportunity for correction. *Ear-AR* calculates the angular offset between the sound direction and the user's actual gazing direction, and uses this offset for resetting the drift. Thus, encounters with annotated objects periodically reset the IMU drifts, permitting a continued AAR.

In building *Ear-AR*, we use a wireless Beats headphone attached with an external IMU (see Figure 2(a)), and a Samsung smartphone. IMU data from both devices are streamed to a laptop, which first estimates the user's location and gazing direction, then synthesizes the 3D directional sound, and finally transmits the sound to the headphone when the user is looking at an annotated object. For experiments, we *pretend* the indoor environment is a museum with annotated objects on the walls/shelves (Figure 2(b)). We assume the object locations $<X_i, Y_i, Z_i>$ are known, so *Ear-AR* knows when a user is in the vicinity of an object.

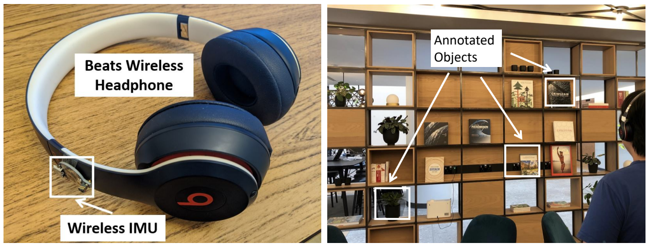

Figure 2: *Ear-AR* evaluation: (a) Beats wireless headphone taped with a wireless IMU. (b) Objects annotated in our lobby send out directional sounds.

During tests, 7 volunteers were asked to wear our *Ear-AR* headphones, carry the smartphone in pocket, and stand at a known starting location. As they begin walking, they hear directional voices that say "find me". Following this voice, a volunteer's task is to mark a location on the wall or shelf that she believes is the annotated object's location (Figure 2(b)). She repeats this process for 5 or 7 "find me" voices in one session. Volunteers perform 60 sessions in total.

We plot the error between the user-marked locations and actual object locations (in addition to various other micro-benchmarks). While exact results depend on configurations, the broad finding is that: *users can identify the correct object (and listen to the correct audio annotation) with >90% accuracy* when they opportunistically reset their locations using spatial audio. Importantly, the IMU drift is slower compared to state-of-the-art schemes and sound-based calibrations are effective. Thus, all in all, the contributions in this paper may be summarized as follows.

- *We recognize that IMUs at the ear is an important vantage point for measuring human motion*, and utilize this opportunity (in conjunction with the smartphone IMU) for indoor localization. We utilize the earphone IMU for head orientation as well, a bonus.
- *We leverage the human brain's 3D sound tracking ability to correct/refine IMU's motion tracking errors*, enabling a sustainable AAR experience. We build a functional proof of concept, evaluated with real volunteers in a (pretend) museum setting.

The rest of the paper expands on each of these contributions, beginning with foundations and measurements, followed by system architecture, design, and evaluation.

2 BASICS OF IMU-BASED TRACKING

Consider the classical problem of IMU-based human localization (also called *pedestrian dead reckoning*, PDR). Figure 3 shows the accelerometer data from a smartphone when a user is walking. The PDR problem is to accept such data as input and output a walking trajectory of the user.

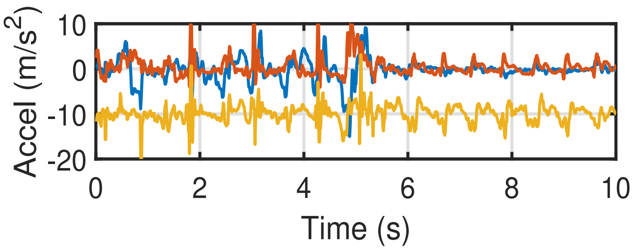

Figure 3: Accelerometer data from smartphone IMU when a user is walking.

Solving PDR translates to inferring 3 parameters from IMU data: step-count n, step-length l, and walking-direction θ. Each of the n steps can be modeled as a vector of magnitude l_i and direction θ_i. Adding n step vectors produces the user's location. Let us intuitively discuss the difficulties in estimating these 3 parameters.

• **Step counting (n)** is the easiest of the 3 problems because walking produces an up/down bounce on the body, which manifests in the phone's IMU[1]. Hence, today's techniques essentially filter out a smooth sinusoid from Figure 3 and count the number of peaks in it [12, 44, 62, 65]. Some difficulties arise when users perform some gestures with the phone (e.g., checking messages). These gestures pollute the sinusoid causing errors in step-counting.

• **Step length (l)**, defined as the forward displacement of the feet for each walking step, is the hardest problem [50, 69, 72]. The major challenge arises from not having a fixed reference frame. This is because the phone's local $\langle X, Y, Z\rangle$ axes are constantly changing due to the swing of the leg, hence the notion of forward is unclear. Moreover, even if this is resolved, estimating displacement δ from acceleration requires a double integration as follows:

$$\delta = v_0\Delta t + \iint_0^{\Delta t} A(t)dt \;= v_0\Delta t + \iint_0^{\Delta t} \big(a^*(t) + n(t)\big)dt$$

Here, v_0 is the initial velocity at the start of a step, $A(t)$ is the measured accelerometer signal, composed of the body's acceleration $a^*(t)$ plus noise $n(t)$ from limb gestures and IMU hardware. Observe that (1) the double integration causes the noise pollution to grow quadratically, severely affecting displacement. (2) Error incurred in the first step gets carried over to the second step through an incorrect v_o. There is little hope to reset these (accumulating) errors since there is no global/independent means of learning the truth. Today's state of the art approach is to train step length as a function of acceleration and step frequency [72]. However, given the position of phone IMUs, even getting an accurate acceleration or step frequency measurement is hard. Moreover, requiring per-user training data is an additional burden.

• **Walking direction** is difficult as well because θ needs to be expressed in a global reference frame (e.g., relative to North) [15, 23, 26, 67, 73]. But again, since IMU measurements are in its local $\langle X, Y, Z\rangle$ reference frame, and since this reference frame keeps rotating due to the leg's swing, it is hard to continuously map the motions to the global framework. Any error (due to noise) will again accumulate, causing the estimated trajectory to diverge over time. Even if we can track the phone's orientation to some degree, it is still difficult to translate this orientation to the user's walking direction. This is because human motions are complex aggregates of swings, up/down bounces, sideward sways, and various jitters – the walking direction is a 2D vector hidden in this complex mix of signals. Extracting out this vector is challenging.

3 ENABLING OPPORTUNITIES

IMUs in the earphones, and the human brain's ability to sense directional sound, are 2 enabling opportunities. Let us discuss them at an intuitive level (with some basic measurements).

[1] Naturally Filtered Signal at the Earphone IMU

Figure 4 plots the accelerometer signal from the earphone from the same scenario as in Figure 3. The advantage is evident – the earphone IMU captures a clean up/down bounce from the walking motion and high frequency randomness from the lower body is almost absent.

[1] For a smartwatch, the swing of the arm also produces such an up/down motion that strongly correlates to each step.

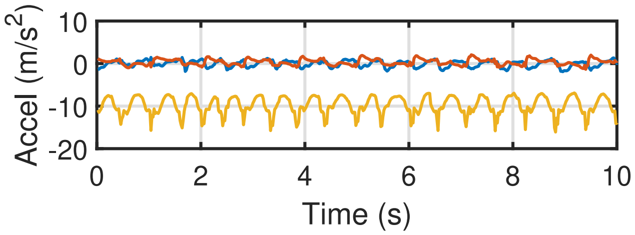

Figure 4: Accelerometer data from the earphone IMU during walking and performing arm gestures.

Importantly, the advantage of this clean earphone signal goes beyond step counting. As detailed later, the upper-body mounting location and the cleanliness of this earphone signal helps to compute the *vertical displacement* of the head when the user is walking. This vertical displacement is implicitly linked to the step length of the user, hence, knowing one leads to the other. This is the key enabler. In contrast, smartphone IMUs can count steps reasonably well, but any kind of displacement measurement – a double integration of acceleration – is highly erroneous, since unlike head IMUs, there are no chances to calibrate (as further detailed in section 5.1).

[2] 3D Sound Resolution in Humans

Human brains measure the *time difference of arrival* (TDoA) and amplitude difference across the two ears to estimate a sound's source location. We intend *Ear-AR* to leverage this capability by artificially injecting delays and amplitude differences in the earphone sounds, creating an illusion that the sound is coming from a specific direction. Such sounds are called "binaural" audio [57], referring to the 2 ears through which they must be played. *Ear-AR* wants to exploit binaural sounds not only for directional voice annotations, but also to correct localization error.

Figure 5(a) plots the angular resolution at which humans perceive sound, while Figure 5(b) plots how synthesized sounds (played through earphones) can approximate it. For Figure 5(a), we blindfolded a volunteer in the middle of a room and played speech signals on a speaker from different angles. We asked him to point a laser pointer in the *perceived direction* of the speaker. Then we repeated the same with synthesized binaural sound from the earphones. The confusion matrices are comparable, implying that artificially produced directional sounds are effective.

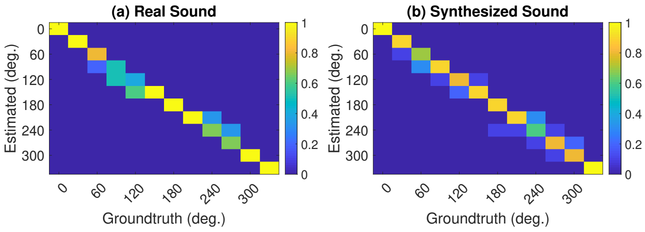

Figure 5: Confusion matrix of human's angular resolution for (a) real sound, and (b) synthesized sound.

Building on these opportunities, we design the *Ear-AR* system. We begin with a high level system architecture before describing the technical details and implementation.

4 SYSTEM ARCHITECTURE

Figure 6 illustrates *Ear-AR*'s overall architecture, composed of 2 main modules. **(1) Motion Tracking** estimates gazing direction from the head's rotation, and the user's location from the walking patterns. **(2) IMU Acoustics Sensor Fusion** combines the human's 3D binaural capabilities with motion tracking to bring together a practical AAR system.

4.1 Motion Tracking

This module receives the IMU sensor data as input and outputs a user location $\langle x, y, z \rangle$, and a gazing direction, θ. The input data is 12 streams of IMU, i.e., 3-axis accelerometer and 3-axis gyroscope data from an earphone and a smartphone. *We do not use the magnetometer since it is easily polluted by ferromagnetic materials in the environment, causing many past proposals to suffer from unpredictable errors* [11].

■ **Gazing Direction** refers to the direction in which the user is looking. Since precise eye-tracking is difficult from earphones, we assume the head's direction approximates the gazing direction. *Ear-AR* tracks the head orientation from the earphone's gyroscope and calculates a vector emanating outward from the center of the face. If this vector extends and intersects with an annotated object, *Ear-AR* assumes the user is looking at that object. In the interest of space, we omit discussing the methods to infer gazing direction since these are established techniques from literature [64, 73].

■ **User Location** is derived from step count, walking direction, and step length estimation modules. We leverage the earphone IMU for step count and step length, while we develop a new technique for walking direction. Through all these techniques, the IMUs from the earphone and smartphone are carefully fused (since none of them are individually adequate). The outcome is a continuous estimate of user location and gazing direction.

4.2 IMU Acoustics Sensor Fusion

This module receives 4 inputs – the user location and gazing direction from the motion tracking module, and the *object locations* and *annotations* from an AAR database. The outputs are 3-fold: (1) A binaural filter to create a directional version of any given annotation. (2) Corrected user location after opportunistic IMU recalibration. (3) Location and annotation of new objects that are not in the AAR database.

■ **Binaural Filtering** modifies a given sound annotation to produce two different versions for the two ears. The modification (or filtering) is a function of the relative direction (and distance) between the object and the user. The filter includes the attenuation and reverberations due to the shapes of the ears and head. As the user moves and turns, the binaural sounds are modified in real time, so the user always hears the sound coming from an absolute source location.

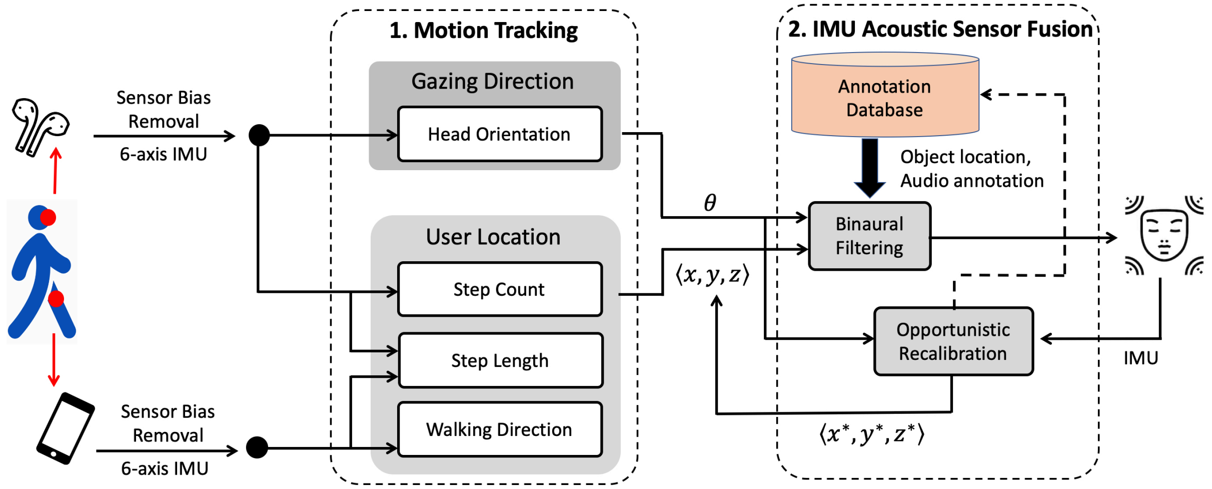

Figure 6: *Ear-AR* is composed of 2 main modules: Motion Tracking and (Application-Specific) Refinements. Motion tracking entails user localization and estimating gazing direction by fusing the earphone and smartphone IMU. Refinements pertain to calibrating IMU drift through binaural sounds and AAR-specific opportunities.

■ **Opportunistic Recalibration.** This module measures the angular offset between the *expected* gazing direction of the user (due to the binaural sound) and her *actual* gazing direction (towards the correct object). This offset reveals the inertial drift of the motion tracking module, and via geometry, corrections are injected in the location estimates. Performed periodically, this slows down the IMU drift, making the system robust and practical.

■ **Object Annotation.** We assume that object locations and annotations are available from a database. However, we propose an optional method for users to annotate objects on the fly. *Ear-AR* uses 3D geometry to calculate the object's relative location from the user, and combined with the user's own location and orientation, the object's location is inferred (detailed in Sec. 5.2). With this overview in place, let us now discuss *Ear-AR* design.

5 *EAR-AR*: SYSTEM DESIGN

We begin the section with algorithmic details on motion tracking, followed by sensor-fusion optimizations, and finally end with engineering details on the overall system.

5.1 Part I: Motion Tracking

Figure 7 illustrates the dynamics of human walking. Time $t = t_1$ is when the right foot just leaves the ground, and $t = t_3$ is when the right foot just lands back on the ground. During this single step, the left foot is fixed, and the left leg undergoes an inverted pendulum motion, i.e., rotation around the left foot hinged on the ground. A phone in the pocket senses this rotational motion, like an arc. Of course, the upper body (i.e., hip, torso, and head) also follow this arc and these arcs repeat with each step. Note that each arc decomposes into a horizontal forward motion and up-and-down oscillation. The earphone senses both on its accelerometer.

■ **Estimating Walking Direction**
Our key intuition is that the smartphone measures the inverted pendulum rotation through its gyroscope. The axis of this rotation can be computed, and given that the walking direction is a 90° rotation of this axis on the horizontal plane, it should be possible to derive it. Observe that past work on walking direction is limited. Existing proposals either use PCA-like statistical approaches that cannot estimate each individual step [15, 23, 26, 73], or rely on the phone accelerometer which is polluted by noise and jitter [67]. *Ear-AR*'s method is robust and applies to each single step.

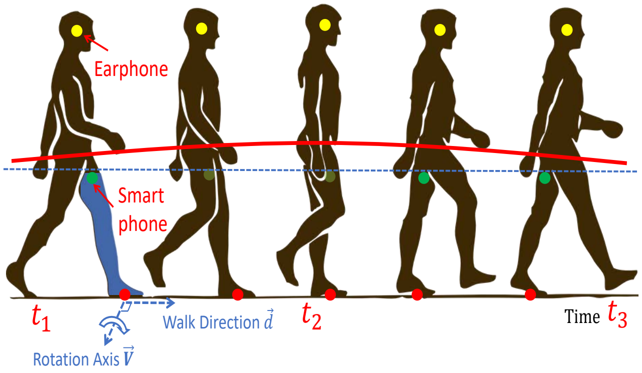

Figure 7: Left leg moves like an inverted pendulum, hinged around the left foot. Hence the head moves in an arc which equals a horizontal plus up/down motion. The earphone IMU senses these motions.

To realize our idea in *Ear-AR*, we first identify time instants t_1 and t_3, i.e., the beginning and end of a step. This is straightforward from the earphone IMU since each step – marked by a strike of the foot on the floor – produces a clear acceleration peak.We then compute the leg's delta rotation, $\Delta\mathbf{R}$, between t_1 and t_3, by integrating the smartphone's gyroscope measurement:

$$\Delta\mathbf{R} = \prod_{t=t_1}^{t_3} \Delta\mathbf{R}_t^{\text{gyro}} \tag{1}$$

where $\Delta\mathbf{R}$ and $\Delta\mathbf{R}_t$ are 3X3 rotation matrices. Finally, we convert $\Delta\mathbf{R}$ into a rotation axis $\vec{\mathbf{V}}$ and a rotation angle θ, a standard mathematical operation:

$$\langle\vec{\mathbf{V}}, \theta\rangle = \text{RotMat2AxisAngle}(\Delta\mathbf{R}) \tag{2}$$

The rotation axis $\vec{\mathbf{V}}$, as shown in Figure 7, is exactly the user's left/right direction. Rotating it by 90° gives the user's walking direction, $\vec{d}$.

When the phone is in hand, and the arm/hand performs random gestures, motion tracking is harder. However, if the user carries the phone in the pocket once, thereafter the earphone's IMU is enough. We will revisit this in Section 5.3 once we have discussed step length estimation next.

■ Estimating Step Length

Obtaining step length (i.e., the foot displacement) is a much harder problem. This requires double-integration of the accelerometer, which suffers from heavy noise accumulation and lack of information about the initial velocity (as discussed in the equation in Sec.2). The key to solving this is to identify stationary instants at which the velocity is zero, and calibrate double-integration methods at these times. Previous researchers benefit from shoe IMUs [63] because when the foot is on the ground, they obtain a good opportunity of "zero-velocity". However, this is not the case for the movement of phones or earphone IMUs – there is not a single static point during human walking for these sensors. However, opportunities appear if we examine the vertical and horizontal movements separately.

This is where the earphone brings unique opportunities. As shown in Figure 8, when the user walks, the head moves horizontally forward (hence no static instants), but vertically, it moves up-and-down periodically. The head is at its highest position when two legs are vertically straight (time t_2 and t_4); and at its lowest position when both feet hit the ground (time t_1, t_3, t_5). At either the highest or the lowest position, the head's vertical velocity is exactly 0, serving as a "landmark" for periodic calibration.

Unfortunately, this calibration opportunity is available only in the vertical dimension, whereas step length is the horizontal displacement. *But luckily, horizontal and vertical motion are strongly dependent as they are both an outcome of the (inverted pendulum-like) leg swing.* From Figure 8, we can derive the following relationship across these parameters: head's vertical movement, δh; leg swing angle, θ; leg length, L; and step length, l:

$$l = L \times \sin\theta \times 2 \tag{3}$$

$$\delta h = L - L\cos\theta \tag{4}$$

Combining these two equations, we compute step length as:

$$l = 2 \times \delta h \times \frac{\sin\theta}{1 - cos\theta} \tag{5}$$

In solving this equation, we already have obtained θ from Equation (2). To obtain δh, we perform double integration on the earphone's vertical Z axis during each step:

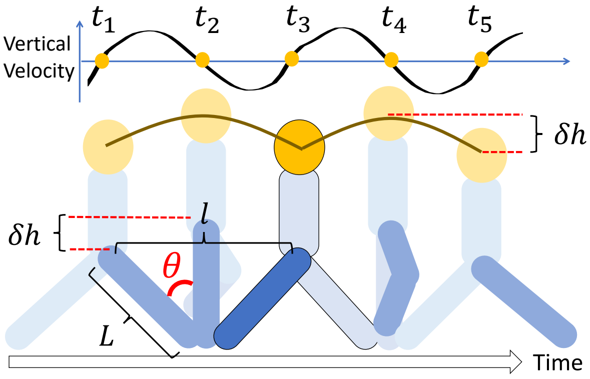

Figure 8: Right leg shown in dark blue, swings for angle θ, leading to half step length $l/2$, and head's up-down motion of δh. The earphone IMU's vertical velocity is 0 when head is at highest or lowest position.

$$\delta h = \int_0^T v_z(t)dt; \qquad v_z(t) = \int_0^t a_z(t)dt \tag{6}$$

To leverage the calibration opportunity, we force the starting vertical velocity $v_z(0)$ and the ending vertical velocity $v_z(T)$ to be both 0. Of course, integration error makes the end velocity not exactly 0 – we evenly distribute the offset back over time:

$$v_z^{\text{corrected}}(t) = v_z(t) - \frac{t}{T}v_z(T) \qquad \text{for } 0 \le t \le T \tag{7}$$

As a result, our vertical displacement, δh, is estimated much more accurately. Now, inserting θ and δh into Equation (5), we have obtained step length l. Together with walking direction $\vec{d}$, we are now able to track the user location as a vectorial addition over time: $\text{Loc}(t) = \sum_t |l_t| \, \vec{d_t}$.

5.2 Part II: IMU Acoustic Sensor Fusion

We introduce binaural filtering (not our contribution) followed by how fusion between IMU and binaural acoustics can optimize user localization.

■ Binaural Filtering

Figure 9 illustrates the core idea in human binaural hearing. The ear closer to the sound source receives audible signals with relative smaller delay and stronger intensity. Moreover, the sound undergoes different echoes and attenuation due to the shapes of the ears and head. Together, these effects are modeled as a function called *Head Related Transfer Function* (HRTF) [27] – a standard technique from literature. *Ear-AR* uses a global HRTF filter from a public medical dataset [5] and generates the binaural sound as a function of the object's relative distance and angle from the human's gazing direction.

To preserve the sound's absolute location during walking and turning, *Ear-AR* re-synthesizes the sound in small time windows (25ms). The boundaries of the windows are smoothened using a low pass filter. Eventually, *Ear-AR* is able to maintain 30Hz update rate for binaural sound, even if the user keeps moving constantly.

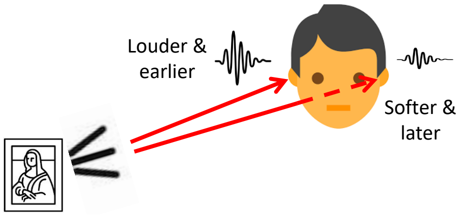

Figure 9: Sounds differ at two ears in (1) delay and (2) intensity, permitting sound-source localization.

■ Opportunistic Recalibration

It is widely agreed that IMU based motion tracking is bound to diverge over time[14, 19, 31, 38, 79]. Yet, IMUs are valuable since they do not rely on any external service or infrastructure. Generally, the way to harness such IMU-based techniques is by curbing the divergence through (periodic) error recalibration. *Ear-AR* proposes two forms of recalibration: (1) *soft recalibration*, which exploits the users' responses to binaural sounds, and (2) *hard recalibration*, performed when the user stands on known landmarks on the ground.

Figure 10 illustrates the key idea in soft recalibration. Say the true user location is right in front of the painting, but *Ear-AR*'s estimated location is shifted to the left due to IMU drift. Thus, when *Ear-AR* generates the binaural sounds in the earphone, it will suffer a $\Delta\phi$ angular error. The user will end up looking away from the object (shown by the solid black arrow labeled "Binaural Direction"). However, assuming $\Delta\phi$ is not large (i.e., the actual object is modestly close to the incorrect binaural direction), we expect the user to still recognize and look at the correct object. This is because annotated audio normally has information about the object; so by listening to the audio content alone, the user should be able to identify the object[2]. Now, the difference between the user's actual gazing direction (known from the earphone IMU) and the binaural direction (derived from the user's estimated location and the known object location) reveals the $\Delta\phi$. This provides the core recalibration opportunity. If soft recalibration is performed periodically, say in an AAR gallery or a museum, the location drift can be bounded.

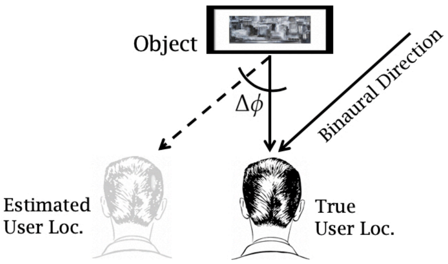

Figure 10: Opportunistic recalibration when user identifies object correctly and looks at it.

The technical steps in soft recalibration is simple. Observe that the true user location lies on a ray emitted from the object, in a direction that is opposite to the user's gazing direction (see Figure 11):

$$\langle x_{\text{user}}, y_{\text{user}}, z_{\text{user}}\rangle = \langle x_{\text{obj}}, y_{\text{obj}}, z_{\text{obj}}\rangle - \alpha\langle \vec{F_x}, \vec{F_y}, \vec{F_z}\rangle \quad (8)$$

where $\vec{F}$ is the true gazing direction. α is a coefficient representing the length of the ray. Since the binaural direction and $\Delta\phi$ are both known, the true gazing direction $\vec{F}$ can be calculated, hence α is the only unknown. Now, assuming the user's height is known, α can be derived as:

$$z_{\text{user}} = H_{\text{user}} = H_{\text{obj}} - \alpha\vec{F_z} \quad (9)$$

Combining Equation (8) and (9), we can derive the exact user location for recalibration.

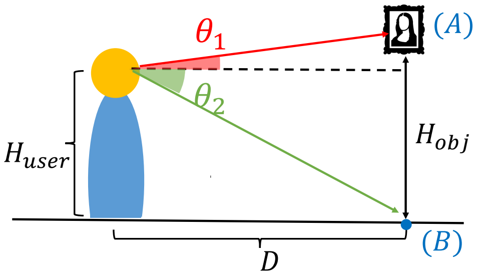

Figure 11: For soft recalibration, the user looks at an object (with known location) with elevation angle θ_1. *Ear-AR* can then correct user location.

Of course, applications where the user cannot identify the correct object (e.g. navigation in a museum or airport where a 3D voice escorts the user saying "follow me"), soft calibration is not feasible. In such settings, it is necessary for hard recalibration landmarks to be deployed in the environment. One example is to have arrow-head stickers on the floor, so that users can periodically stand on these landmarks, align their gazing direction with the arrow direction, and actively ask the earphone to recalibrate. This will recalibrate both user location and gazing direction.

■ Object Annotation

As an optional feature in *Ear-AR*, we want users to annotate objects on the fly by gazing at it and recording a voice clip. Thus, *Ear-AR* needs to estimate the object location from the knowledge of user's location and gazing direction. To this end, we propose a method that requires very little user effort. The core idea is similar to *Ear-AR* soft recalibration. As shown in Figure 11, the user will first look at the object (point A), then looks at the projection of the object on the ground (point B). Assuming the head elevations in these two cases are θ_1 and θ_2, we have the following geometrical relationship:

$$\tan\theta_1 = \frac{H_{\text{obj}} - H_{\text{user}}}{D} \qquad \tan\theta_2 = \frac{H_{\text{user}}}{D} \quad (10)$$

where H_{user} and H_{obj} are the heights of the user and the object, and D is the horizontal offset between the object and the user. Since θ_1 and θ_2 are the two elevation angles from gazing direction, and

[2] For instance, the annotated audio can be"The Mona Lisa was painted by Leonardo Da Vinci in 1517".

H_{user} is the known user's height, we can then compute H_{obj} and D, respectively. Finally, the object location can be computed as:

$$\langle x_{\text{obj}}, y_{\text{obj}}, z_{\text{obj}} \rangle = \langle x_{\text{user}}, y_{\text{user}}, z_{\text{user}} \rangle + \langle \vec{D}_x, \vec{D}_y, H_{\text{obj}} \rangle \tag{11}$$

The computed object location, plus the user's audio annotation, can now be stored in the *Ear-AR* annotation database.

5.3 System-Level Questions

■ **Correlated leg and head motion enables step length estimation, but what if the phone is in the hand?**
Arm and hand movements can be arbitrary, hence, motion tracking is difficult with the phone in hand. However, if the phone has been placed in the pocket once, *Ear-AR* estimates the leg-length L. Since the length does not vary, the phone's IMU is not needed anymore. From Equation 4, θ can be estimated directly from L and δh (from the earphone IMU). Thus, *step length* $l = 2L \sin\theta$. Finally, we will estimate *walking direction* using PCA on the earphone IMU, which is better than the smartphone IMU (polluted by the user's hand gestures).

■ **Will localization error slowly drift over time? How often does the user need to calibrate**
We are pushing the performance of IMU dead reckoning, however any kind of dead reckoning will still drift eventually. How often users need to calibrate is naturally a function of how fast their location estimations drift. Evaluation results (Figure 20) will show that *Ear-AR*'s localization error grows to ≈ 2.5m after 50m of walking, and ≈ 3.3m after 100m. This means, to identify objects that are 2.5m apart, users can walk uncalibrated for 50m. In a normal indoor *Ear-AR* setting, there are frequent calibration opportunities, (e.g., a 3D audio in the airport saying "Free WiFi at Pete's coffee"). Users can use their own binaural sensing capabilities to calibrate. In a long corridor-like environment, where there are not enough audio annotated objects, we will need hard calibration points (e.g., floor stickers) or utilize some sensor landmark opportunities like exploited in UnLoc[71].

■ **Will human's ability to perceive binaural sound affect performance?**
Binaural audio offers hints about where the target object is located. In reality, these audios should have the object description, making users unlikely to identify the wrong object during calibration. In the rare case where the user does calibrate using the wrong object, the error correction would get delayed until she encounters the next calibration opportunity. Finally, we used general *head related transfer functions* (HRTF) instead of a personalized one to generate binaural audios. Estimating the *personal* HRTF is part of our future work and will of course reduce binaural perception errors.

■ **Power consumption with *Ear-AR***
In the newest firmware update for Airpods Pro, Apple enabled head tracking and spatial audio [8]. Compared with Apple's technique, the only extra energy consumption is computation energy caused by indoor pedestrian dead reckoning. It is generally well-known that IMU dead reckoning is not energy hungry [43]. In light of this, we believe energy consumption would not be a hurdle for *Ear-AR*.

This concludes our system design section; next, we evaluate the performance of *Ear-AR*.

6 EVALUATION

We begin with experiment design, followed by end-to-end system performance and micro benchmarks.

6.1 Experiment Design

■ **Setup**: We evaluate *Ear-AR* in two settings shown in Figure 12 – a large 24m X 10.5m lobby of our engineering building and a small 8m X 6m lab space. These spaces are deliberately chosen to test different aspects of the system. Small settings require lots of turns which challenges the step length and walking direction components of *Ear-AR*. Large areas prompt longer walks, causing errors to accumulate over time. We demonstrate robustness to both settings.

Volunteers wear a Beats [2] Bluetooth headphone taped with a wireless 100Hz IMU (Figure 2(a)). They also carry a phone that records IMU data; all data is wirelessly streamed to a laptop. The laptop runs MATLAB which computes user location and generates a 48kHz binaural audio. This audio is re-synthesized at a rate of 30Hz as the user moves around.

■ **Methodology**: An experiment session begins with a volunteer wearing the headphone+IMU, carrying the smartphone in the pant pocket, and standing at a known starting location. 15 objects are annotated in the lobby, and 9 in the lab (Figure 12) including plants and books on the shelves, a CCTV camera, a refrigerator, a clock, a wall painting, etc. As the volunteer begins walking, she hears a binaural voice that says "find me". She follows the direction of the voice and her task is to identify the source object. If her location estimation or gazing direction drifts too much, the binaural sounds become misleading, and she ends up choosing a wrong object from the surroundings.

We measure "*object identification error* (OIE)" defined as the error between the true object location and the user-identified object location. Once done, the user walks around until a new "find me" voice plays in her headphones and she must now find the location of this new object. Importantly, none of our audio annotations describe the object, hence users must find the objects solely based on IMU + binaural acoustics. Of course, soft calibration is still feasible since every time a user identifies the object (correct or wrong), a recalibration is performed. Thus, our results are a conservative estimate; in real settings, descriptive annotations are likely to improve recalibration leading to better overall results.

6.2 Results: End-to-End Performance

■ **Object Identification**: Figure 13 plots the CDF of *object identification error* (OIE) in meters. The large lobby experiment is composed of 36 sessions/trajectories and the small lab composed of 24. Sessions are defined as a random sequence of 5 to 7 objects that must be identified. Each session is around 8 minutes long and does *not* include any hard recalibration. Our main results are as follows.

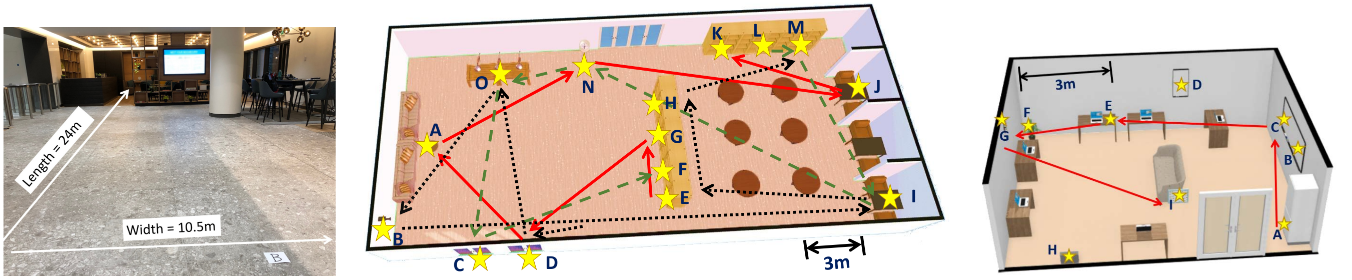

Figure 12: (a) A large lobby of our engineering building; (b) a 3D model of the lobby showing a map of annotated objects set up on the walls and shelves, as well as example walking paths of volunteers; (c) a 3D model of a small lab space in which we also perform AAR experiments.

(1) For large settings, *and with no recalibration*, 63% of the objects were identified correctly (hence OIE=0m). The remaining 37% errors resulted in a median OIE=3.9m, suggesting that errors mostly happened with nearby objects. When soft calibration was applied, the correct cases increased to >90% and the median OIE for the wrong cases reduces to 2.0m.

(2) For small settings, objects were correctly identified in >71% cases, while the remaining 29% cases produced a median OIE of 3.2m. Soft recalibration was deliberately turned off since annotated objects were very densely packed in the lab. Without descriptive annotations, the risks with recalibration are greater than the benefits. In general, *Ear-AR* avoids recalibration when annotations are generic and adjacent objects are within 2m apart, which is the case here. In the real world, such situations should be infrequent.

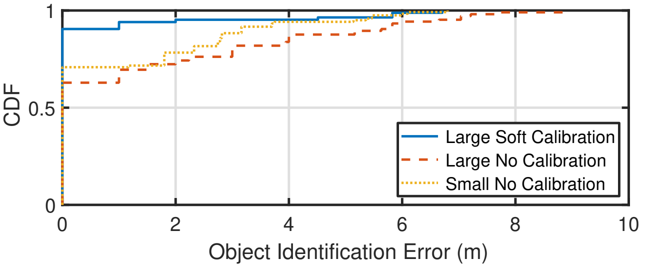

Figure 13: CDF of object identification error (OIE) for (1) large setting with soft calibration, (2) large setting w/o calibration, and (3) small setting w/o calibration.

In sum, the overall performance is promising, particularly given that the above is a conservative evaluation. Real world annotations are expected to be descriptive, implying far greater accuracy with soft recalibration, in turn leading to improved OIE and localization.

Figure 14 zooms into the results from 6 randomly picked sessions, first without recalibration, and then with soft recalibration. Each row represents one session/trajectory. For example, consider the first row in Figure 14(a). The user hears binaural sounds from objects ⟨A, D, E, G, J, K, N⟩, *although not necessarily in that sequence.* In this example, the user correctly identifies the 5 objects, indicated by the check marks. However, for object A, she incorrectly identifies it as object O, which is 5.2 meters away. Similarly, K gets incorrectly classified to L, 1.5m away. Figure 14(b) shows the same results but *with soft recalibration.* Every time the user looks at an object, *Ear-AR* is able to correct user location error. As a result, the error margin is constantly low, leading to fewer incorrect identifications.

6.3 Results: Micro Benchmarks

■ **Walking Direction**: Figure 15 plots the estimated walking direction over time, for two different pre-defined trajectories: a rectangle, and a triangle. We choose the widely-used PCA technique [26] as our baseline. PCA works by first projecting the accelerometer measurements into global reference frame, and then identifying the direction of maximum variance as the walking direction. Figure 15 shows that our method follows the ground truth more closely, essentially because we leverage the physics model to achieve per-step granularity. On the other hand, PCA needs at least a few steps for statistical convergence.

Figure 16 shows the CDF of walking direction errors for 4 users. For this, users walk naturally along a longer and curved path for 2 minutes (without any calibration in between). Errors are smaller for two of the users who exhibit less sideward sway than the others. Overall, the median error is less than 8° for all users, implying reasonable robustness in our walking direction algorithm.

■ **Step Length**: We compare our step length estimation with a baseline algorithm called Weinberg method [72], which essentially assumes that step length can be estimated with an upper-body IMU using the following formula:

$$L = K \times (a_{max} - a_{min})^{\frac{1}{4}} \tag{12}$$

where K is a constant that is trained per user, and a_{max} (a_{min}) is the maximum (minimum) acceleration within a step. For Weinberg method, we ask each user to walk 30 meters to train her K value.

Figure 17 shows the comparison of step length estimation accuracy, between the baseline Weinberg method and *Ear-AR* (without any training). For comparison, we also plot the results of *Ear-AR* after training a scaling factor (similar to the Weinberg method). Users are asked to walk at an even speed. We classify steps into three categories: small (< 0.6m), medium (0.6m - 0.8m), and large (> 0.8m). Ground truth is computed by dividing distance over the total

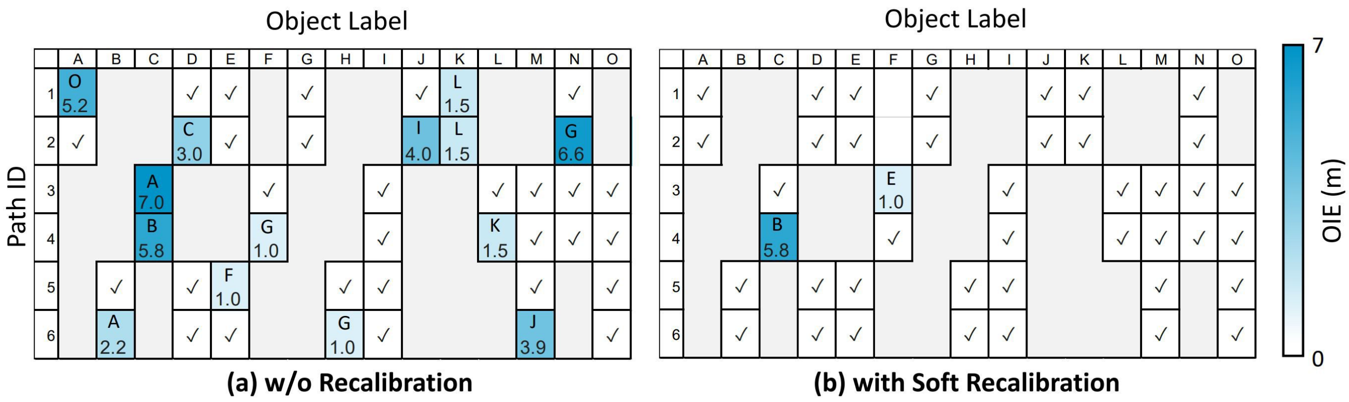

Figure 14: Visualization of the results from 6 example trajectories, (a) w/o and (b) with recalibration opportunities. Each row is one trajectory. A box with a tick represents correct object identification; a box with label and number means this object is incorrectly identified as a different object with this error; a box with no border means it doesn't belong to this trajectory.

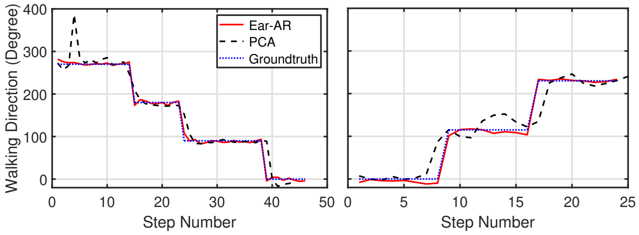

Figure 15: Walking direction over the number of steps for two different trajectories: (a) rectangle, (b) triangle. Comparison shown with past techniques using PCA.

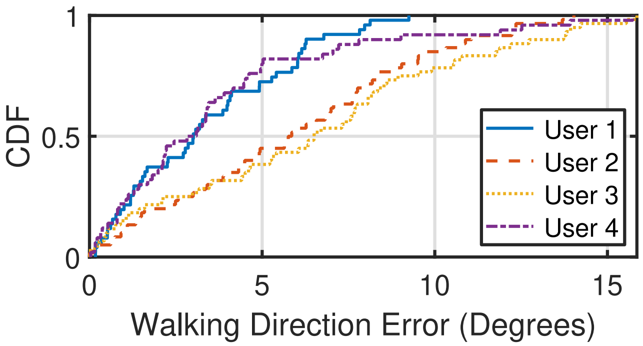

Figure 16: CDF of walking direction errors, across four different users.

number of steps. For small and large steps, even without training, *Ear-AR* performs better than the baseline algorithm by > 4*X*. For medium steps, *Ear-AR* performs worse than baseline, but better if training is also allowed for *Ear-AR*. On average, *Ear-AR* achieves 9.0% step length error without training, and 5.4% error with simple training.

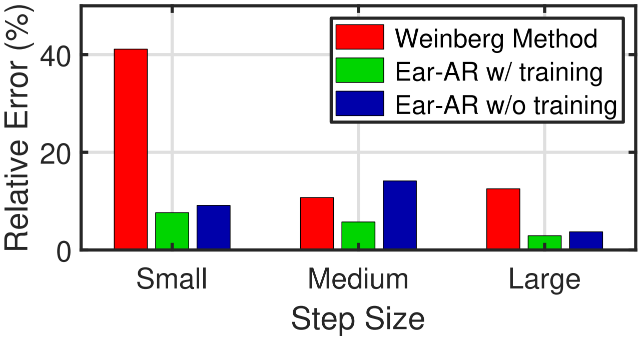

Figure 17: Avg. step length error for varying step size

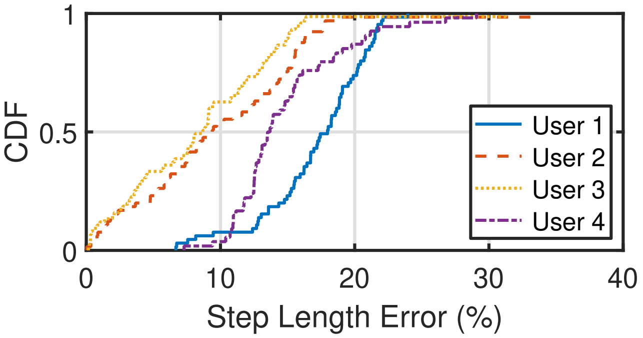

Figure 18: Step length estimation error across different users, for a medium (natural) step size.

Figure 18 further decouples the blue bar in Figure 17 (*Ear-AR* without training, medium steps) into a CDF graph across different users. Overall, the variation of *Ear-AR*'s step length median errors is within 8.1% − 17.4% across users.

■ **IMU-based Tracking (called Dead Reckoning)**:

With walking direction and step length in place, we now evaluate the net dead reckoning error. Figure 19 shows the quantitative results by plotting the dead reckoning errors (in meters) over varying walking distances. We plot the results for 4 different methods: (1) *Ear-AR*, with no recalibration; (2) *Ear-AR*, with soft recalibration; (3) *Ear-AR*, with hard recalibration; and (4) baseline, which is PCA walking direction + fixed (averaged) step length. Even after a user has walked for 150 meters inside the lobby, the average worst case error (averaged across all test sessions) is 7.9m with no calibration, 7.1m with hard recalibration, and 5.25m with soft calibration. This is encouraging, especially when combined with opportunistic (soft and hard) recalibration.

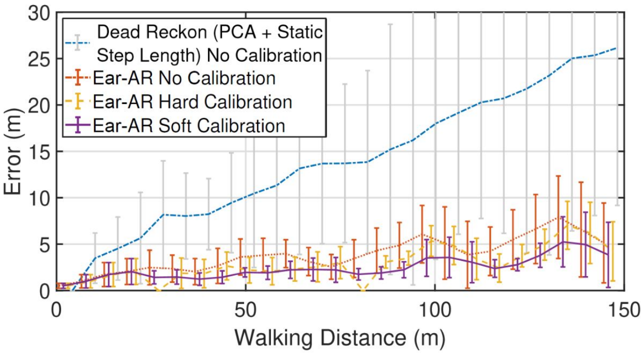

Figure 19: Localization error over variants of *Ear-AR*, plus a comparison with conventional dead reckoning.

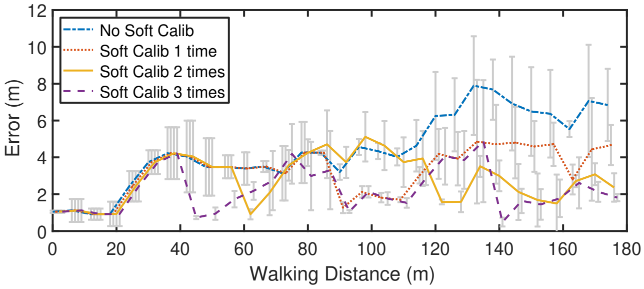

Figure 20: Dead Reckoning Error with Different Soft Calibration Point Density

■ **Soft Calibration Frequency**: Figure 20 shows the effect of soft calibration on dead reckoning error. We plot the error over a distance of 180m of random walking inside our building for (1) no soft calibration, (2) one soft calibration in the middle, (3) two evenly distributed soft calibrations (60 meters apart) and (4) three evenly distributed calibration points (45 meters apart). *Ear-AR*'s average worst case error is 7.9m with no calibration, and 5.1m with 3 calibrations. The average error is obviously less: 4.4m with no calibration, and 3.2m, 2.8m and 2.2m with 1, 2 and 3 soft calibrations, respectively. Evidently, *Ear-AR*'s localization is reasonably robust even without frequent calibrations. With one calibration every 90m, *Ear-AR* should be able to perform well.

■ **Dead Reckoning Stress Test**: Figure 21 shows the dead reckoning performance of *Ear-AR* under challenging environments where normal dead reckoning will falter. These environments include changing walking speed, some side steps, and constant head movement. Our results shows that *Ear-AR*'s dead reckoning error is still less than 12% during a 200m walk. If we have calibration opportunity every 50m, the average error is still below 7.0m. This error is mainly due to step length error – changing walking speed creates additional challenge in estimating step length accurately. However, as shown in Figure 17, *Ear-AR* is already much better than current heuristics because *Ear-AR* builds on concrete geometric relationships between leg movement and step length. Finally, random head rotation will hardly affect *Ear-AR* because once the step length is estimated (and the phone is in the pocket), we need not rely on the ear IMU anymore for dead reckoning.

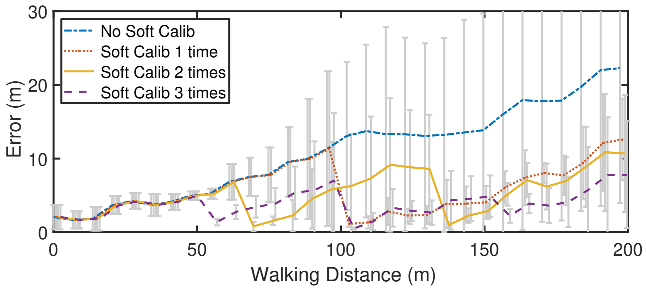

Figure 21: Dead Reckoning Under Stress Test

■ **Object Annotation**: We evaluate *Ear-AR*'s ability to localize objects on the wall (by gazing at it and its vertical projection on the floor). Figure 22 shows the median annotation error with increasing distance between the user and the wall (as depicted in Figure 11). The error bars represent standard deviation. As expected, larger user-to-wall distance causes larger annotation errors, essentially because small angular error (in gazing direction) translates to large annotation displacement on the walls. On average, the annotation error is 15*cm* per 1 meter of distance.

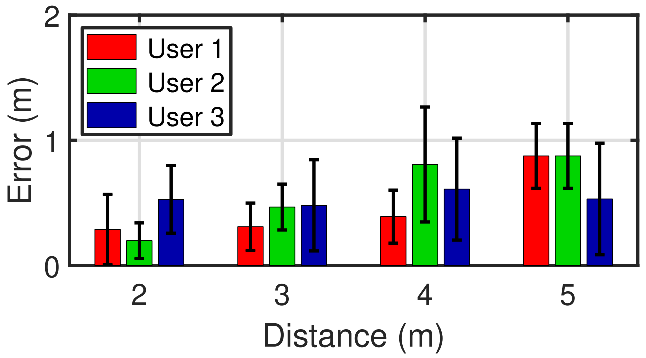

Figure 22: Errors in annotating objects on the wall, when users stand at different distances from the wall.

■ **Gazing Direction**: Finally, Figure 23 shows the performance of gazing direction tracking, when the user walks along a circular trajectory for three times, so that her head orientation slowly rotates from 0° to 1080°. We plot two types of errors: (1) gyroscope integration drift; and (2) gazing direction error, which is the combination of gyroscope drift and the untracked eyeball movement. On average, the gazing direction error is less than 8°, even after three full rounds of rotation.

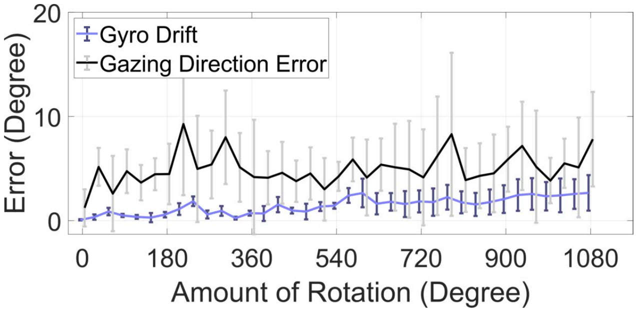

Figure 23: Angular error in gyroscope integration and gazing direction, as the amount of head rotation increases.

7 LIMITATIONS AND DISCUSSIONS

We discuss a few limitations with our current system.

■ **Crowdsourcing Object Annotation:** This paper assumes that object locations are known from an offline database. Ideally, object annotations should be produced seamlessly, e.g., people visiting a new object, looking at it, and recording the annotation. We have designed and evaluated such object localization separately (Fig. 22) but have not used it in the system since soft recalibration would get affected. One improvement is to utilize the average of the crowd's estimates to refine the object's location. This raises questions about system convergence, hence opens to future research.

■ **Floor Semantics:** *Ear-AR* needs the rough floor plan (mainly wall locations) to know if an annotated object is in the user's line of sight. Otherwise, an object may be in another room and *Ear-AR* would believe the user is gazing at it. Our evaluation did not require such floorplans since all objects were in a single lobby or lab. In a building-wide deployment like airports, libraries, malls, such floor semantics are necessary.

■ **HCI Factors in *Ear-AR*:** A real-world deployment would need to regulate how, when, and which annotations are played, to minimize disturbance or information overload. Such policies need to be designed with considerations of human factors, a topic unaddressed in this paper.

■ **Using Earphones Alone for *Ear-AR*:** Future earphones are being envisioned as stand-alone devices that do not rely on the smart phone. From the perspective of *Ear-AR*, we believe this is viable, however, this means the walk estimation technique would no longer have the leg rotation information. Reliably inferring walking direction from earphone IMUs alone is a critical but challenging problem. This remains an open question for follow-up research.

8 RELATED WORK

Earables and Acoustic AR: Bose AR [20, 21] and Microsoft Soundscape [55, 66] are the closest to our work. Both offer AAR to users via earphones (including binaural sounds) but are entirely for outdoor use (via GPS). *Ear-AR* can be viewed as an enabler of indoor experiences for Bose and Microsoft's applications. Additional ear-worn devices are on the rise from both industry [13, 28] and academic projects [25, 59], however, none offer the location context necessary for AR. *Ear-AR*'s indoor localization technique, coupled tightly with binaural acoustics, is well suited for AR applications.

Motion Tracking and Localization: Motion tracking and localization are classical research questions in the mobile and wireless community [10, 18, 30, 39–41, 45, 49, 51–54, 58, 61, 68, 70, 75–78, 80]. Past work on IMU based tracking have proposed various creative ideas [24, 37, 42, 67, 71, 74], but none of them solve the entire dead reckoning problem. This is essentially because phone or watch IMUs are not in a "good position" to accurately estimate human walks, even with advancements in algorithm design. IMU-embedded earphones [3, 4, 6] are bringing new opportunities as leveraged by *Ear-AR*. Recent researches have looked at specific aspects of the problem, such as counting steps [4, 6, 62], measuring head rotation [32, 73], detecting walking stages [33], or fusion with other sensors [16]. In fact, STEAR [62] takes advantage of cleaner IMU signal on the ears but only for step counting. In contrast, *Ear-AR* fully utilizes the unique opportunities from earphone IMUs to fill in the missing pieces in pedestrian dead reckoning.

Binaural Acoustics: The fact that human brain is capable of resolving the direction of the incoming sound is well known [57]. Past works have utilized the binaural effect for different purposes, including sound recording and reproduction [1, 22, 29], entertainment [7, 17, 46, 48], and localization [36, 56, 60]. We instead leverage binaural sounds to recalibrate IMU-based motion tracking errors, a novel usage of the brain's binaural capability.

9 CONCLUSION

The popularity of sensor-embedded earphones is ushering new opportunities. This paper demonstrates how earphone IMUs capture "naturally filtered" signals related to the human walk, improving over state-of-the-art solutions in pedestrian dead reckoning (PDR) and localization. The IMUs also capture head movements, which when combined with the human ability to sense 3D sounds, enables new kinds of applications. This paper demonstrates one such application in acoustic augmented reality (AAR), pointing to a future where voice assistants like Siri would be context-aware, both in terms of the user's interest and the surrounding environment. *Ear-AR* is an early step in this direction.

10 ACKNOWLEDGMENTS

We sincerely thank the anonymous shepherd and reviewers for their insightful comments and suggestions. We are also grateful to NSF (award numbers: 1918531, 1910933, 1909568, and 1719337), NIH (award number: 1R34DA050262-01), Google, and Samsung for partially funding this research.

REFERENCES

[1] 2019. 3Dio: Professional Binaural Microphones. Retrieved Sept 18, 2019 from https://3diosound.com/

[2] 2019. Beats Solo3 Wireless – Beats by Dre. Retrieved Sep 13, 2019 from https://www.beatsbydre.com/uk/headphones/solo3-wireless

[3] 2019. Bragi Earbuds. Retrieved Sept 18, 2019 from https://bragi.com/

[4] 2019. Jabra Wireless Workout Headphones. Retrieved Sept 18, 2019 from https://www.jabra.com/sports-headphones/jabra-sport-coach-wireless#/

[5] 2019. Listen HRTF Database. Retrieved Sep 19, 2019 from http://recherche.ircam.fr/equipes/salles/listen/

[6] 2019. Samsung Gear IconX. Retrieved Sept 18, 2019 from https://www.samsung.com/global/galaxy/gear-iconx/

[7] 2019. Sennheiser's short film shows the power of binaural audio. Retrieved Sept 18, 2019 from https://thenextweb.com/plugged/2018/09/17/sennheisers-short-film-shows-the-power-of-binaural-audio/

[8] 2020. Airpods Pro. Retrieved Jul 21, 2020 from https://www.apple.com/airpods-pro/

[9] 2020. Apple spatial audio: what is it? How do you get it? Retrieved Jul 21, 2020 from https://www.whathifi.com/advice/what-is-apple-spatial-audio

[10] Fadel Adib, Zach Kabelac, Dina Katabi, and Robert C Miller. 2014. 3D tracking via body radio reflections. In *11th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 14)*. 317–329.

[11] Muhammad Haris Afzal, Valérie Renaudin, and Gérard Lachapelle. 2010. Assessment of indoor magnetic field anomalies using multiple magnetometers. In *ION GNSS*, Vol. 10. 21–24.

[12] Hyun-Sung An, Gregory C Jones, Seoung-Ki Kang, Gregory J Welk, and Jung-Min Lee. 2017. How valid are wearable physical activity trackers for measuring steps? *European journal of sport science* 17, 3 (2017), 360–368.

[13] Apple. 2020. Apple Airpods. https://www.apple.com/airpods/ [Online; accessed 20-March-2020].

[14] C Ascher, C Kessler, M Wankerl, and GF Trommer. 2010. Dual IMU indoor navigation with particle filter based map-matching on a smartphone. In *2010 International Conference on Indoor Positioning and Indoor Navigation*. IEEE, 1–5.

[15] Haitao Bao and Wai-Choong Wong. 2013. Improved PCA based step direction estimation for dead-reckoning localization. In *2013 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*. IEEE, 325–331.

[16] Stéphane Beauregard. 2006. A helmet-mounted pedestrian dead reckoning system. In *3rd International Forum on Applied Wearable Computing 2006*. VDE, 1–11.

[17] Durand R Begault and Leonard J Trejo. 2000. 3-D sound for virtual reality and multimedia. (2000).

[18] Alastair R Beresford and Frank Stajano. 2003. Location privacy in pervasive computing. *IEEE Pervasive computing* 1 (2003), 46–55.

[19] Johann Borenstein, Lauro Ojeda, and Surat Kwanmuang. 2009. Heuristic reduction of gyro drift in IMU-based personnel tracking systems. In *Optics and Photonics in Global Homeland Security V and Biometric Technology for Human Identification VI*, Vol. 7306. International Society for Optics and Photonics, 73061H.

[20] BOSE. 2020. Bose Developer Portal | NaviGuide Around Your World: Heads Up, Hands Free. https://developer.bose.com/bose-ar/get-inspired/naviguide-around-your-world-heads-hands-free [Online; accessed 16-March-2020].

[21] BOSE. 2020. Bose Developer Portal | Tap Into Otocast, Your Friendly AR-enabled Tour Guide. https://developer.bose.com/bose-ar/get-inspired/tap-otocast-your-friendly-ar-enabled-tour-guide [Online; accessed 17-March-2020].

[22] C Phillip Brown and Richard O Duda. 1998. A structural model for binaural sound synthesis. *IEEE transactions on speech and audio processing* 6, 5 (1998), 476–488.

[23] Christophe Combettes and Valerie Renaudin. 2017. Walking direction estimation based on statistical modeling of human gait features with handheld MIMU. *IEEE/ASME Transactions on Mechatronics* 22, 6 (2017), 2502–2511.

[24] Alejandro Correa, Estefania Munoz Diaz, Dina Bousdar Ahmed, Antoni Morell, and Jose Lopez Vicario. 2016. Advanced pedestrian positioning system to smartphones and smartwatches. *Sensors* 16, 11 (2016), 1903.

[25] Daniel de Godoy, Bashima Islam, Stephen Xia, Md Tamzeed Islam, Rishikanth Chandrasekaran, Yen-Chun Chen, Shahriar Nirjon, Peter R Kinget, and Xiaofan Jiang. 2018. Paws: A wearable acoustic system for pedestrian safety. In *2018 IEEE/ACM Third International Conference on Internet-of-Things Design and Implementation (IoTDI)*. IEEE, 237–248.

[26] Zhi-An Deng, Guofeng Wang, Ying Hu, and Di Wu. 2015. Heading estimation for indoor pedestrian navigation using a smartphone in the pocket. *Sensors* 15, 9 (2015), 21518–21536.

[27] Richard O Duda. 1993. Modeling head related transfer functions. In *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*. IEEE, 996–1000.

[28] Google. 2020. Pixel Buds. https://store.google.com/product/pixel_buds [Online; accessed 20-March-2020].

[29] Dorte Hammershøi and Henrik Møller. 2002. Methods for binaural recording and reproduction. *Acta Acustica united with Acustica* 88, 3 (2002), 303–311.

[30] Fabian Höflinger, Rui Zhang, and Leonhard M Reindl. 2012. Indoor-localization system using a micro-inertial measurement unit (imu). In *2012 European Frequency and Time Forum*. IEEE, 443–447.

[31] JC Hung, JR Thacher, and HV White. 1989. Calibration of accelerometer triad of an IMU with drifting Z-accelerometer bias. In *Proceedings of the IEEE National Aerospace and Electronics Conference*. IEEE, 153–158.

[32] Tong-Hun Hwang, Julia Reh, Alfred Effenberg, and Holger Blume. 2016. Real-time gait event detection using a single head-worn inertial measurement unit. In *2016 IEEE 6th International Conference on Consumer Electronics-Berlin (ICCE-Berlin)*. IEEE, 28–32.

[33] Tong-Hun Hwang, Julia Reh, Alfred O Effenberg, and Holger Blume. 2018. Real-time gait analysis using a single head-worn inertial measurement unit. *IEEE Transactions on Consumer Electronics* 64, 2 (2018), 240–248.

[34] Ian Dickson. 2020. Bose and HERE – creating the future of AR. https://360.here.com/bose-and-here-creating-the-future-of-ar [Online; accessed 24-March-2020].

[35] Ian Dickson. 2020. Bose and HERE – creating the future of AR. https://developer.bose.com/content/here-maps-and-bose-ar-integration [Online; accessed 24-March-2020].

[36] Lloyd A Jeffress. 1948. A place theory of sound localization. *Journal of comparative and physiological psychology* 41, 1 (1948), 35.

[37] Antonio R Jimenez, Fernando Seco, Carlos Prieto, and Jorge Guevara. 2009. A comparison of pedestrian dead-reckoning algorithms using a low-cost MEMS IMU. In *2009 IEEE International Symposium on Intelligent Signal Processing*. IEEE, 37–42.

[38] Antonio Ramón Jiménez, Fernando Seco, José Carlos Prieto, and Jorge Guevara. 2010. Indoor pedestrian navigation using an INS/EKF framework for yaw drift reduction and a foot-mounted IMU. In *2010 7th Workshop on Positioning, Navigation and Communication*. IEEE, 135–143.

[39] Haojian Jin, Jingxian Wang, Zhijian Yang, Swarun Kumar, and Jason Hong. 2018. Rf-wear: Towards wearable everyday skeleton tracking using passive rfids. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. 369–372.

[40] Haojian Jin, Jingxian Wang, Zhijian Yang, Swarun Kumar, and Jason Hong. 2018. Wish: Towards a wireless shape-aware world using passive rfids. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. 428–441.

[41] Haojian Jin, Zhijian Yang, Swarun Kumar, and Jason I Hong. 2018. Towards wearable everyday body-frame tracking using passive rfids. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–23.

[42] Wonho Kang and Youngnam Han. 2014. SmartPDR: Smartphone-based pedestrian dead reckoning for indoor localization. *IEEE Sensors journal* 15, 5 (2014), 2906–2916.

[43] Kleomenis Katevas, Hamed Haddadi, and Laurissa Tokarchuk. 2016. Sensingkit: Evaluating the sensor power consumption in ios devices. In *2016 12th International Conference on Intelligent Environments (IE)*. IEEE, 222–225.

[44] Maan Khedr and Nasser El-Sheimy. 2017. A smartphone step counter using IMU and magnetometer for navigation and health monitoring applications. *Sensors* 17, 11 (2017), 2573.

[45] Manikanta Kotaru, Kiran Joshi, Dinesh Bharadia, and Sachin Katti. 2015. Spotfi: Decimeter level localization using wifi. In *ACM SIGCOMM computer communication review*, Vol. 45. ACM, 269–282.

[46] Zeqi Lai, Y Charlie Hu, Yong Cui, Linhui Sun, Ningwei Dai, and Hung-Sheng Lee. 2019. Furion: Engineering high-quality immersive virtual reality on today's mobile devices. *IEEE Transactions on Mobile Computing* (2019).

[47] Teesid Leelasawassuk, Dima Damen, and Walterio W Mayol-Cuevas. 2015. Estimating visual attention from a head mounted IMU. In *Proceedings of the 2015 ACM International Symposium on Wearable Computers*. 147–150.

[48] Tobias Lentz, Dirk Schröder, Michael Vorländer, and Ingo Assenmacher. 2007. Virtual reality system with integrated sound field simulation and reproduction. *EURASIP journal on advances in signal processing* 2007, 1 (2007), 070540.

[49] Hui Liu, Houshang Darabi, Pat Banerjee, and Jing Liu. 2007. Survey of wireless indoor positioning techniques and systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 37, 6 (2007), 1067–1080.

[50] Yu Liu, Yanping Chen, Lili Shi, Zengshan Tian, Mu Zhou, and Lingxia Li. 2015. Accelerometer based joint step detection and adaptive step length estimation algorithm using handheld devices. *Journal of Communications* 10, 7 (2015), 520–525.

[51] Yunhao Liu, Zheng Yang, Xiaoping Wang, and Lirong Jian. 2010. Location, localization, and localizability. *Journal of Computer Science and Technology* 25, 2 (2010), 274–297.

[52] Kieran Mansley, Alastair R Beresford, and David Scott. 2004. The carrot approach: encouraging use of location systems. In *International Conference on Ubiquitous Computing*. Springer, 366–383.

[53] Wenguang Mao, Mei Wang, and Lili Qiu. 2018. Aim: acoustic imaging on a mobile. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 468–481.

[54] Wenguang Mao, Zaiwei Zhang, Lili Qiu, Jian He, Yuchen Cui, and Sangki Yun. 2017. Indoor follow me drone. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 345–358.

[55] Microsoft. 2020. Microsoft Soundscape: A Map Delivered in 3D Sound - Microsoft Research. https://www.microsoft.com/en-us/research/video/microsoft-soundscape-map-delivered-3d-sound/ [Online; accessed 17-March-2020].

[56] Pauli Minnaar, S Krarup Olesen, Flemming Christensen, and Henrik Møller. 2001. Localization with binaural recordings from artificial and human heads. *Journal of the Audio Engineering Society* 49, 5 (2001), 323–336.

[57] Henrik Møller. 1992. Fundamentals of binaural technology. *Applied acoustics* 36, 3-4 (1992), 171–218.

[58] Rajalakshmi Nandakumar, Vikram Iyer, and Shyamnath Gollakota. 2018. 3D Localization for Sub-Centimeter Sized Devices. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*. ACM, 108–119.

[59] Anh Nguyen, Raghda Alqurashi, Zohreh Raghebi, Farnoush Banaei-Kashani, Ann C Halbower, and Tam Vu. 2016. A lightweight and inexpensive in-ear sensing system for automatic whole-night sleep stage monitoring. In *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM*. ACM, 230–244.

[60] David R Perrott and AD Musicant. 1977. Minimum auditory movement angle: Binaural localization of moving sound sources. *The Journal of the Acoustical Society of America* 62, 6 (1977), 1463–1466.

[61] Swadhin Pradhan, Ghufran Baig, Wenguang Mao, Lili Qiu, Guohai Chen, and Bo Yang. 2018. Smartphone-based Acoustic Indoor Space Mapping. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 75.

[62] Jay Prakash, Zhijian Yang, Yu-Lin Wei, and Romit Roy Choudhury. 2019. STEAR: Robust Step Counting from Earables. In *Proceedings of the 1st International Workshop on Earable Computing*. 36–41.

[63] Sujatha Rajagopal. 2008. Personal dead reckoning system with shoe mounted inertial sensors. *Master's Degree Project, Stockholm, Sweden* (2008).

[64] Eike Rehder, Horst Kloeden, and Christoph Stiller. 2014. Head detection and orientation estimation for pedestrian safety. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2292–2297.

[65] Zhi-Hong Ren, Gui-Lin Wang, and David Ho. 2006. Mobile phone with pedometer. US Patent App. 11/136,265.

[66] Rico Malvar. 2020. A new Soundscape experience with Bose Frames - Microsoft Accessibility Blog. https://blogs.microsoft.com/accessibility/soundscape-boseframes/ [Online; accessed 17-March-2020].

[67] Nirupam Roy, He Wang, and Romit Roy Choudhury. 2014. I am a smartphone and I can tell my user's walking direction. In *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*. ACM, 329–342.

[68] Sheng Shen, Daguan Chen, Yu-Lin Wei, Zhijian Yang, and Romit Roy Choudhury. 2020. Voice localization using nearby wall reflections. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–14.

[69] SH Shin, CG Park, JW Kim, HS Hong, and JM Lee. 2007. Adaptive step length estimation algorithm using low-cost MEMS inertial sensors. In *2007 ieee sensors applications symposium*. IEEE, 1–5.

[70] Anran Wang and Shyamnath Gollakota. 2019. MilliSonic: Pushing the Limits of Acoustic Motion Tracking. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 18.

[71] He Wang, Souvik Sen, Ahmed Elgohary, Moustafa Farid, Moustafa Youssef, and Romit Roy Choudhury. 2012. No Need to War-drive: Unsupervised Indoor Localization. In *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services (MobiSys '12)*. ACM, New York, NY, USA, 197–210. https://doi.org/10.1145/2307636.2307655

[72] Harvey Weinberg. 2002. Using the ADXL202 in pedometer and personal navigation applications. *Analog Devices AN-602 application note* 2, 2 (2002), 1–6.

[73] Jens Windau and Laurent Itti. 2016. Walking compass with head-mounted IMU sensor. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 5542–5547.

[74] Zhuoling Xiao, Hongkai Wen, Andrew Markham, and Niki Trigoni. 2014. Robust pedestrian dead reckoning (R-PDR) for arbitrary mobile device placement. In *2014 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. IEEE, 187–196.

[75] Yaxiong Xie, Jie Xiong, Mo Li, and Kyle Jamieson. 2019. mD-Track: Leveraging Multi-Dimensionality for Passive Indoor Wi-Fi Tracking. In *The 25th Annual International Conference on Mobile Computing and Networking*. ACM, 1–16.

[76] Jie Xiong and Kyle Jamieson. 2013. Arraytrack: A fine-grained indoor location system. In *Presented as part of the 10th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 13)*. 71–84.

[77] Zheng Yang, Chenshu Wu, and Yunhao Liu. 2012. Locating in fingerprint space: wireless indoor localization with little human intervention. In *Proceedings of the 18th annual international conference on Mobile computing and networking*. ACM, 269–280.

[78] Ali Yassin, Youssef Nasser, Mariette Awad, Ahmed Al-Dubai, Ran Liu, Chau Yuen, Ronald Raulefs, and Elias Aboutanios. 2016. Recent advances in indoor localization: A survey on theoretical approaches and applications. *IEEE Communications Surveys & Tutorials* 19, 2 (2016), 1327–1346.

[79] Ji Zhang and Sanjiv Singh. 2017. Low-drift and real-time lidar odometry and mapping. *Autonomous Robots* 41, 2 (2017), 401–416.

[80] Yanzi Zhu, Zhujun Xiao, Yuxin Chen, Zhijing Li, Max Liu, Ben Y. Zhao, and Haitao Zheng. 2018. Et Tu Alexa? When Commodity WiFi Devices Turn into Adversarial Motion Sensors. *arXiv e-prints*, Article arXiv:1810.10109 (Oct 2018), arXiv:1810.10109 pages. arXiv:cs.CR/1810.10109