Room-scale Location Trace Tracking via Continuous Acoustic Waves

JIE LIAN, University of Louisiana at Lafayette, Lafayette, USA
XU YUAN, University of Delaware, Newark, USA
JIADONG LOU, University of Delaware, Newark, USA
LI CHEN, University of Louisiana at Lafayette, Lafayette, USA
HAO WANG, Louisiana State University, Baton Rouge, USA
NIANFENG TZENG, University of Louisiana at Lafayette, Lafayette, USA

The increasing prevalence of smart devices spurs the development of emerging indoor localization technologies for supporting diverse personalized applications at home. Given marked drawbacks of popular chirp signal-based approaches, we aim at developing a novel device-free localization system via the continuous wave of the inaudible frequency. To achieve this goal, solutions are developed for fine-grained analyses, able to precisely locate moving human traces in the room-scale environment. In particular, a smart speaker is controlled to emit continuous waves at inaudible $20kHz$, with a co-located microphone array to record their Doppler reflections for localization. We first develop solutions to remove potential noises and then propose a novel idea by slicing signals into a set of narrowband signals, each of which is likely to include at most one body segment's reflection. Different from previous studies, which take original signals themselves as the baseband, our solutions employ the Doppler frequency of a narrowband signal to estimate the velocity first and apply it to get the accurate baseband frequency, which permits a precise phase measurement after I-Q (i.e., in-phase and quadrature) decomposition. A signal model is then developed, able to formulate the phase with body segment's velocity, range, and angle. We next develop novel solutions to estimate the motion state in each narrowband signal, cluster the motion states for different body segments corresponding to the same person, and locate the moving traces while mitigating multi-path effects. Our system is implemented with commodity devices in room environments for performance evaluation. The experimental results exhibit that our system can conduct effective localization for up to three persons in a room, with the average errors of $7.49cm$ for a single person, with $24.06cm$ for two persons, with $51.15cm$ for three persons.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing;**

Additional Key Words and Phrases: ultrasonic sensing, smart home, localization

ACM Reference Format:
Jie Lian, Xu Yuan, Jiadong Lou, Li Chen, Hao Wang, and Nianfeng Tzeng. 2024. Room-scale Location Trace Tracking via Continuous Acoustic Waves. *ACM Trans. Sensor Netw.* 20, 3, Article 61 (April 2024), 23 pages. https://doi.org/10.1145/3649136

This work was supported in part by the National Science Foundation (NSF) grants 2348452, and 2315613, 2019511, 2153502, 2315612, and 2327480.

Authors' addresses: J. Lian, L. Chen, and N. Tzeng, University of Louisiana at Lafayette, Lafayette, LA 70503; e-mails: fuc192012@gmail.com, li.chen@louisiana.edu, nianfeng.tzeng@louisiana.edu; X. Yuan (Corresponding author) and J. Lou, University of Delaware, Newark, DE 19716; e-mails: xyuan@udel.edu, loujd@udel.edu; H. Wang, Louisiana State University, Baton Rouge, LA 7080; e-mail: haowang@lsu.edu.

ACM 1550-4859/2024/04-ART61
https://doi.org/10.1145/3649136

1 INTRODUCTION

The indoor localization has spurred extensive personalized applications. The prior report has revealed that a person may spend almost 88.9% of the day indoors [20]. Recently, [8] has predicted that the market value of indoor positioning and indoor navigation is expected to exceed 23.6 billion dollars in 2023, suggesting the tremendous demand for reliable indoor localization technologies. Among the emerging technologies, the device-free indoor localization without requiring a user to carry any device for sensing, is most appealing in many scenarios, prompting diverse applications such as smart homes, healthcare services, fitness tracking, and so on. The acoustic-based device-free sensing approach is promising and affordable, given the increasing prevalence of in-house smart devices. Its low frequency range sensing is readily achieved by the **commercial off-the-shelf** (**COTS**) devices. Various human-computer interaction services based on the acoustic signal have been exploited, such as activity recognition [14, 15, 37], lip-reading [27, 45, 46], respiration and heartbeat detection [35, 44], localization [12, 13, 19], and so on. However, it remains challenging to develop a reliable system for room-scale location trace in real home environments.

Although previous studies in device-free acoustic sensing have achieved the millimeter level accuracy [22, 36] and extended the sensing range to $4.5m$ [19], they mainly strive to explore inaudible chirp signals, which typically entice two fundamental issues. First, chirp signals require to have short durations and large bandwidth for achieving satisfactory performance. However, when the COTS devices are controlled to generate a short chirp signal of wide bandwidth under the inaudible frequency, the electronic burst can occasionally cause an audible "Beep" sound on the speaker, considered to be rather annoying to humans. In addition, due to the limited space of an indoor environment, the chirp signal will be reflected multiple times before it completely attenuates, resulting in multiple echoes. As these echoes are amenable to environmental changes, any human motion may alter the echoes significantly. Therefore, the echoes of chirps can be different, difficult to be eliminated completely, thus degrading the signal quality and incurring the range estimation error. In sharp contrast to chirp signals, the continuous waves emitted by COTS devices do not generate the perceptible sounds to the human as they operate in the inaudible frequency. In addition, the frequencies of reflections from stationary objects in an environment are relatively stable, causing little interference to target signals reflected from a moving person, making it possible to eliminate those static reflection frequencies.

In this article, we explore the continuous wave as the sensing signal and leverage the phase change of Doppler signals for tracking the motions of moving targets. Although device-free tracking based on the phase measurement has been pursued previously [25, 36, 42], the sensing ranges achieved therein are considerably short (only up to 1 meter), unsuitable for location trace tracking in a room. This is due to the fact that they ignored the Doppler effect of signal and directly extracted the phase from the reflected signals by referring to the original signal as the baseline. Albeit simple, such a method can be effective just for a short distance. However, in a room-scale environment, the original signal will be transmitted over multiple paths to the microphone. The multiple path signals can be viewed as several time-delay versions of the original signal, and after those signals are aggregated, the original signal experiences phase distortion. Thus applying the original signal as the baseline tends to yield wrong phase estimation in a room-scale environment. Instead, we examine the phase change of reflected Doppler signals, permitting to acquire the proper baseline signal from the body reflection signals. Since this way makes it possible to obtain very accurate phase estimation over a relatively large range, the sensing range can thus be significantly extended for applications to room-scale environments.

To this end, we strive to develop a novel device-free indoor localization system by relying on continuous waves emitted from the speaker's built-in smart devices and performing the fine-grained

phase and frequency analyses of Doppler signals received at the built-in microphone array to implement precise indoor location trace tracking. Specifically, we control a speaker to generate the continuous wave ultrasound at $20kHz$ and leverage a co-located microphone array to record reflected signals for analysis. We first perform the **Short-time Fourier transform** (**STFT**) to get the spectrogram of reflected Doppler signals and then apply a set of interference cancellation techniques to remove the interference/noise caused by both surrounding objects and system defects. Considering different body segments of a moving person may generate Doppler signals at different frequencies, we propose a new solution by slicing signals into a set of narrowband signals, with each narrowband likely to include the reflections from only one body segment. We next perform the fine-grained analysis on each narrowband signal by a series of developed solutions. First, we estimate the velocity via analyzing the original signal from the speaker and the received signal at the microphone array, followed by utilizing such a velocity to estimate the baseband frequency of the narrowband signal. The I-Q (i.e., in-phase and quadrature) decomposition is then applied to accurately measure its phase. After that, a signal model is proposed to formulate the phase correlated to the range, angle, and velocity. The range and angle of each narrowband signal are then estimated by solving an optimization problem, aiming at getting each body segment's motion state. Next, a solution based on K-mean clustering is applied to cluster the motion states associated with a given person. Finally, we apply the L1-norm (i.e., the sum of the magnitudes of the vectors) to refine our estimation and remove the interference from multi-path effects, able to get the number of moving persons and their associated location traces accurately.

Our contributions are summarized as follows:

- We develop a new device-free localization system, which leverages the continuous wave emitted from the smart speaker as the sensing signal for precise localization. As far as we know, this study is the very first to demonstrate the possibility of tracking multiple individuals using continuous waves. To achieve our goal, we develop solutions to perform the fine-grained analysis of Doppler signals by extracting their frequency and phase details for localization. Unlike the chirp-based localization counterparts, our system only relies on continuous waves at $20kHz$, which utilize a small bandwidth and are both imperceptible to humans and robust to environmental factors.
- Novel solutions to enable the phase measurement of Doppler signals are developed by slicing the Doppler signals into narrowband signals to have each narrowband contain the reflection of a single body segment for the fine-grained analysis of human body patterns. So, the narrowband signals can be viewed as single tone signals, enabling to estimate the phase of each narrowband signal determined by estimating its baseband signal. We are the first to extract the phase of a Doppler signal by dividing it into narrowband signals, which can facilitate the fine-grained analysis for multiple-person location tracing. A new signal model is then proposed for each narrowband signal to fuse the target velocity, range, and angle, enabling the estimation of the motion states of different body segments. Motion state clustering and location trace solutions are also proposed to precisely group together reflections from different body segments of the same person and to further eliminate the multi-path effect.
- We implement our system with COTS devices and conduct extensive experiments for performance evaluation. The experimental results exhibit that our system: (1) can achieve the precise location tracking for a single person, with an average error of $7.49cm$; (2) can support the multi-target tracking, having the average errors of $24.06cm$ and of $51.15cm$, respectively, for two and three persons; (3) is robust to various environmental factors; (4) can be extended for locating a stationary person with a swing arm.

2 RELATED WORK

Acoustic signal applications have gained increasing interest in the research community. Among them, acoustic-based tracking has become prevalent in recent years, by taking advantage of widely available speakers and microphones built in commodity devices. This section discusses the current state-of-the-arts in acoustic localization and differentiates them from our work.

2.1 Acoustic-Based Tracking

Some research efforts [3, 7, 17, 29, 34, 41, 41, 43] entail to develop the device-based tracking by requiring a user to carry the device for localization. This line of solutions indirectly tracks human motions by sensing the movement of carried devices. But, they are inconvenient and impractical in many scenarios, since a person may forget or be reluctant to carry the device. In contrast, our system belongs to device-free tracking, freeing a user from carrying any device to make it more attractive.

Previous device-free tracking mainly relies on controlling the speaker to transmit chirp signals (i.e., OFDM, FMCW, **Zadoff-Chu** (**ZC**) sequence) and analyzing reflected signals for tracking. For example, [22, 23] have exploited the correlation of OFDM signals for measuring the **time-of-flight** (**ToF**) for finger and activity tracking. [12, 13, 49] leveraged the FMCW signal and developed a series of signal processing techniques for motion tracking in the room-scale environment and indoor floor plan mapping. [21] leveraged the ZC sequence to infer user's hand patterns when tapping the PIN code. In contrast, our system achieves target localization in approximately 0.63 seconds. Also, they are using FMCW signals, and the chirp signal-based solutions typically require large bandwidth for accurate tracking (i.e., 16khz to 20khz). Although they work on the inaudible ultrasound range, some people may still perceive such wideband ultrasound disturbingly, as a result of frequency changes. Another potential drawback is that the audible "Beep" sound may be produced, annoying human life. Differently, our system relies on emitting the continuous wave for sensing, which can substantially avoid such drawbacks.

Some earlier methods [18, 19] have demonstrated the viability of neural networks in aiding localization. For example, [19] utilized both the 2D MUSIC algorithm and beamforming to extend the sensing range, coupled with a **recurrent neural network** (**RNN**) to determine target locations. [18] directly inputted data into a **deep neural network** (**DNN**) for finger tracking. One noteworthy distinction in our approach is its ability to operate directly without collecting additional data and labeling processes while achieving room-scale tracking. In addition, our approach takes into account practical considerations, such as the limited hardware and computational constraints of **commercial off-the-shelf** (**COTS**) devices, which might raise challenges for the applicability of existing methods. Although taking longer than earlier methods (i.e., 0.63 sec versus some 0.05 sec in [18, 19]) for localization, our approach nonetheless is still considered to exhibit near real-time localization.

Our research is closely related to [36], where demonstrated the feasibility of tracking the hand movement for the first time by utilizing the phase information of continuous waves. That work analyzed the signal phase by acquiring the difference between the baseband signal and the reflected signal via I-Q decomposition [36]. However, it considered the frequency of baseband signals to be the same as that of original signals, potentially resulting in excessive interference between them. Since the acoustic signal experiences fast attenuation with the distance, the reflection signals over a long range become rather weak and thus are hard to be separated from the strong baseband signal, rendering it suitable only for limited-range sensing. In contrast, our work leverages the phase of Doppler signals for analysis, which has a large difference from that of the baseband signal to yield clean phase information after I-Q decomposition. The tracking range can then be

substantially extended, applicable to room-scale environments. Note that the phase of chirp signals was analyzed in certain prior studies for motion tracking. For example, [42] relied on the phase change of OFDM signals to track fine-grained chest movements for respiration detection. [47] leveraged the phase of FMCW signals to infer the finger position, whereas [25] tracked the finger movements by combining the ZC sequence phase and amplitude information. However, it's worth noting that these systems are primarily designed for finger tracking and may not be suitable for room-scale localization. Meanwhile, they underperform our approach, which directly measures the phase of continuous waves over a set of narrowband signals that incur low noise for sound phase estimation.

2.2 RF-Based Localization

Next, we briefly review RF-based solutions for sensing and tracking target movement, contrasting them with our approach in the acoustic domain. RF-based solutions can be summarized into three categories: Radar-based sensing, RFID-based sensing, and Wi-Fi based sensing.

Many efforts on radar-based sensing [1, 6, 10, 11, 38, 38, 48, 50, 50] have been conducted for single/multi-target tracking. They all call for large bandwidth and the antenna array(s) to acquire signals, requiring specialized hardware to transmit and receive signals to incur considerable extra costs. RFID-based solutions [9, 16, 24, 30–32] for human motion tracking were also considered, by leveraging the RFID tags and readers for localization. However, they typically require users to wear the RFID tags for localization, deemed inconvenient and cumbersome. Although some device-free tracking solutions [5, 40] have been proposed, they experience limited sensing ranges, usually in tens of centimeters to make them unsuitable for the room-scale environment. Location trace tracking techniques based on the Wi-Fi signals have been widely studied [1, 2, 4, 10, 26, 33, 39]. However, they utilize large bandwidth to achieve satisfactory localization accuracy, occupying the communication channels at 2.4 GHz or 5 GHz and thus negatively impacting the operations of nearby Wi-Fi-based devices to a certain extent. In addition, most of them involve multiple transmission links, often calling for large antenna arrays and customized hardware for processing to hinder their adoption for commodity device-based applications.

3 PROBLEM STATEMENT

This article aims at developing a new device-free localization system via inaudible acoustic sensing by harnessing pervasively available **commercial off-the-shelf** (**COTS**) smart home devices. The speaker of a smart device is controlled to emit the inaudible acoustic signals at $20kHz$, while the built-in microphone array receives the reflected signals from the moving target for analyses in order to track its moving trace. Although the chirp signal-based solutions have been extensively pursued for localization, they have inherent drawbacks. First, the COTS devices are originally manufactured for sending the continuous waves, but when controlled to generate the chirp signals in the inaudible frequency band (i.e., $18kHz$ to $23kHz$), the electric burst at the speaker may bump noise audible to a human. Such noise sounds like "Beep", which can annoy people's life. Second, the chirp signals require large bandwidth for accurate tracking [17], but their associated ultrasonic signal is more likely to be heard by some sensitive people. Third, due to the frequency fluctuation of chirp signals, their reflection echoes from the environment will overlap with those from the moving target, making them hard to be differentiated and thus incurring potential location estimation errors.

3.1 Our Goal and Challenges

To overcome the limitations of chirp-based localization, we control the speaker to emit the continuous waves of inaudible acoustic signals while measuring the phase changes of Doppler shifts for

tracking the walking trace of a human in room-scale environments. Our approach is expected to have three salient features when compared with chirp-based methods. First, the continuous waves are generated by following the original design of COTS devices, so their controlled transmissions in the inaudible frequency will not generate audible noise. Second, the static environment reflection will have a relatively stable frequency, which can be easily differentiated from shifted frequencies caused by a moving target. Third, the distance resolution is superior, given that the phase measurement can track the subtle distance variation via a small phase shift. Besides, analyzing the phase shift of Doppler signals can eliminate any potential error caused by the Doppler effect. Nonetheless, a set of technical challenges surfaces in designing such a system for deployment in home environments, outlined as follows:

— The previous solutions for finger, hand, or chest tracking by measuring the phase change of continuous waves, are inapplicable here since they consider the original signal as a baseband signal. However, the human moving often has a speed high enough to cause the clear Doppler effect of original signal. For example, a continuous acoustic wave with a frequency of $f_c = 20$kHz reflected by a walking person with a speed of $v = 0.5m/s$ incurs $29Hz$ Doppler shift on the baseband signal. Such a frequency shift will cause a considerable error in phase estimation, if the original signal is considered as a baseband signal. Hence, it is necessary to design new methods toward identifying the shifted baseband signals for accurate phase estimation.
— Different body segments will generate various Doppler shifts, with their phases mixed at the receiver side. It is challenging to isolate them and identify the respective baseband signal of each body segment, while grouping those shift phases associated with the same person in a room where multiple persons exist.
— The multipath reflections will cause strong interference with target signals of interest in room environments and they fluctuate with the movement of a target. It is challenging to mitigate multi-path effects, calling for new solutions for differentiating the target reflections.

This article aims at overcoming the aforementioned challenges and develop the first continuous wave-based localization system based on the phase measurement, to be deployable for use in home environments. We briefly outline how we overcome the challenges mentioned above.

— To address the first challenge, our proposed system will take into account the Doppler shift as the baseband frequency. We achieve this by calculating the velocity of the body for determining the associated Doppler shift, and then combining it with the original frequency emitted by the speaker to serve as the baseband frequency.
— To tackle the second challenge, we'll segment the baseband signal into a series of narrowband signals, based on the fact that each narrowband is likely to contain reflections from at most one body segment. Subsequently, we compute the motion states, encompassing velocity, range, and angle of arrival, for the body segments within these narrowband signals. Ultimately, a clustering algorithm is employed to group together reflection signals originating from the same individual.
— The third challenge will be addressed by spectral subtraction, where we record the background reflection of the environmental objects and then subtract it from the recorded signals to get a clean signal of the human reflection.

Our solution is expected to work not only for the single person tracking, but also applicable for tracking multiple persons. Before presenting our design details in the next section, we briefly provide some preliminary knowledge relevant to our design next.

3.2 Preliminary Knowledge

Parameters of Human Motion. To track the moving trace of a human, the following parameters are needed: (1) *Range,* indicating the absolute distance between the target and the device; (2) *Velocity,* representing the moving speed of a target (with a target signifying one body segment); (3) ***Angle-of-Arrival (AoA)***, denoting the angle of a target corresponding to the device.

Doppler Effect. In our system, the movement of a target will generate the frequency shift of the original signal, which can be captured for sensing its motion status. The frequency of Doppler signal is determined by the signal frequency and velocity of the signal transmitter and receiver.

If the signal transmitter is stationary and the receiver is moving at a velocity of v, the frequency of its Doppler shift can be represented as

$$f_1 = \frac{c+v}{c} f_0, \tag{1}$$

where c is the sound propagation speed in the air. If the signal receiver is stationary and the transmitter is moving at a velocity of v, the frequency of its Doppler shift can be represented as

$$f_2 = \frac{c}{c-v} f_0. \tag{2}$$

Hence, the motion state of the signal transmitter or receiver significantly impacts the Doppler frequency. In our problem, a moving person can be seen as a moving signal receiver, if the signal reaches the body from the speaker, or as a virtual moving signal transmitter, if the signal is reflected from the body to the microphone array. As such, the Doppler frequency of a moving target can be expressed by

$$f_v = \frac{c+v}{c} \frac{c}{c-v} f_0 = \frac{c+v}{c-v} f_0. \tag{3}$$

4 SYSTEM DESIGN

In this section, we elaborate our design of the continuous-wave based localization system. Figure 1 presents the workflow of our system, consisting of four component modules: *Sensing, Signal Processing, Signal Modeling*, and *Target Localization*. The *Sensing* module involves a speaker to generate the continuous wave at 20 kHz and a microphone array to receive the reflected signals. In the *Signal Processing* module, we first perform the STFT on received signals to generate their spectrogram. Then, we develop a solution to eliminate the interference signals that (1) come from the direct transmission and (2) are reflected from surrounding objects. What remains is a clean spectrogram, which is next divided into a set of narrowband signals, aiming at segregating signal components that are reflected from different human body segments. The *Signal Modeling* employs a series of our developed solutions for estimating the baseband signal in each narrowband signal, calculating the velocity according to its frequency, and analyzing the phase change of each narrowband signal for obtaining the signal phase with the aid of range, AoA, and velocity. Finally, the *Target Localization* module relies on our developed estimator for estimating the location parameters of body segments. It applies K-means clustering and L1-norm to fuse the estimation of a moving person, with the multipath effect mitigated to get the moving trace of a target. The details of each design component are provided next.

4.1 Sensing

The speaker is controlled to keep sending the $20kHz$ continuous waves. We select $20kHz$ because it is imperceptible by most people and is also large enough to produce a significant Doppler effect. The signal will be transmitted at the low power for saving energy while the microphone array keeps sensing the sound pressure between $19kHz$ and $21kHz$. Once it senses the power

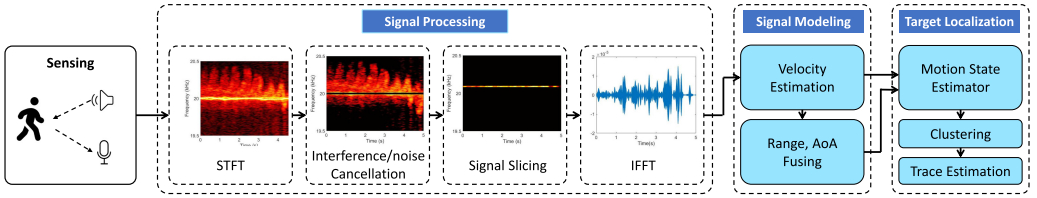

Fig. 1. The workflow of our continuous waves-based localization system via inaudible acoustic sensing.

level exceeding a certain threshold due to Doppler signals caused by a moving target, our system is triggered to transmit the high powered continuous wave at frequency $f_0 = 20kHz$ with received signals processed for localization. The transmitted continuous wave can be represented as follows:

$$F(t) = A\cos(2\pi f_0 t)\,. \tag{4}$$

The microphone array receives the Doppler shift signals, which are reflected from different body segments of a moving person, at a sampling frequency of $44.1kHz$. Such a sampling frequency ensures the reflected signals are entirely reconstructed from the signals recorded by the microphone array, according to the Nyquist Sampling Theorem.

4.2 Signal Processing

The recorded signals via the microphone array are inputted to the *Signal Processing* module for eliminating the interference and then split into a set of narrowband signals.

4.2.1 Short Time Fourier Transform Processing. We apply the **short-time Fourier transform (STFT)** to generate the spectrogram of signals. In particular, the signals are sliced by a set of small time windows with a length of 0.3s. Two consecutive time windows are overlapped with 95%, meaning an $1s$ signal will be sliced into 66 bins. Each small bin is multiplied by a Hamming window, and then a 21000-point **Fast Fourier transform (FFT)** is applied to each bin. We divide the frequency into 21000 sub-bands, having a frequency resolution of $1Hz$.

4.2.2 Interference Cancellation. A moving person can be considered as a cluster of moving body segments (e.g., head, torso, arms, and legs), generating a collection of disparate reflection signals to be received by the microphone array. Besides, the direct transmission signals from the speaker and reflections from stationary objects are also received by the microphone array. Hence, the composite signals received by the microphone array can be modeled by

$$R(t) = A\cos(2\pi f_0 t) + \sum S_r + \sum A_n \cos(2\pi f_0 (t-\tau_n)) + N(t), \tag{5}$$

where $A\cos(2\pi f_0 t)$ is the signal directly transmitted from the speaker to the microphone array. $\sum S_r$ represents the combined signal reflected from different moving body segments. Since the move of each body segment causes a Doppler effect, the frequency of each S_r differs among one another and also differs from f_0. The term $A_n \cos(2\pi f_c(t-\tau_n))$ represents the signal reflected from surrounding objects. Since these objects are stationary, their reflections have the same frequency as the transmitting signal f_0. $N(t)$ indicates the noise caused by system defects. In this step, we aim at eliminating three categories of signals: (1) the direct transmission signals, (2) signals reflected from surrounding stationary objects, and (3) the noise.

Given the first two categories of signals have the frequency of $f_0 = 20kHz$, we can directly set the spectrogram power at $20kHz$ to 0 for elimination. However, in practice, the smart home speaker is not designed to generate the high-frequency ultrasound, so the frequency actually fluctuates slightly over time. We cannot completely remove them by setting the spectrogram power at $20kHz$

to 0. Our empirical studies exhibit that such signal fluctuation typically ranges from $19.99kHz$ to $20.01kHz$. Hence, we set the spectrogram power in this range to be 0 instead, for fully removing direct transmission signals and reflections from surrounding objects.

On the other hand, the noise in the ultrasound domain is caused by system defects, seen as the points spreading over the spectrogram. Such noise does not vary with time, allowing us to subtract it by spectral subtraction directly. In particular, we let the microphone periodically record the noise in the static environment and then subtract the latest recorded noise from the current signal.

4.2.3 Signal Slicing. After the above process, we get a relatively clean spectrogram for $\sum S_r(t)$, which can be considered as the pure Doppler signal, expressed as

$$\sum S_r(t) = \sum \alpha_m \cos\left(2\pi f_m(t - \tau_m)\right), \tag{6}$$

where f_m represents each baseband signal frequency and τ_m indicates its ToF. Notably, the multipath reflection is also included therein, and its effect elimination will be described in Section 4.4.3. Each body segment can be viewed as a virtual signal transmitter. Hence, the baseband signal represents the signal transmitted from the virtual signal transmitter to the microphone array. The virtual signal senders (corresponding to different body segments) have different speeds during human walks or moves, thus generating baseband signals at different frequencies. As such, Equation (6) can be considered as the superimposition of multiple baseband signals with different delays. Our goal is to separate the $\sum S_r$ into multiple narrowband signals corresponding to different body segments.

We next analyze the spectrogram of $\sum S_r(t)$. Considering the human walking/moving speed in a room is no more than $4m/s$, we only need to take into account signals at the frequency range of $19.5Hz$ to $20.5Hz$, to sufficiently cover all Doppler shifts from a moving person. Assume $F(f, t)$ represents the Doppler energy of frequency f at a certain time t on the spectrogram. Since $F(f, t)$ changes with the frequency, by analyzing its peaks, we can identify the frequency components of $\sum S_r(t)$ that signify the Doppler frequencies of different body segments. Our empirical study exhibits the averaged difference between two consecutive frequency components is about $20Hz$. Thus, we divide the entire spectrogram into a set of $20Hz$ narrowband signals. The $20Hz$ band results in $0.05m/s$ velocity resolution, sufficiently to differentiate the Doppler shifts from different body segments. As such, each narrowband signal can be represented as

$$S_r(t) = \alpha_m \cos\left(2\pi f_m(t - \tau_m)\right), \tag{7}$$

Notably, τ_m is a one-way ToF that corresponds to the time taken by the signal to travel from a virtual signal sender to the microphone. Since only the frequency range of $19.5Hz$ to $20.5Hz$ is taken into account with the narrowband signal of $20Hz$ bandwidth, the microphone covers 48 narrowband signals. But, not all narrowband signals contain the Doppler reflections from moving body segments, so those narrowband signals whose averaged amplitudes below a predefined threshold are discarded.

4.3 Signal Modeling

Each narrowband signal is considered to be associated with one specific virtual signal transmitter. We then perform I-Q decomposition to analyze its phase, which is affected by τ_m and determined by the range, angle, and velocity, according to Equation (7). After the phase of the narrowband signal is analyzed, we then build a signal model to formulate the narrowband signal in terms of three parameters, i.e., range, angle, and velocity.

I-Q decomposition: Figure 2 illustrates the process of I-Q decomposition. The narrowband signal $S_r(t)$ is multiplied by the continuous wave of $cos2\pi f_m(t)$ and its 90-degree phase-shifted version

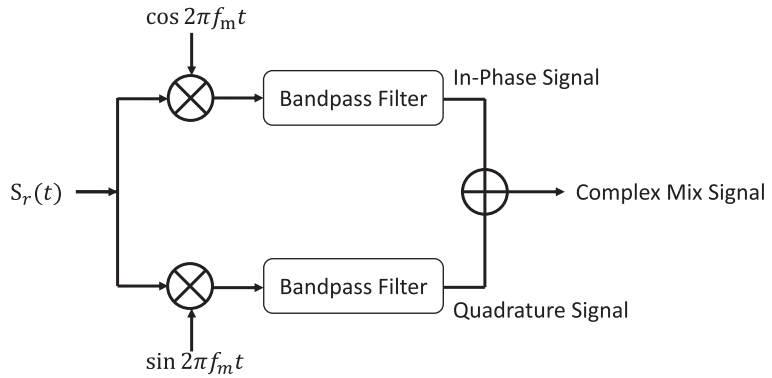

Fig. 2. I-Q decomposition.

of $sin2\pi f_m(t)$, where f_m denotes the frequency when the original $20kHz$ signal is received by the moving body segment, and it can be calculated via Equation (1). Since v is unknown yet, so we need to calculate v first, done by directly applying FFT to the narrowband signal to get its frequency f_v. According to Equation (3), we can derive the velocity v of the narrowband signal and then apply Equation (1) to obtain f_m.

With f_m, we next perform I-Q decomposition. According to the fact of $\cos A \cos B = \frac{1}{2}(\cos(A+B) + \cos(A-B))$, we get $\cos(A-B)$ by filtering out the high frequency component of $\cos(A+B)$ through a low pass filter. Thus, In-Phase signal $I(t)$ and Quadrature signal $Q(t)$ can be represented as

$$I(t) = \frac{1}{2}\alpha_1 \cos 2\pi f_m \tau,\ Q(t) = \frac{1}{2}\alpha_1 \sin 2\pi f_m \tau. \tag{8}$$

Figures 3 and 4 present the 0.1s examples of In-Phase signal and of Quadrature signal, respectively, with the baseband frequency equal to $20.045kHz$. From the two figures, we clearly observe a phase change regarding the waveforms of I-signal and Q-signal. By combing them, we arrive at a complex signal, denoted by

$$S^M(t) = I(t) + jQ(t) = \frac{1}{2}\alpha_1 e^{j2\pi f_m \tau}. \tag{9}$$

Figure 5 shows the I-Q trace of a complex signal, in which the I-Q trace moving around one circle corresponds to a 2π phase change. From Equation (9), the phase change is caused by the change of τ.

Phase Analysis: We next analyze the phase of a complex signal by first obtaining τ, since it determines the phase. Notably, τ is the one-way traveling time from an associated moving body segment to the receiver, and it varies according to the change of distance between the segment and the receiver. Suppose that a target is moving across the distance of $r(t)$ and with the angle of θ. Then, $r(t)$ causes a one-way time of $\frac{r(t)}{c}$, where c is the sound speed. Our experiment employs a liner 4-microphone array, with the distance between its two adjacent microphones equal to d. As Figure 6 shows, for the k^{th} microphone, the respective ToF of a received signal travels for an extra time of $\frac{(k-1)d\cos\theta}{c}$ compared to that received by the first microphone (the rightmost one). Hence, ToF for the signal received at the kth microphone can be computed by

$$\tau = \frac{r(t)}{c} + \frac{(k-1)d\cos\theta}{c}. \tag{10}$$

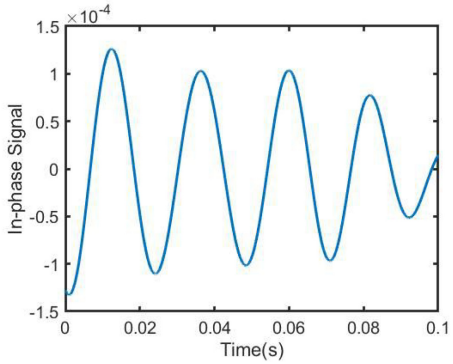

Fig. 3. I-signal waveform.

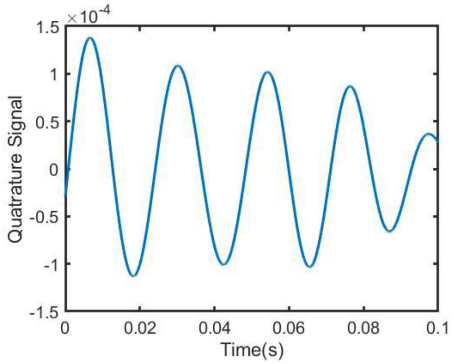

Fig. 4. Q-signal waveform.

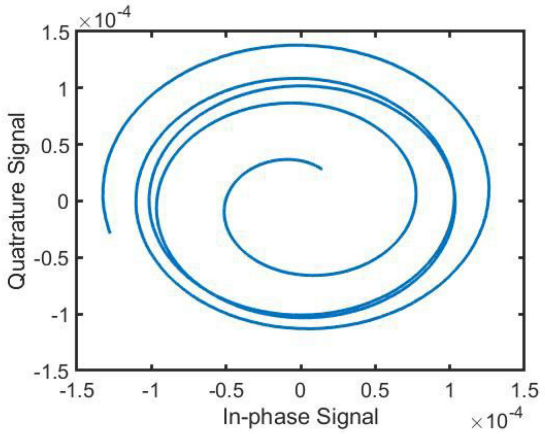

Fig. 5. Complex I-Q trace.

With Equation (9), we have the complex signal at the kth microphone:

$$S^M(t) = I(t) + jQ(t) = \frac{1}{2}\alpha_1 e^{j2\pi f_m(\frac{r(t)}{c} + \frac{(k-1)d\cos\theta}{c})}, \tag{11}$$

where f_m is the baseband frequency that already has been estimated. In Figure 5, the I-Q trace combines 0.1s worth of I-single and Q-signal, and the baseband frequency f_m is $20.045kHz$. There are about 4.3 circles in the figure, and the phase change of τ in 0.1s is $4.3 \times 2\pi = 8.6\pi$. In this small time period, the velocity and AoA are deemed constant. So, the distance change is due solely to a change in $r(t)$. We can calculate the distance change Δr through the accumulated phase change of τ given by $2\pi\frac{f_m \Delta r}{c} = 8.6\pi$ to yield $\Delta r = 0.073m$.

To mathematically model this, we consider the velocity v of each moving body segment to be constant. Then, $r(t)$ can be denoted as $r(t) = r + vt$, where r is the initial range. $S^M(t)$ is thus expressed by

$$S^M(t) = I(t) + jQ(t) = \frac{1}{2}\alpha_1 e^{j2\pi f_m(\frac{r}{c} + \frac{vt}{c} + \frac{(k-1)d\cos\theta}{c})}. \tag{12}$$

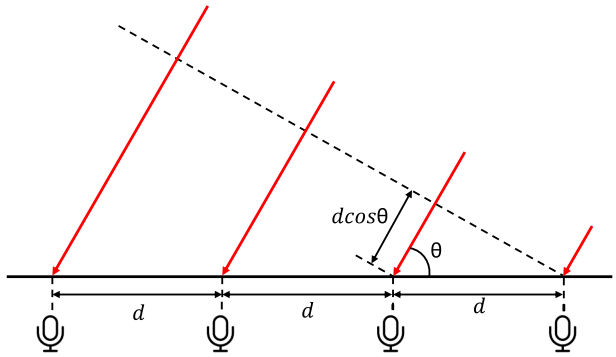

Fig. 6. Additional distance caused by the microphone array structure.

We represent this signal model as $S^M = \frac{1}{2}\alpha \cdot R \cdot V \cdot \Theta_k$, where R, V, and Θ_k are defined as follows:

$$\begin{aligned} R(r) &= e^{j2\pi \frac{r}{c} f_m}, \\ V(v) &= e^{j2\pi \frac{vt}{c} f_m}, \\ \Theta_k(\theta) &= e^{j2\pi \frac{(k-1)d\cos\theta}{c} f_m}. \end{aligned} \tag{13}$$

Since v is already calculated by the previous step, we only need to consider R and Θ components when analyzing r and θ of each body segment in its associated narrowband signal.

4.4 Target Estimating

In this section, we first estimate r and then θ of each narrowband signal and cluster the estimated parameters corresponding to the same person.

4.4.1 Motion State Estimator. In each round of estimation, we take into account those narrowband signals that last for $0.1s$. Given the sampling rate of $44.1kHz$, each narrowband signal contains $N = 4410$ samples. As discussed in Section 4.2.3, the narrowband signals that contain no Doppler reflections are discarded. We denote the number of remaining narrow bands as B and represent all narrowband signals by $\Sigma = [S_{r1}, S_{r2}, S_{r3}, \ldots, S_{rb}, \ldots, S_{rB}]^T$. On each narrowband, since we have signals from M microphones, after I-Q decomposition, S_{rb} can be represented as $S_{rb} = \left[S^{b1}, S^{b2}, \ldots, S^{bk}, \ldots, S^{bM}\right]^T$, where S^{bk} denotes the signals received at the kth microphone. Since $S^{b1}, S^{b2}, \ldots, S^{bM}$ are considered to have the same r, θ, and v, according to Equation (13), we can formulate an optimization problem for each narrowband signal S_{rb} as follows:

$$(\theta, r) = \arg\max \left| \sum_{k=1}^{M} S^{bk} \cdot R^*(r) \cdot \Theta_k^*(\theta) \right|, \tag{14}$$

where $(\cdot)^*$ indicates the conjugate operation. With the correct θ and r are estimated, Equation (14) gives rise to the local maximum value. Due to the searching range of $(-90°, 90°)$ for θ, from Equation (13), we can see $\Theta^*(\theta)$ is not a periodic function. Hence, θ has a unique solution corresponding to each S_{ri}. However, since r is in the range from $0m$ to several meters, $R^*(r)$ is a periodic function, making the optimal solution of r nonunique and calling for a further process to estimate the correct r.

Consider the two body segments of upper limbs with their motion states (r_1, v_1, θ_1) and (r_2, v_2, θ_2), respectively. Although their speeds and angles may be different due to the motions

of body segments, their horizontal ranges are similar. As such, to get the correct r, we can consider that some body segments from the same person have the same distance to the microphone array. Equation (12) reveals that each narrowband signal's initial phase is affected by r and θ. Since we can remove Θ from a narrowband signal by multiplying its corresponding Θ_k^*, the initial phase of the remaining signal is determined solely by r. Denote ϕ_1 and ϕ_2, as the initial phases of two narrowband signals, which can be extracted directly from signals themselves. Assuming their corresponding frequencies are f_{m1} and f_{m2} after removing Θ. According to Equation (12), for two narrowband signals with the same distance, we have $\phi_1 = 2\pi f_{m1}\frac{r}{c}$ and $\phi_2 = 2\pi f_{m2}\frac{r}{c}$, which give rise to $\frac{\phi_1}{f_{m1}} = \frac{\phi_2}{f_{m2}}$, implying that the ratio of the initial phase over the baseband frequency is identical for all signals with the same distance. Based on this result, we calculate phase-frequency ratios with respect to all Z narrowband signals. For each narrowband signal, we identify another one with the closest phase-frequency ratio for conjecture to have the similar distance. This results in a total of Z pairs, with each pair considered to have the same distance. For each pair of narrowband signals with their frequencies of (f_{m1}, f_{m2}) and their initial phases of (ϕ_1, ϕ_2), we have $2\pi f_{m1}\frac{r}{c} - 2\pi f_{m2}\frac{r}{c} = \phi_1 - \phi_2$. Then, the distance r can be calculated by $r = \frac{\Delta\phi c}{\Delta f}$, where $\Delta\phi$ is the initial phase difference and Δf is the baseband frequency difference of two baseband signals in this pair. Next, r is fed to the optimizer characterized by Equation (14) for refinement. That is, we slightly shift r's value to find the closest optimal solution. In the end, we can have Z sets of (r, v, θ) values in total.

Until now, we get the motion states from different body segments and are yet to identify the moving person. The next subsection describes a clustering algorithm to associate all proper (r, v, θ) values with the same person, given that those (r, v, θ) values represent the person's motion states.

4.4.2 Clustering Motion States. The reflections from different body segments of a moving person have relatively identical θ and r values, even their v values are different. On the other hand, the reflections from body segments of different persons possess markedly different (θ, r, v) values. Based on these facts, it is possible to cluster (θ, r, v) values associated with the same person together. Here, we apply the K-means algorithm to perform such clustering over Z sets of (r, v, θ) values, with each set treated as one point in the 3-D space and inputted to the algorithm.

We employ the silhouette value [28] to help improve our K-means clustering. That is, corresponding to each point i, the silhouette value can be defined as follows:

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}},$$

where a_i indicates the averaged distance from the ith point to all other points in the same cluster, and b_i is the minimum averaged distance from the ith point to points in each other cluster. So, the silhouette value ranges from -1 to +1, with a high value signifying the object well matched with its own cluster and poorly matched to other clusters. If the majority of points have the high silhouette values, good clustering results. Otherwise, the clustering configuration is inappropriate, likely to have too many or too few clusters. Here, the averaged silhouette value over all data points quantifies the clustering effectiveness degree. We explore K in a proper range to select the value which yields the highest averaged silhouette value.

After getting a clustering configuration, potential errors may still exist. For example, one person's corresponding points may be separated into multiple clusters. In this case, we rely on the inter-cluster distance to further mitigate errors, by calculating the euclidean distance between two clusters' centroids. If the inter-cluster distance of two clusters is below a pre-defined threshold, they are merged into one cluster.

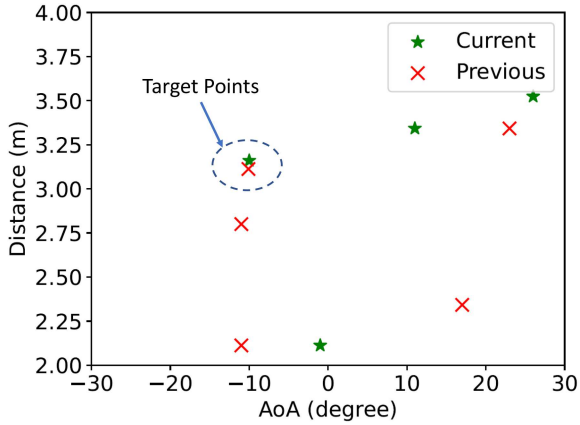

Fig. 7. Identifying the target points with the help of the previous estimation.

Ideally, we aim at clustering all points corresponding to the same person together through this step. However, the multipath effect may cause wrong (r, v, θ) values. Besides, the same narrowband may include the signals of different body segments, albeit to a low probability but warranting further refinement. The next subsection outlines a trace estimation solution for addressing aforementioned issues to enhance localization through refining our results.

4.4.3 Trace Identification. Given that a person's movement is continuous, the two consecutively estimated (r, v, θ) values of the same person should be similar. Since a rather short signal of 0.1s is considered at a time, the changes in a person's angle, range, and velocity are very small. Hence, we can take into account a sequence of two consecutive 0.1s estimations to measure their difference per pair for the trace identification of a moving person. For each cluster j at the time point of $t-1$, we calculate its centroid, i.e., the averaged r, v, and θ values over all points, indicated as $r_{t-1}(j)$, $v_{t-1}(j)$ and $\theta_{t-1}(j)$, respectively. Similarly, for each cluster i at the time point of t, we calculate the averaged r, v, and θ values over all points, denoted as $r_t(i)$, $v_t(i)$, and $\theta_t(i)$, respectively. For each cluster pair of i and j, their L-1 distance is obtained by

$$L1(i,j) = |r_t(i) - r_{t-1}(j)| + |v_t(i) - v_{t-1}(j)| + |\theta_t(i) - \theta_{t-1}(j)| \,. \tag{15}$$

For each i, we find corresponding j that has the minimum $L1(i,j)$. Notably, if i and j are associated with the same person, the L-1 distance should be very small. Otherwise, if i is from a multi-path reflection, its corresponding L-1 distance tends to be large due to the quick change of reflection paths. This way identifies the moving trace of a person, starting from a cluster associated with the person. Figure 7 show an example for cluster centroids of two consecutive 0.1s signals, plotted in a 2-D plane. In the figure, the green points show the currently estimated points, and red points show the previously estimated points, plotted in a 2-D plane. Comparing the previous points and current points, we can see the positions of the circled points have almost no changes, whereas the positions of other points change largely. Thus the circled points are considered as the target points from the same person. Other points are considered to be caused by multi-path reflections.

5 EXPERIMENT

We have implemented our motion tracking system for conducting extensive experiments to evaluate its performance.

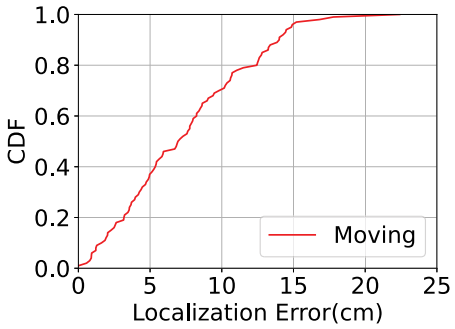

Fig. 8. Localization errors when locating a single person.

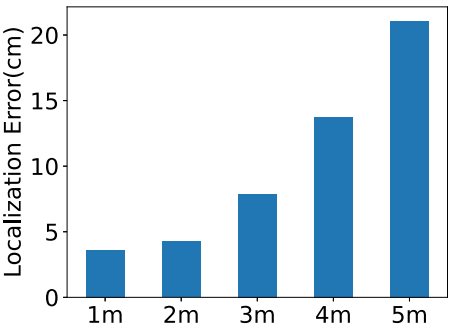

Fig. 9. Localization errors at different ranges.

5.1 Experiment Setup

Our system comprises a ReSpeaker 4-Mic Linear Array, which is connected to a Raspberry Pi 4 and a co-located Edifier R1280DB speaker. Note that smart devices do not release their APIs, so we cannot directly program their microphones for our purpose. Instead, we are using the ReSpeaker 4-Mic Linear Array, which has a layout similar to those of current smart devices. This microphone array has four microphones, with the distance between two adjacent microphones is $5.08cm$. We place the microphone array atop the speaker to ensure that both the microphone and the speaker are located in the same place. When our system is triggered, the speaker is controlled to send continuous sinusoid wave at $20kHz$ with a sampling frequency of $44.1kHz$. The transmission power is tuned to 80% of the maximum volume, so the measured power at 1 meter away from the speaker is 45dB. The microphone array records reflected signals, which are saved via the Raspberry Pi 4 at the sampling rate of $44.1kHz$. The Raspberry Pi 4 is connected to a laptop via Wi-Fi connection and all signals are processed via Matlab in this laptop. We measure the located positions of moving persons in the 2-D plane, by creating their corresponding trajectories on the floor to serve as the ground truth. The localization error is used as our evaluation metric, defined by the Euclidean distance between each located position and its ground truth coordinate.

5.2 Performance on Locating a Single Moving Person

In this experiment, we let a person walk at his natural speed along the predefined trajectory. Our system processes the reflected signals to get this person's location at each 0.1s. The localization error in each 0.1s is calculated, with the CDF result depicted in Figure 8. This figure reveals that the localization error of 40% (or 80%) positions is less than $5.4cm$ (or $12.5cm$), with the mean localization error equal to $7.49cm$. Such results are promising and demonstrate that our system can accurately track a person's moving trace.

Further experiments are conducted to evaluate the localization errors of different distances away from the speaker. We truncate the collected data at five distance ranges, i.e., $(1m \sim 2m)$, $(2m \sim 3m)$, $(3m \sim 4m)$, $(4m \sim 5m)$ and $(5m \sim 6m)$, indicating by $1m$, $2m$, $3m$, $4m$, and $5m$. Figure 9 shows results of mean errors for distances of $1m$, $2m$, $3m$, $4m$, and $5m$, equal to $3.6cm$, $4.3cm$, $7.9cm$, $13.7cm$, and $21.09cm$, respectively. Obviously, a larger distance leads to a bigger error, as expected since the acoustic signals experience fast attenuation when the distance increases. The reflected signals are weakened for a bigger distance, causing Doppler signals harder to be extracted and thus yielding bigger errors in calculating phase and velocity to deteriorate performance. Nonetheless, the maximum error of $21.09cm$ at the range of $(5m - 6m)$ is far smaller than the human body size, making our system suffice for use in the room scale environment.

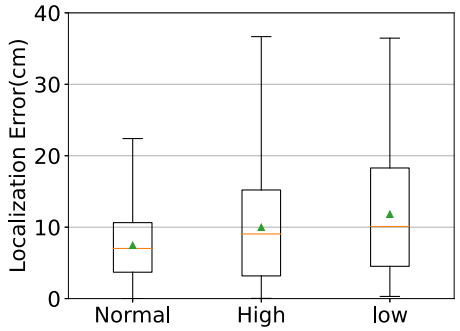

Fig. 10. Impact of different speeds.

5.3 Robustness for Single Person Localization

To show the robustness of our system, we take into account different factors and examine their impacts on our system performance.

Impact of Moving Speeds. We let a person walk at different speeds for examining system performance. Three different walking speeds are considered: (1) low speed ($< 1m/s$), (2) normal speed (around $1m/s$), and (3) high speed. Figure 10 shows the quartiles figure under the three moving speeds. From this figure, we observe our system to achieve the worse performance when a person is moving at the low or high speed, in comparison to moving at the normal speed. Specifically, the averaged errors are $7.4cm$, $10.1cm$, and $13.2cm$ respectively, with respect to the normal speed, low speed, and high speed. The reason is as follows. When a person is walking at the high speed, the Doppler effect is more apparent to make our system's assumption of a person's velocity considered as a constant in a short time period to yield a large error. On the other hand, a slow walking speed weakens the Doppler effect, resulting in light narrowband signals for localization, thereby degrading system performance.

Impact of Audible Sounds. Since our system works at the $20kHz$, inaudible to human, we next examine the possible impacts of pervasively existing audible sounds. Two audible sounds common in the home environments are considered: (1) humans talking and (2) music playing. The two sound sources are placed at $0.5m$ away from the speaker. Figure 11 shows the performance of our system under different scenarios, where *silent* indicates that no audible sound exists in the environment. When comparing three scenarios, we can observe the audible sounds (i.e., human talking and music playing) only have slight impacts on system performance. The averaged errors in a silent environment, a human talking environment, and a music playing environment are $7.4cm$, $9.3cm$, and $10.4cm$, respectively. The reason is that the two audible sounds do not generate any frequency component in the ultrasound domain, without causing interference to our system. The slight performance difference comes from frequency leakage caused by imperfect STFT, which produces certain unexpected high-frequency components, interfering with the target signal. But such frequency leakage typically is insignificant due to the use of continuous wave, yielding a minor impact.

Impact of Different Devices. We next show our system is transferable to different devices. Three different speakers are examined, i.e., Edifier R1280DB, Logitech z200, and Amazon Echo, indicated respectively as Speaker 1, Speaker 2, and Speaker 3. In our experiment, speakers' volumes are tuned to the same level and we let them emit the $20kHz$ continuous wave. Our system performance results of those three different speakers are shown in Figure 12. We observe that Speakers 2 and 3 underperform Speaker 1. The reason is that Amazon Echo generates signals in all directions,

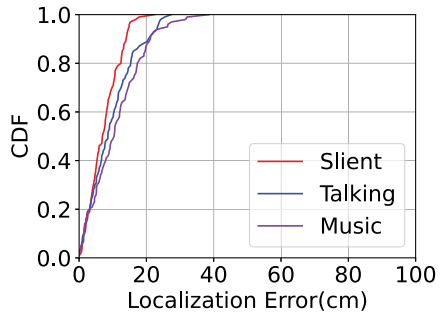

Fig. 11. Impact of audible sounds.

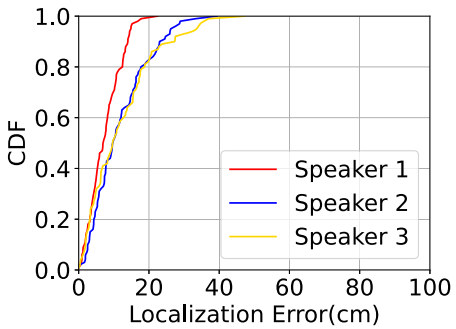

Fig. 12. Performance with different speakers.

making their reflected signals have relatively low power to result in a low SNR. Logitech z200 is the cheapest device, so its generated signals have lower quality than that of Edifier R1280DB's signals. However, their performance levels are still acceptable, i.e., 80% of localization errors being less than 18.8cm. Overall, the averaged localization errors of Speaker 1, Speaker 2, and Speaker 3 are 7.4cm, 11.7cm, and 12.1cm, respectively. Such experimental results demonstrate the good transferability of our system to different smart home devices.

Impact of Different Environments. We next conduct experiments in three rooms with different layouts to quantify the robustness of our system, with experimental results demonstrated in Figure 13. Our system is found to achieve the similar performance in three rooms, with the mean localization errors of 7.4cm, 9.6cm, and 9.0cm, respectively. This is due to the fact that reflections from all environments have the same frequency as the originally transmitted signal, which is fully eliminated via our noise cancellation step. It is thus concluded that our system is robust to the environmental changes.

5.4 Tracking Multiple Moving Persons

Our system is also applicable for locating multiple moving persons. We conduct experiments to show its performance when multiple moving persons co-exist in a room.

Localization Performance. Two persons walk on two predefined trajectories (i.e., straight lines) separated by 2m. Figure 14 shows the CDFs of localization errors under two moving persons, with mean errors equal to 22.6cm and 25.4cm, respectively. When comparing to Figure 8, we find the performance results are degraded since some Doppler signals from two bodies have identical

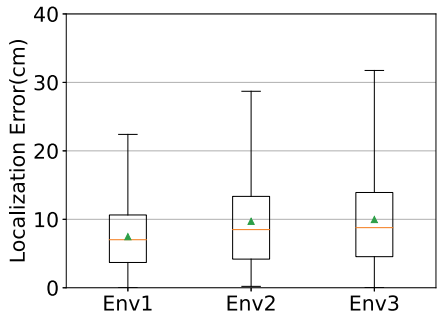

Fig. 13. Impact of different environment.

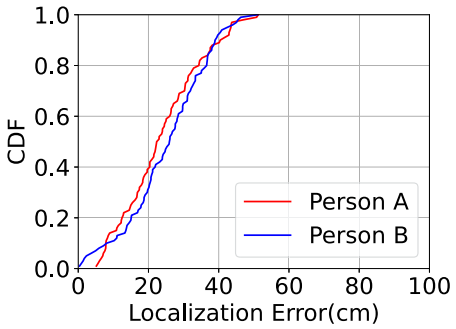

Fig. 14. Localization errors when locating two moving persons.

frequencies, overlapped in the same narrowband, and making them hard to be separated. Since the maximum error is less than 50cm, our system is accurate enough for tracking two persons' traces.

Impact of Distances between Two Targets. The impact of the distance separating two persons is assessed. We let two persons be separated 3m, 2m, 1m, and 0.5m, respectively, when walking toward the speaker from 5m away to 1m away. Figure 15 plots the localization errors under various separation distances. We observe that our system performance degrades as the separation distance drops.

Our system performs the best at the separation distance of 3m, with the averaged errors of 21.2cm, and 21cm, respectively, for the two persons. Under the separation distance of 0.5m, our system performs the worst, with the mean errors of 102cm and 99cm, respectively. The reason is that when two persons are closer to each other, their mutual interference hike, making our system harder to cluster them.

Performance under Different Target Counts. We increase the number of targets from 1 to 3 for examining the system performance results. Three sets of experiments are conducted, respectively under single person, two persons, and three persons. In each experimental set, every target walks toward the speaker at his/her natural speed. Figure 16 shows the CDF curves of localization errors with respect to the three target counts, with different color curves denoting the three target counts. Their averaged errors are 7.4cm, 24.1cm, and 51.2cm, respectively. It is observed that our system performance degrades faster with a larger target count. Especially, with three targets, more than 40% of measured points have their localization errors exceeding 50cm. The reason is that more targets increase the chance that Doppler signals from different targets overlap in the same narrowband and incur more complex multipath reflections, making it harder to separate signals

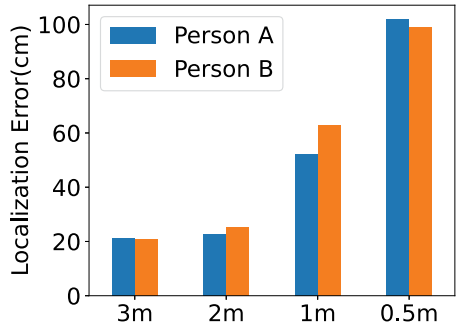

Fig. 15. Localization errors under different separation distances.

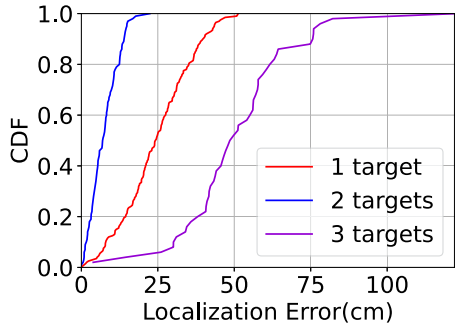

Fig. 16. Localization errors on different target counts.

from different persons. Nonetheless, the localization errors are still less than $1m$, useful for coarse location tracking.

5.5 Extension to Locate a Stationary Person

Since our system relies on the phase change of Doppler signals for localization, it cannot localize a stationary object. However, we can have an aid of arm swings to localize a stationary person. To this end, an experiment is undertaken to gauge the performance of localizing static targets, by letting the participants stand at predefined points with their arms swing. Figure 17 depicts the CDF curves, with the red curve (or blue curve) indicating the mean localization errors of a single person (or two persons). When comparing to Figures 8 and 14 (respectively for locating the single moving person and two moving persons), we find degraded performance. This is due to the fact that the reflections from arm swings are much weaker than those from the entire body segments, giving rise to inferior performance. However, the mean errors are of $21.1cm$ (or $35.4cm$) for a single person (or two persons). Such results are still promising, suggesting that our system is applicable to track the movement of a relatively small target or fine-grained movement, such as gesture tracking.

6 DISCUSSION

As the first work to explore the feasibility of continuous wave-based localization via the phase measurement of Doppler signals for room-scale location tracking, our current design exhibits some limitations that will be further addressed in the future.

First, our system performance will substantially degrade if a target moves extremely slowly due to its resulting weak Doppler effects. In this case, the Doppler signals are to have similar

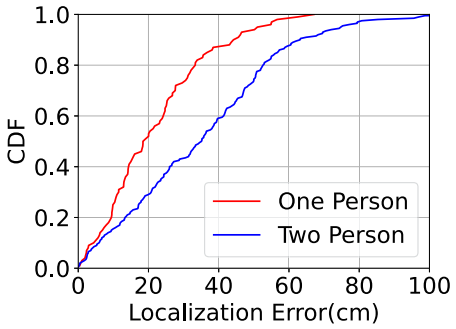

Fig. 17. Localization errors when locating the stationary persons.

frequencies as that of the direct transmission signals, but their strengths are much weaker. The current signal cancellation method fails to work since these two types of signals cannot be differentiated. Besides, the amplitude of direct transmission signals is far higher than that of Doppler signals, further preventing the proposed system from extracting correct phase information. To overcome such a limitation, one plausible method is to impose an additional requirement by letting the target swing his/her arm(s) in walking. Another direction is to develop new signal processing techniques able to separate weak Doppler signals from the direct transmissions. Such a problem remains challenging, deferred to our future work.

Second, our system can perform well for tracking up to two targets if their walking trajectories are in the constrained range, with noticeable degradation expected for tracking three or more targets. To overcome such limitations, one direction is to increase the microphone amounts while implementing a beamforming algorithm to differentiate multiple targets for higher resolution. Another direction is to leverage the circular microphone array, which can cover a larger range better. However, its omnidirectional transmission will incur intensive multi-path interference, calling for a more advanced signal processing solution to mitigate it.

Third, the system's performance on AoA estimation is limited, due to its need for uniform microphone spacing that often causes sidelobes to appear when estimating the AoA. Typically, the sidelobes may have their values close to the correct AoA, thus introducing ambiguity in AoA estimation. To mitigate the sidelobes effect, we can utilize the non-uniform microphone array [19] to mitigate AoA "side lobes" for better performance, yielding a higher AoA resolution. New signal processing solutions for dealing with such an effect are also planned in our future work.

Forth, our system cannot resolve multiple targets if their body segments have exactly the same velocity, albeit rather unlikely. Our experiment unveils that the proposed system can still discern participants walking at similar paces, provided that their body segments move at different speeds. If all body segments have exactly same speeds, however, their reflections possess the same Doppler frequencies, making them unable to be differentiated. In this case, we have to rely solely on the phase, which is different due to various ranges and angles. A new signal processing solution is required to separate two signals with the same frequency but different phases.

Fifth, the system cannot guarantee consistent performance across different environments. Its performance tends to decrease in environments with a higher density of objects due to more dynamic reflections from the complex surrounding objects, making it harder to isolate the signal from the human body. Meanwhile, our system's effectiveness tends to drop in larger spaces due to its limited working range. Nevertheless, the experimental result reveals that, even at a distance of 5 meters, our system still exhibits a localization error of $21.09cm$, sufficing for use in most indoor settings. In our future research, we intend to enhance the system by incorporating multiple devices

or utilizing various tone frequencies. This way aims at enriching the Doppler signals and capture more detailed localization information, especially in scenarios involving multiple individuals simultaneously. Another research direction involves the implementation of a robust signal processing algorithm, allowing for the direct separation of signals from distinct targets. Additionally, the integration of deep learning approaches to further refine the system's localization capabilities can be considered. We are to thoroughly investigate them in our future work.

7 CONCLUSION

This article has proposed a novel indoor location tracking system via the inaudible continuous waves, by leveraging the speaker and microphone array built in a commodity device. The proposed system controls the speaker of such a device to emit inaudible continuous waves for sensing and utilizes the co-located microphone array to record the reflected Doppler signals for analysis. A set of solutions has been developed for removing environmental interference, analyzing the phase change, and clustering reflections from a given person, plus result refinement, toward achieving accurate location tracking in room-scale settings. We have implemented our system with commodity devices and conducted extensive experiments to evaluate the performance of our system. Experimental results have demonstrated that our system is promising in effectively tracking the traces of up to three moving persons and is robust to various factors. The benefits of our system include being (1) reliant on the continuous wave, which is inaudible to humans, (2) implemented on COTS devices, and (3) generally applicable for room-scale tracking. Our proposed solutions and approaches are valuable and provide affluent insights, to benefit future work on tracking fine-grained movements.

REFERENCES

[1] Fadel Adib, Zachary Kabelac, and Dina Katabi. 2015. Multi-person localization via {RF} body reflections. In *12th {USENIX} Symposium on Networked Systems Design and Implementation*. 279–292.

[2] Fadel Adib, Zachary Kabelac, Dina Katabi, and Rob Miller. 2014. WiTrack: Motion tracking via radio reflections off the body. In *Proceedings of the NSDI*.

[3] Md Tanvir Islam Aumi, Sidhant Gupta, Mayank Goel, Eric Larson, and Shwetak Patel. 2013. Doplink: Using the doppler effect for multi-device interaction. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 583–586.

[4] Roshan Ayyalasomayajula, Aditya Arun, Chenfeng Wu, Sanatan Sharma, Abhishek Rajkumar Sethi, Deepak Vasisht, and Dinesh Bharadia. 2020. Deep learning based wireless localization for indoor navigation. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–14.

[5] Ziyang Chen, Panlong Yang, Jie Xiong, Yuanhao Feng, and Xiang-Yang Li. 2020. TagRay: Contactless sensing and tracking of mobile objects using COTS RFID devices. In *Proceedings of the IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 307–316.

[6] Tianbo Gu, Zheng Fang, Zhicheng Yang, Pengfei Hu, and Prasant Mohapatra. 2019. Mmsense: Multi-person detection and identification via mmwave sensing. In *Proceedings of the 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems*. 45–50.

[7] Wenchao Huang, Yan Xiong, Xiang-Yang Li, Hao Lin, Xufei Mao, Panlong Yang, and Yunhao Liu. 2014. Shake and walk: Acoustic direction finding and fine-grained indoor localization using smartphones. In *Proceedings of the IEEE INFOCOM 2014-IEEE Conference on Computer Communications*. IEEE, 370–378.

[8] IndustryArc. 2017. Indoor Positioning and Navigation Market - Forecast(2020 - 2025). Retrieved October 6, 2021 from https://www.industryarc.com/Report/43/global-indoor-positioning-navigation-market.html

[9] Guang-yao Jin, Xiao-yi Lu, and Myong-Soon Park. 2006. An indoor localization mechanism using active RFID tag. In *Proceedings of the IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing*. IEEE, 4–pp.

[10] Kiran Joshi, Dinesh Bharadia, Manikanta Kotaru, and Sachin Katti. 2015. Wideo: Fine-grained device-free motion tracing using {RF} backscatter. In *Proceedings of the 12th {USENIX} Symposium on Networked Systems Design and Implementation*. 189–204.

[11] Pei H. Leong, Thushara D. Abhayapala, and Tharaka A. Lamahewa. 2013. Multiple target localization using wideband echo chirp signals. *IEEE Transactions on Signal Processing* 61, 16 (2013), 4077–4089.

[12] Dong Li, Jialin Liu, Sunghoon Ivan Lee, and Jie Xiong. 2020. FM-track: Pushing the limits of contactless multi-target tracking using acoustic signals. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 150–163.
[13] Jie Lian, Jiadong Lou, Li Chen, and Xu Yuan. 2021. EchoSpot: Spotting your locations via acoustic sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–21.
[14] Jie Lian, Xu Yuan, Ming Li, and Nian-Feng Tzeng. 2021. Fall detection via inaudible acoustic sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–21.
[15] Kang Ling, Haipeng Dai, Yuntang Liu, Alex X. Liu, Wei Wang, and Qing Gu. 2020. UltraGesture: Fine-grained gesture sensing and recognition. *IEEE Transactions on Mobile Computing* 21, 7 (2020), 2620–2636.
[16] Jullawadee Maneesilp, Chong Wang, Hongyi Wu, and Nian-Feng Tzeng. 2012. RFID support for accurate 3D localization. *IEEE Transactions on Computers* 62, 7 (2012), 1447–1459.
[17] Wenguang Mao, Jian He, and Lili Qiu. 2016. Cat: High-precision acoustic motion tracking. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. 69–81.
[18] Wenguang Mao, Wei Sun, Mei Wang, and Lili Qiu. 2020. Deeprange: Acoustic ranging via deep learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–23.
[19] Wenguang Mao, Mei Wang, Wei Sun, Lili Qiu, Swadhin Pradhan, and Yi-Chao Chen. 2019. RNN-based room scale hand motion tracking. In *Proceedings of the 25th Annual International Conference on Mobile Computing and Networking*. 1–16.
[20] Carlyn J. Matz, David M. Stieb, Karelyn Davis, Marika Egyed, Andreas Rose, Benedito Chou, and Orly Brion. 2014. Effects of age, season, gender and urban-rural status on time-activity: Canadian human activity pattern survey 2 (CHAPS 2). *International Journal of Environmental Research and Public Health* 11, 2 (2014), 2108–2124.
[21] Santiago Murano, M. Carmen Pérez, Jesús Ureña, Chris J. Bleakley, and Carlos De Marziani. 2018. Comparison of Zadoff-Chu encoded modulation schemes in an ultrasonic local positioning system. In *Proceedings of the 2018 International Conference on Indoor Positioning and Indoor Navigation*. IEEE, 206–212.
[22] Rajalakshmi Nandakumar, Vikram Iyer, Desney Tan, and Shyamnath Gollakota. 2016. Fingerio: Using active sonar for fine-grained finger tracking. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 1515–1525.
[23] Rajalakshmi Nandakumar, Alex Takakuwa, Tadayoshi Kohno, and Shyamnath Gollakota. 2017. Covertband: Activity information leakage using music. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–24.
[24] Lionel M. Ni, Yunhao Liu, Yiu Cho Lau, and Abhishek P. Patil. 2003. LANDMARC: Indoor location sensing using active RFID. In *Proceedings of the 1st IEEE International Conference on Pervasive Computing and Communications*. IEEE, 407–415.
[25] Ke Sun, Ting Zhao, Wei Wang, and Lei Xie. 2018. Vskin: Sensing touch gestures on surfaces of mobile devices using acoustic signals. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. 591–605.
[26] Li Sun, Souvik Sen, Dimitrios Koutsonikolas, and Kyu-Han Kim. 2015. Widraw: Enabling hands-free drawing in the air on commodity wifi devices. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*. 77–89.
[27] Jiayao Tan, Cam-Tu Nguyen, and Xiaoliang Wang. 2017. SilentTalk: Lip reading through ultrasonic sensing on mobile phones. In *Proceedings of the IEEE INFOCOM 2017-IEEE Conference on Computer Communications*. IEEE, 1–9.
[28] T. Thinsungnoen, N. Kaoungku, P. Durongdumronchai, K. Kerdprasop, and N. Kerdprasop. 2015. The clustering validity with silhouette and sum of squared errors. In *Proceedings of the 3rd International Conference on Industrial Application Engineering*. 44–51.
[29] 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*. 1–11.
[30] Chong Wang, Hongyi Wu, and N.-F. Tzeng. 2007. RFID-based 3-D positioning schemes. In *Proceedings of the IEEE INFOCOM 2007-26th IEEE International Conference on Computer Communications*. IEEE, 1235–1243.
[31] Jue Wang, Fadel Adib, Ross Knepper, Dina Katabi, and Daniela Rus. 2013. RF-compass: Robot object manipulation using RFIDs. In *Proceedings of the 19th Annual International Conference on Mobile Computing and Networking*. 3–14.
[32] Jue Wang and Dina Katabi. 2013. Dude, where's my card? RFID positioning that works with multipath and non-line of sight. In *Proceedings of the ACM SIGCOMM 2013 Conference on SIGCOMM*. 51–62.
[33] Ju Wang, Jie Xiong, Hongbo Jiang, Kyle Jamieson, Xiaojiang Chen, Dingyi Fang, and Chen Wang. 2018. Low human-effort, device-free localization with fine-grained subcarrier information. *IEEE Transactions on Mobile Computing* 17, 11 (2018), 2550–2563.
[34] Junjue Wang, Kaichen Zhao, Xinyu Zhang, and Chunyi Peng. 2014. Ubiquitous keyboard for small mobile devices: Harnessing multipath fading for fine-grained keystroke localization. In *Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services*. 14–27.

[35] Tianben Wang, Daqing Zhang, Yuanqing Zheng, Tao Gu, Xingshe Zhou, and Bernadette Dorizzi. 2018. C-FMCW based contactless respiration detection using acoustic signal. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–20.

[36] Wei Wang, Alex X Liu, and Ke Sun. 2016. Device-free gesture tracking using acoustic signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. 82–94.

[37] Yanwen Wang, Jiaxing Shen, and Yuanqing Zheng. 2020. Push the limit of acoustic gesture recognition. *IEEE Transactions on Mobile Computing* 21, 5 (2020), 1798–1811.

[38] Teng Wei and Xinyu Zhang. 2015. mtrack: High-precision passive tracking using millimeter wave radios. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*. 117–129.

[39] Yaxiong Xie, Jie Xiong, Mo Li, and Kyle Jamieson. 2019. mD-Track: Leveraging multi-dimensionality for passive indoor Wi-Fi tracking. In *Proceedings of the 25th Annual International Conference on Mobile Computing and Networking*. 1–16.

[40] Panlong Yang, Yuanhao Feng, Jie Xiong, Ziyang Chen, and Xiang-Yang Li. 2020. RF-Ear: Contactless multi-device vibration sensing and identification using COTS RFID. In *Proceedings of the IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 297–306.

[41] Sangki Yun, Yi-Chao Chen, and Lili Qiu. 2015. Turning a mobile device into a mouse in the air. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*. 15–29.

[42] Sangki Yun, Yi-Chao Chen, Huihuang Zheng, Lili Qiu, and Wenguang Mao. 2017. Strata: Fine-grained acoustic-based device-free tracking. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. 15–28.

[43] Cheng Zhang, Qiuyue Xue, Anandghan Waghmare, Sumeet Jain, Yiming Pu, Sinan Hersek, Kent Lyons, Kenneth A. Cunefare, Omer T. Inan, and Gregory D. Abowd. 2017. Soundtrak: Continuous 3d tracking of a finger using active acoustics. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 1–25.

[44] Fusang Zhang, Zhi Wang, Beihong Jin, Jie Xiong, and Daqing Zhang. 2020. Your smart speaker can" hear" your heartbeat! *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–24.

[45] Qian Zhang, Dong Wang, Run Zhao, and Yinggang Yu. 2021. SoundLip: Enabling word and sentence-level lip interaction for smart devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–28.

[46] Yongzhao Zhang, Wei-Hsiang Huang, Chih-Yun Yang, Wen-Ping Wang, Yi-Chao Chen, Chuang-Wen You, Da-Yuan Huang, Guangtao Xue, and Jiadi Yu. 2020. Endophasia: Utilizing acoustic-based imaging for issuing contact-free silent speech commands. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–26.

[47] Yunting Zhang, Jiliang Wang, Weiyi Wang, Zhao Wang, and Yunhao Liu. 2018. Vernier: Accurate and fast acoustic motion tracking using mobile devices. In *Proceedings of the IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 1709–1717.

[48] Peijun Zhao, Chris Xiaoxuan Lu, Jianan Wang, Changhao Chen, Wei Wang, Niki Trigoni, and Andrew Markham. 2019. mID: Tracking and identifying people with millimeter wave radar. In *Proceedings of the 2019 15th International Conference on Distributed Computing in Sensor Systems*. IEEE, 33–40.

[49] Bing Zhou, Mohammed Elbadry, Ruipeng Gao, and Fan Ye. 2017. BatMapper: Acoustic sensing based indoor floor plan construction using smartphones. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. 42–55.

[50] Yanzi Zhu, Yibo Zhu, Ben Y. Zhao, and Haitao Zheng. 2015. Reusing 60ghz radios for mobile radar imaging. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*. 103–116.

Received 6 May 2023; revised 4 December 2023; accepted 4 February 2024