EchoSpot: Spotting Your Locations via Acoustic Sensing

JIE LIAN, University of Louisiana at Lafayette, USA
JIADONG LOU, University of Louisiana at Lafayette, USA
LI CHEN, University of Louisiana at Lafayette, USA
XU YUAN*, University of Louisiana at Lafayette, USA

Indoor localization has played a significant role in facilitating a collection of emerging applications in the past decade. This paper presents a novel indoor localization solution via inaudible acoustic sensing, called EchoSpot, which relies on only one speaker and one microphone that are readily available on audio devices at households. We program the speaker to periodically send FMCW chirps at 18kHz-23kHz and leverage the co-located microphone to capture the reflected signals from the body and the wall for analysis. By applying the normalized cross-correlation on the transmitted and received signals, we can estimate and profile their time-of-flights (ToFs). We then eliminate the interference from device imperfection and environmental static objects, able to identify the ToFs corresponding to the direct reflection from human body. In addition, a new solution to estimate the ToF from wall reflection is designed, assisting us in spotting a human location in the two-dimensional space. We implement EchoSpot on three different types of speakers, e.g., Amazon Echo, Edifier R1280DB, and Logitech z200, and deploy them in real home environments for evaluation. Experimental results exhibit that EchoSpot achieves the mean localization errors of 4.1cm, 9.2cm, 13.1cm, 17.9cm, 22.2cm, respectively, at 1m, 2m, 3m, 4m, and 5m, comparable to results from the state-of-the-arts while maintaining favorable advantages.

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

Additional Key Words and Phrases: Inaudible Acoustic Sensing, Localization, Device-free, Kalman Filter

ACM Reference Format:
Jie Lian, Jiadong Lou, Li Chen, and Xu Yuan. 2021. EchoSpot: Spotting Your Locations via Acoustic Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 3, Article 113 (September 2021), 21 pages. https://doi.org/10.1145/3478095

1 INTRODUCTION

Indoor localization has attracted wide research attention due to its potential of facilitating a variety of applications in smart homes such as security surveillance, elderly care, crowd monitoring, fitness tracking, etc. A report has shown that a person may spend almost 88.9% of the day indoors [32]. Also, the market value of indoor positioning and indoor navigation is expected to exceed $23.6 billion dollars in 2023 [18], substantiating that there is a large demand for effective indoor localization technology. The commonly used localization systems based on Global Positioning System (GPS) are not applicable to indoor environment, due to significant signal attenuation when

*Corresponding author.

Authors' addresses: Jie Lian, University of Louisiana at Lafayette, Louisiana, Lafayette, 70504, USA; Jiadong Lou, University of Louisiana at Lafayette, Louisiana, Lafayette, 70504, USA; Li Chen, University of Louisiana at Lafayette, Louisiana, Lafayette, 70504, USA; Xu Yuan, University of Louisiana at Lafayette, Louisiana, Lafayette, 70504, USA.

2474-9567/2021/9-ART113 $15.00
https://doi.org/10.1145/3478095

penetrating the wall that leads to meter-level localization error. This is unacceptably large in indoor environment, and reducing localization error at decimeter or centimeter level is highly desirable.

In the past decade, diverse technologies have been developed for indoor localization and tracking. While extensive methods have been developed for localization via camera [48, 66], motion sensor [24], inertial measurement unit (IMU) [64], floor sensor [33], or light sensor [17, 71], they either require a user to carry/wear sensors, or require to purchase and deploy the dedicate devices/sensors. These methods have the drawbacks of inconvenience for use or causing privacy issues. For example, wearable solutions may need the user to wear the device all day for continuous monitoring. The elders, sometimes, are likely to forget to wear devices, making it unsuitable for elderly monitoring. Also, the wearable solutions are not suitable for some scenarios, such as localizing undefined people. The camera-based solution could be accurate but likely to invade the user's privacy and make users feel uncomfortable. The RFID localization [19, 36, 51, 53], has become popular recently due to its low cost and ease of use. However, most RFID-based localization techniques are based on the assumption of knowing the tags' coordinates, which is impractical. Besides, many RFID-based systems heavily rely on ideal propagation models of RF phase or the received signal strength indicator (RSSI), which may not be feasible.

On the other hand, the radio-frequency (RF) sensing solutions leveraging the Wi-Fi and mmWave for localization have been extensively explored, producing promising results, i.e., decimeter-level accuracy in the several meters sensing range. However, the Wi-Fi-based solutions [1, 12, 20, 21, 41, 45, 50, 52, 60, 61] occupy the data communication channels of 2.4GHz or 5GHz which are already crowded with data traffic, inevitably impacting the nearby devices to some extent, especially for those using multiple channels so as to achieve good performance. In addition, most of such systems require regular maintenance, or some of them need specialized signals, hindering their wide deployment. The mmWave-based solutions [14, 39, 49, 69] do not cause interference to home devices, but they require the specialized mmWave radar which is typically expensive.

In contrast to the aforementioned solutions, acoustic sensing is promising for localization which can take advantage of the ubiquitously available audio devices without competing for radio resources with other home devices. The low sampling frequency of the audio signal enables signal processing to be implemented on a smart device. Existing efforts [7, 26, 27, 30, 35, 40, 68, 70] for human localization via acoustic sensing are device-dependent, requiring users to carry smartphones. Although device-free (without carrying the devices) acoustic solutions [5, 15, 31, 34, 54, 55, 65] have been proposed for tracking the movement of hand, finger, or mouth, for the purposes of localization, activity recognition, or authentication, their effective sensing ranges are extremely limited, no more than 1 meter.

In this paper, we propose a novel device-free localization solution via acoustic sensing, named EchoSpot, that can be implemented in the commercially available off-the-shelf (COTS) audio devices to work in home environments. Different from the existing work, we only rely on one speaker and one microphone with the reliance of wall reflection to precisely spot a human location in the two-dimensional space. Specifically, we program the audio device to control its built-in speaker to periodically emit inaudible FMCW (Frequency Modulated Continuous Wave) signals at 18kHz∼23kHz and use the co-located microphone to receive the reflected signals from objects for analysis. Based on the reflected signals, we generate time-of-flight (ToF) profile and aim to identify the peaks corresponding to the body reflection. By eliminating the influence from the device imperfection and from the environmental objects' reflection, we can identify ToF corresponding to the person reflection, allowing us to estimate the distance between the human body and the device. To locate a person's position, we continue to develop a new solution to identify the ToF corresponding to the wall reflection and obtain the path of human-wall-device. Then, we can calculate the position information of a person in a two-dimensional space. Considering the potential impact of the multipath effect, we further apply the Kalman filter to correct the position information. In the end, EchoSpot is implemented on the commercial speaker and microphone, and deployed in the real home environments for performance evaluation.

Comparing to the existing approaches, our contributions can be summarized as follows.

- We design a novel device-free localization system EchoSpot, by leveraging only one speaker and one microphone for precisely locating a human. It is a software-based system and can be implemented on the COTS audio devices for conducting the localization services, rendering a wider application range in the general house environment. While leveraging the acoustic signals at $18kHz \sim 23kHz$, it does not cause interference to the home Wi-Fi devices and is inaudible to human.
- A collection of acoustic signal processing techniques are developed, including generating the ToF profile, removing the impact from the device imperfection, environment objects, and multipath effect. In addition, a new solution is also proposed to identify the reflection path from the wall. After applying these techniques, EchoSpot can work suitably to spot a person's location in the home environment.
- We implement EchoSpot with the COTS speaker and microphone for proof-of-concept validation. We conduct experiments to demonstrate that EchoSpot can work effectively both to locate the static person and to track the moving person. Experimental results exhibit that EchoSpot can achieve the median errors of $8.5cm$ and $19.8cm$ for locating the static person and the moving person, respectively, comparable to the reported results from the state-of-the-art. In addition, EchoSpot achieves the mean errors of $4.1cm$, $9.2cm$, $13.1cm$, $17.9cm$,, $22.2cm$, respectively, at the distances of $1m$, $2m$, $3m$, $4m$, and $5m$, for locating the static person.

2 RELATED WORK

We review the state-of-the-art solutions for indoor localization that fall into three categories: 1) Sensor-based, 2) RF-based, and 3) Acoustic sensing-based localization.

Sensor-based Localization. A number of solutions for indoor localization have been developed by relying on camera, floor sensor, IMU sensor, light sensor, etc. In particular, the IMU sensor-based solution [64] tracked human motion by the captured acceleration and gyroscope data. Such methods are inconvenient and cumbersome as they need users to wear/carry the sensors, and it is likely that users will feel uncomfortable or forget to wear/carry the devices. Also, the IMU-based solutions could not directly provide the accurate location information, because they need to double integration of acceleration information, which introduce large error [10]. Many works also employed the camera for localization [2, 48, 62, 66]. However, such a method [47] highly depends on lighting conditions and features, thus is not robust in different environments. In addition, the dedicated camera is required for localization and tracking, which will incur non-negligible costs and raise privacy concerns. The floor sensor [33, 38] and light sensor [11, 17, 22, 28, 56, 59] are also used for localization, but they again require the purchase of dedicated sensors and incur considerable deployment overhead.

RF-based Localization. RF-based sensing for localization has widely attracted research attention, which leverages the variation of RF signals. Extensive works have developed solutions by relying on RFID for localization [19, 36, 51, 53], but they require users to wear RFID tags for detection, which is inconvenient and cumbersome. Although some contactless methods [6, 63] have been proposed, the sensing range is limited within 2 meters.

A series of prominent solutions have been proposed for localization with Wi-Fi signals, including but are not limited to [1, 12, 20, 21, 41, 45, 50, 52, 60, 61]. However, such Wi-Fi-based solutions occupy the data communication channels, especially for those using multiple channels at 2.4GHz or 5GHz by design to achieve good performance, inevitably competing for radio resources with other smart home applications. In addition, most of them require multiple transmission links, either using custom hardware (e.g., USRP or WARP) or deploying multiple access points (APs), to form large antenna arrays, which are typically not available at home. Moreover, custom hardware processing (e.g., software radio) that is unavailable in commodity devices may also be required. On the other hand, the mmWave-based solutions [14, 39, 49, 69] are also proposed for indoor location tracking, but they require the dedicated devices, i.e., the specialized mmWave radar, for emitting the ultra frequency signals, incurring non-negligible costs.

Acoustic Sensing. Recently, acoustic sensing has attracted considerable interest. This line of research takes advantage of the ubiquitous availability of speakers and microphones built in commercial devices, which is closely related to our work. Specifically, some research efforts have been undertaken to explore the device-dependent localization via acoustic sensing [7, 26, 27, 30, 40, 68, 70]. The nature of these solutions is to locate smartphones, requiring users to carry them to receive the acoustic signals for localization. On the other hand, some researchers have focused on exploring the device-free acoustic sensing for activity tracking, recognition, and authorization, which do not require users to carry sensors or smartphones. For instance, in [5, 15, 31, 34, 54, 55], acoustic systems have been developed to track hand/finger movement and recognize gestures by generating ultrasonic signal from speakers; the reflected signals from human body are captured by microphones, and the frequency shifts/phase shifts are then analyzed. In addition, [67] analyzed the Doppler shifts of reflected signals caused by user's articulatory gestures to achieve liveliness detection, while [29] further analyzed the uniqueness of Doppler profiles caused by mouth movement for the purpose of user authentication. BreathPrint [4] was proposed to sense user's breathing patterns via analyzing the reflected acoustic signals for authentication. The aforementioned device-free systems can achieve effective activity recognition, but they all require the devices to be close to the target body part and all rely on the strict multipath assumption, which cannot support our application scenarios at home use with the sensing range desired to be several meters. Covertband [35] implemented the device-free localization for tracking users and capturing various categories of human motions, but its localization relied on at least two microphones and one speaker. In contrast, our system only relies on one microphone and one speaker. Strata [65] was proposed for tracking the finger movement based on the phase change of a signal. It only works within a limited range (i.e., 0.5 meters), and the proposed phase-based method is not suitable for our application scenario since we target the room-scale localization, whereas the phase of signals can be significantly affected by the long-range multipath propagation. Instead, we apply the time-of-flight (i.e., TOF) to measure the signal propagation delay for calculating the distance. VoLoc [44] was proposed to localize a user's voice for spotting his position. Only when a user is talking, his location can be identified; otherwise, VoLoc cannot work. This system can be easily interfered by environmental noise, such as music playing, washer sound, refrigerator sound, and many others. In addition, the microphone array has to be leveraged in his system. In contrast, our EchoSpot relies on using the speaker to emit inaudible ultrasound and using the microphone to capture the reflected signals from the human body for localization. It is implemented on only one pair of speaker and microphone without requiring the microphone array while being robust against environment sounds, thereby more promising.

3 SYSTEM DESIGN

In this section, we present our design of the EchoSpot system, aiming to leverage only one speaker and one microphone available on off-the-shelf (COTS) device as the sensing tool to implement precise human localization. We program the audio device to control its built-in speaker to emit inaudible signals at 18kHz~23kHz and use the co-located microphone to receive the reflected signals from objects for analysis. For localization, a series of solutions need to be developed to identify the time-of-flight (ToF) for the signals of interests. In this context, a number of technical challenges are encountered to be addressed, briefly summarized as follows.

- The signal modulation is important to determine the resolution of capturing a person's subtle movement, which further limits localization precision. While we only rely on the COTS devices, the suitable method to modulate the emitting signals that can be implemented on while having the potential of achieving high resolutions for serving our purpose has to be determined, which is important while challenging.
- The received signals are weak and mix the superposition of signals reflected from environment objects and multipath. How to locate the signals of interest (i.e. direct reflection) from the human body is a challenging problem, requiring us to develop a collection of techniques for effectively mitigating the interference from environment objects and multipath reflection.

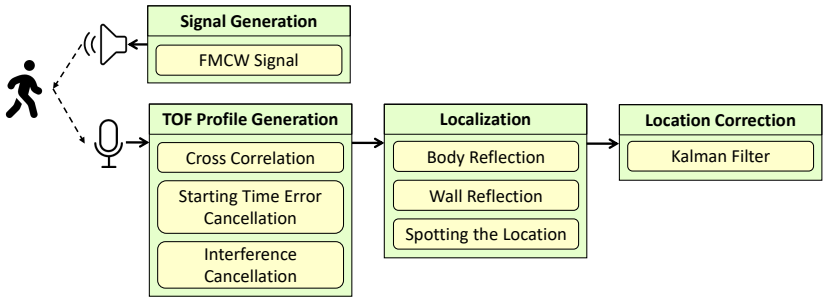

Fig. 1. The workflow of EchoSpot.

- The target ToF can only help identify the distance of a human to the device rather than the position. In general, multiple pairs of speakers and microphones are required to pinpoint the human position by finding the intersection points of their respective ellipse curves after getting the distances. However, EchoSpot is expected to only employ one speaker and one microphone, which is more general in home devices. It remains challenging to determine the position information in the two-dimensional space.

3.1 System Overview

Our design of EchoSpot consists of four components: *Signal Generation*, *ToF Profile Generation*, *Localization*, and *Location Correction*, as shown in Figure 1. In the *Signal Generation* module, EchoSpot programs the speaker to emit inaudible acoustic signals at 18kHz~23kHz. Specifically, it modulates the FMCW signal and periodically sends FMCW chirps while using the microphone to record the reflected FMCW signals. The sampling frequency of the speaker is set to 48kHz. According to the Nyquist Sampling Theorem [43], the reflected signal can be entirely reconstructed from the recorded signals.

The signal is then sent to the *ToF Profile Generation* module to generate the ToF profile for further analysis. This module contains three components, i.e., *Cross Correlation*, *Starting Time Error Cancellation*, and *Interference Cancellation*. Particularly, in the first component, the cross-correlation is applied on the received signal and the transmitted signal, creating the raw TOF profiles. We next identify the starting point of the profile and eliminate the starting time error caused by device imperfection. As the TOF profiles contain peaks of static reflections, we cancel these peaks in the *Interference Cancellation* module, with the direct reflection remained.

In the *Localization* module, we build a location model that utilizes the distances of direct reflection and of wall reflection to calculate the location. We pick up the peaks of the human reflection and wall reflection from the residual TOF profile after the interference cancellation aforementioned, and roughly estimate the distance to be fed into the location model. Finally, in the *Location Correction* module, we apply Kalman filter to further eliminate the potential measurement errors caused by multipath effect, system defects, etc., to improve the accuracy of location estimation.

3.2 Signal Generation

In *Signal Generation module*, EchoSpot controls the speaker to generate the FMCW (Frequency Modulated Continuous Wave) [3, 46] signals for sensing. In particular, it periodically transmits chirps and within the duration of a chirp, the operating frequency keeps changing from $f_{\min}$ to $f_{\max}$. So, for each chirp with the duration of time T, the frequency of signals can be expressed as

$$f(t) = f_{\min} + \frac{Bt}{T}, \tag{1}$$

where B is the signal bandwidth, defined as $B = f_{\max} - f_{\min}$. The phase of the transmitted FMCW signal can be expressed as the integration of $f(t)$ over time, i.e.,

$$\lambda(t) = 2\pi \int_0^t f(t)\, dt = 2\pi(f_{\min} t + B\frac{t^2}{2T}). \tag{2}$$

Then, the FMCW signal could be expressed as $cos(\lambda(t))$.

After the speaker transmits the FMCW signal $cos(\lambda(t))$, the microphone will receive a reflected signal $\cos(\lambda(t-\tau))$, which can be considered as a time shifted version of the transmitted one with a delay of τ. Here, τ is called as the time-of-flight (ToF), which is the time from when the signal is generated at the speaker to when the signal reflected from the object is received at the microphone. In FMCW signal, according to the characteristic of frequency change, τ can also be measured as follows:

$$\tau = \frac{\Delta f}{\left(\frac{\delta(f)}{\delta(t)}\right)}, \tag{3}$$

where Δf represents the frequency shift and $\delta f/\delta t$ denotes the frequency shift per unit of time. Here, $\delta f/\delta t = B/T$ in each time period. The distance R for the object that causes the direct reflection can then be calculated by

$$R = \frac{c\tau}{2} = \frac{c\Delta f}{2\left(\frac{\delta(f)}{\delta(t)}\right)}, \tag{4}$$

where c is the speed of sound.

Since the duration of each FMCW chirp is T, the minimum frequency resolution of the FFT is $1/T$. The range resolution δR, which shows the ability of FMCW signal to capture the minimum movement, is determined by the frequency resolution. Then, we have

$$\delta R = \frac{c/T}{2\left(\frac{\delta(f)}{\delta(t)}\right)} = \frac{c}{2B}. \tag{5}$$

Obviously, the range resolution is determined by the bandwidth B.

Considering that the commercial microphones can only record the signals below the 24kHz and the majority of background noises (such as human conversation, music, FM radio wave, etc.) have frequencies up to 14kHz, EchoSpot assigns the frequency of chirp ranging from $f_{\min}$=18kHz to $f_{\max}$ =23kHz, with the bandwidth of 5kHz. As such, the range resolution can reach $3.4cm$ according to Eqn. (5), giving the sound speed c of $340m/s$.

Typically, the longer the duration time T (chirp length), the more the overlapped parts among reflected signals, which brings difficulty in differentiating them. However, the long duration would help the system to find the big reflector in the environment. On the other hand, [42] has shown that a shorter duration of FMCW can lead to higher tracking accuracy since it will result in a smaller Doppler shift caused by the movement. But considering the limits of home devices, if the duration time is too short, the sound energy becomes too weak so that the reflected signals may not be detected due to the low SNR. Hence, choosing a duration time is very important to help identify the echo reflected from the body. We have conducted extensive experiments in different environments, aiming to identify an appropriate chirp duration T, which works in the home environment while minimizing the overlapping among reflected signals. We found that if the duration is less than $0.02s$, less overlapping is observed; if it is longer than $0.005s$, the target signal is stronger to be detected. As such, we set the duration of T as $0.01s$, resulting in strong enough signal for detecting the target while having less overlapping from the reflections.

For the time interval between adjacent chirps, a short one would result in a high resolution of localization. However, the received signals may be severely affected by multipath effect from the previous chirp. On the other hand, considering a house environment, the maximum distance from a person to the device is around 7 meters. Theoretically, the time interval should be larger than $\frac{14\text{m}}{340\text{m/s}} \approx 41.2\text{ms}$, for sufficiently receiving the reflection

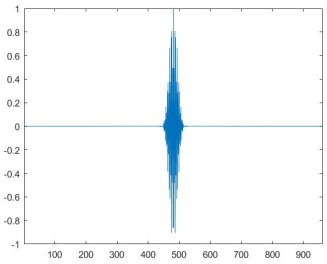

Fig. 2. Autcorrelation of the windowed signal, with less sidelobes.

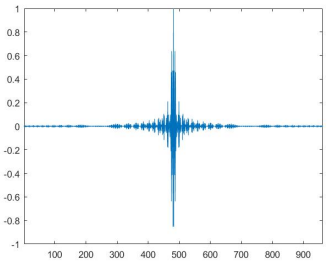

Fig. 3. Autcorrelation of the unwindowed signal, with many sidelobes.

signals from the body. In EchoSpot, we set the time interval to be 200ms to ensure that it is sufficient to receive the reflection signals while the multipath effect from one chirp will not impact the next chirp.

FMCW signal windowing. Inspired by [13], we then apply the Hanning window to reshape the FMCW signal envelopes to increase the signal-to-noise ratio (SNR) by improving the peaks to the sidelobe ratio. The windowed signal would have fewer sidelobes comparing to the raw signal without applying the window function. We conduct an experiment to validate this point. Figure 2 and Figure 3 exhibit the autocorrelation of the windowed signals and the raw signals, respectively. It is obvious that the autocorrelation from the windowed signal in Figure 2 becomes stronger and has fewer sidelobes, resulting in a higher SNR. This phenomenon would be similar when we perform the cross-correlation between the windowed signal and the reflected signal. Fewer sidelobes would result in a higher SNR, thus improving the accuracy.

To conclude, EchoSpot will control the speaker to periodically send FMCW wave at the frequency of 18kHz~23kHz with the time duration of 40ms for each chirp, the bandwidth of 5kHz, and the time interval of 200ms between two chirps.

3.3 TOF Profile Generation

After receiving the reflected signals, EchoSpot will measure the time of flight (ToF) of the respective signals, which will be further used to calculate the distances. Since FMCW signals range from 18kHz to 23kHz, we use a band-pass filter to extract the useful signals. We will first generate the time-of-flight (ToF) profile and then develop a series of solutions to remove the potential measurement errors from the system starting time error and environmental interference.

ToF Profile. We pick up a series of N points from the transmitted signals and the received signals. Denote $v_{tx}(m)$ and $v_{rx}(m)$ as the transmitted and received signals, respectively, at a point m. The normalized cross-correlation is then applied to measure the similarity of transmitted signals and its n-sample shifted version of received ones, expressed as follows:

$$P(n) = \frac{\frac{1}{N}\sum_{m=0}^{N}[v_{rx}(m) - \bar{v}_{rx}][v_{tx}(m-n) - \bar{v}_{tx}]}{\left\{\sum_{m=0}^{N}[v_{rx}(m) - \bar{v}_{rx}]^2 \sum_{m=0}^{N}[v_{tx}(m-n) - \bar{v}_{tx}]^2\right\}^{0.5}}, \tag{6}$$

where $\bar{v}_{tx}$ and $\bar{v}_{rx}$ indicate the average values of the transmitted and received signals, respectively, over the N points. By picking up the corresponding lag of the peak on $P(n)$, we can transform it into the ToF value, so as to further calculate the distance value of the respective reflection object, i.e.,

$$\tau = \frac{n}{F_s}, \tag{7}$$

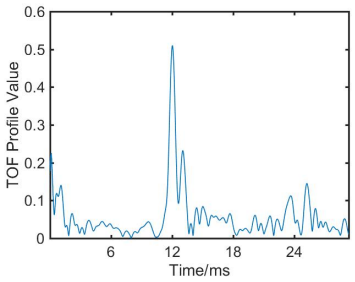

Fig. 4. TOF profile after eliminating the time error.

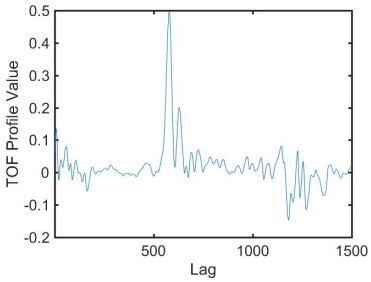

Fig. 5. Denoised TOF profile, where strong peaks represents the human reflection. Some other multipath reflections from the human also remain.

where F_s is the sampling frequency, which is set as 48kHz in EchoSpot.

Figure 6 shows an example of the $P(n)$. After getting the $P(n)$, we apply Hilbert Transform [16] to calculate the envelope $E(n)$ of $P(n)$, as shown in Figure 7. where the x-axis represents the TOF, and the y-axis represents the cross-correlation value which measures the similarity of the transmitted and the received signals. In the remaining of this paper, $E(n)$ will be called as the ToF profile. Since we emit the FMCW signal every 200ms, each ToF profile $E(n)$ will be generated for this period.

From this step, we can roughly calculate the ToF profile, expressed by $E(n)$. However, the unsychronization of the speaker and microphone along with the environmental interference will significantly affect the measurement of ToF, making the ToF profile inaccurate.

Starting time error. Since the speaker and the microphone are co-located, the direct transmission time can be ignored. However, due to the unsynchronization between speaker and microphone (i.e., device imperfection), there is a certain delay for the signals to be received by the microphone. This delay is called the start time error. Since the direct transmission signal is directly received without reflection, it is stronger than all other reflected signals, allowing us to identify it by selecting the lag with the largest peak value. We denote this point as the start time point, whose corresponding time of flight indicates the start time error. This point has to be correctly identified so as to accurately perform the calculations in remaining steps. As shown in Figure 7, the start time error is $0.107s$. By removing the peak value before the start time point in the Figure 7, we get the TOF profile without the start time error, as shown in Figure 4. In Figure 4 the lag values directly correspond to the time of flight of the reflection signals, so we could use it to calculate the time of flight.

Interference Cancellation The environment object will also cause reflection, which thus generate a set of peaks in ToF profile, misleading our selection of these reflected signals from the human body. In Figure 4, we observe two main peaks, however, not all of them are from body reflection. We will show how to remove the peaks that do not correspond to the body reflections so as to mitigate the interference from the environment reflection.

Since the positions of objects in a room are fixed, their respective ToF files are relatively same within a certain time. Hence, EchoSpot records 10 ToF profiles in the static environment without human movement and calculates the averaged correlation value for each corresponding lag within these 10 ToF profiles. We use a vector $\bar{\mathbf{p}}$ to indicate a series of averaged values corresponding to all lag points. By subtracting $\bar{\mathbf{p}}$ from the peak values of ToF profile $E(n)$, we can eliminate those peaks corresponding to the environmental interference, resulting in a denoised ToF profile (Figure 5) with a vector of peak values denoted as $\mathbf{E}_d$. Comparing to Figure 4, we could see the second peak is totally removed, with only one peak remaining which is the body reflection peak.

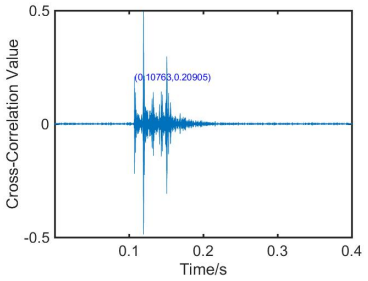

Fig. 6. Example of cross correlation. 0.10763 and 0.20905 indicate the start time error and its corresponding peak value.

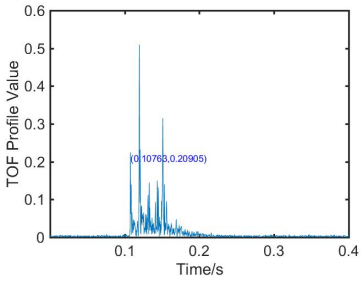

Fig. 7. Example of TOF profile, where the peaks refer to the reflections in the environment.

In the real-world deployment, it is not necessary to let the speaker keep sending the higher power FMCW signals for sensing. Considering that the environment state typically is static, EchoSpot can periodically transmit high power FMCW signal for capturing the environment ToF in each 30 minutes for updating. Meanwhile, EchoSpot can periodically (say 10s) transmit the FMCW signals at a low rate and low signal power (50% of its working power) to sense the environment for detecting the appearance of a person. Once a person is detected from the low power signals, it will be immediately triggered to transmit the higher power signal at a higher rate to start working. Then, EchoSpot can perform the subtraction of the environments from the current ToF profiles, which can result in the pure ToF profiles used for localization. When the person leaves the room, the reflection signals will become weaker, so EchoSpot will resume to the low power state.

3.4 Localization

After preprocessing in Section 3.3, we next show how to choose the correct peaks on the residual ToF profile for calculating the distances, which can be eventually used for localization. Since our solution will rely on only one speaker and one microphone, it is not sufficient to use only the ToF from a human for localization. Here, the ToF from wall reflection will assist us to spot a person's location. In what follows, we will first show how to spot the location of a human with the assistance of wall and then illustrate how to choose the ToFs corresponding to the human reflection and the wall reflection.

Localization via One Speaker and One Microphone. We will show how to model the location of a person when just using one speaker and microphone. We assume the distance between the wall and the device are known, which can be obtained through the following way: letting the speaker periodically send acoustic signals and the microphone receive the direct reflection, then we measure the time-of-flight from the reflected signal for calculating such a distance. As shown in Figure 8, the distance between the wall and the device is known, denoted as w. R and S represent the person and the device, respectively. We draw a point S_v in the opposite side of the wall with a distance of w. In this figure, (x, y) represents the relative distance of the person to the microphone, which we aim to determine. The position of the speaker is assumed to be known, then we could model the position of (x, y) as follows:

$$\begin{cases} x^2 + y^2 = d^2\,, \\ d + d_v = D\,, \\ (2w - x)^2 + y^2 = d_v^2\,, \end{cases} \tag{8}$$

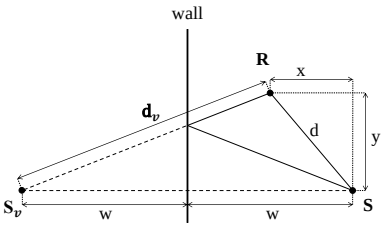

Fig. 8. Location model, R indicate the people and S indicate the speaker.

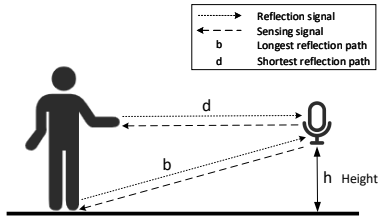

Fig. 9. Different reflection path when the signal reflect from the body, where d is the shortest path, b is the longest path, and h is the height of speaker.

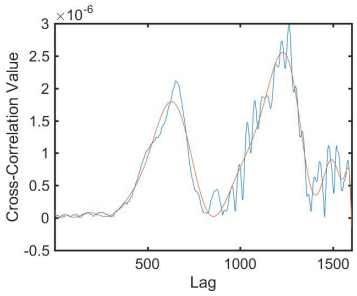

Fig. 10. Reflection from body and wall, the second peak is the reflection from the wall.

where d is the distance between the device and the person, d_v is the distance from the person to S_v. D represents the distance from the device S to the target R and then to S_v. This formula can be solved with solution of:

$$\begin{aligned} x &= \frac{-D^2+2Dd+4w^2}{4w}, \\ y &= \sqrt{d^2 - x^2}. \end{aligned} \tag{9}$$

Hence, x and y can be seen as the functions of D and d. Notably, d can be calculated by the ToF corresponding to the body reflection and D can be calculated by that corresponding to the wall reflection through the human body.

Identification of ToFs for Body Reflection. Through interference cancellation in Section 3.3, the interference of reflected signals from the static object has been eliminated. As shown in figure 5, we could see a strong peak that corresponds to the human reflection. Hence, we can directly pick up the strongest peak $\mathbf{E}_d$ from figure 5, and treat the associated signal as being from the body reflection. The corresponding lag value, assuming l_1, can be used to calculate the time delay τ_1 and the distance d, i.e.,

$$\tau_1 = \frac{l_1}{F_s}, \; d = c\frac{\tau_1}{2} \tag{10}$$

However, the residual multipath reflection will travel a longer distance and have a smaller peak value on ToF profile. From the residual ToF profile on figure 5, it is hard to directly pick up the peak from the wall reflection.

Identification of ToFs for Wall Reflection. We have two observations that could help us pick up the exact location of the wall reflection. The first observation is that the wall reflection signals are strong, which may be due to the large size of the wall, where different multipath signals aggregate together on the wall, resulting in a strong reflection. That means the signal reflected from the body then to the wall is detectable. The second observation is there would be a similarity between the wall reflection and body reflection. When we apply the cross-correlation, we observe a relatively small peak on the TOF profile at the wall reflection distance. The small peak means the wall reflection is similar to the original signal. Also, the strong peak on the body reflection means the body reflection is similar to the original signal. Since both the wall reflection and body reflection are similar to the original FMCW signal, this further implies there would be a similarity between the wall reflection and body reflection.

Based on the two observations, we propose a new solution, which can help us identify the reflection distance from the wall passed through the human body. We consider the human body as a virtual sender and the reflected signals from human body to the wall as the virtual sending signals. Due to the similarity between the wall reflection and body reflection signal, we could apply the cross correlation to identify the wall reflection signal.

Figure 9 shows how the signal is reflected from the body. Typically there are several paths. In the figure, d is the shortest reflection path from the body to microphone in the horizon direction. b is the longest reflection path which is from feet to the microphone. The body reflection signal length can be calculated as *original signal length + (TOF of the longest path - TOF of the shortest path)*. Since the body reflection signals are too long, which shall contain the reflection from other objects in the environment, we cannot directly take them as the virtual signals. Instead, we only take signals at the time interval between the longest and short reflection as our virtual signals, to ensure that the majority are reflected from the body.

As shown in Figure 9, the start time is determined by the shortest signal path, i.e., the distance from the microphone to the human d, which has been calculated in the last step. The possible longest path can be considered as the reflection from the feet, with the distance denoted as $b = \sqrt{d^2 + h^2}$. If we get the height of the speaker h, we could get the length of the longest path. Then, the delay of this reflection comparing to the shortest path is $\hat{t} = 2(b - d)/c$. We set the duration to be around $\hat{t}$ for the virtual sending signals.

We take the FWCM signal with the duration of $\hat{t}$ and apply the cross-correlation on the received signals at the microphone via Eqn. (6) to get a new ToF profile, as shown in Figure 10. Two sets of strong peaks appear on such TOF profile. The first peak cluster can be considered as the direct transmission from the virtual sender (i.e., the reflection from the human body), while the second cluster can be considered as the nearby wall reflection. We apply the polynomial fit on the peak to estimate the body and wall reflection location. As the figure shows, the orange curve is the polynomial fit curve. We take the second strongest peak on polynomial fit curve as the wall reflection and assume its lag value as l_2. Then, the time delay and distance (i.e., D) from the body-wall-device can be calculated as follows:

$$\tau_2 = \frac{l_2}{F_s}, D = c\tau_2 \tag{11}$$

After having the values of d and D, we can apply them to Eqn. (9) and get the position (x, y) for the person R.

Notably, this solution relies on the location of wall. However, in practice, there may be two walls in the opposite position of the person. Our solution still works as follows. Assume the distances from the device to the two walls are w_1 and w_2, respectively. Through the correlation, there will be three strongest peaks in the new ToF profile. While the first one is still the direct transmission from the virtual sender to the device, for the following two, it is necessary to decide their correspondences to w_1 and w_2. We can assume the first one and the second one are corresponding to w_1 and w_2, respectively, and use the aforementioned method to calculate two respective positions. If the two calculated positions coincide, then we locate the person; otherwise, the first one and the second one correspond to w_2 and w_1, respectively.

3.5 Correcting the Location

Ideally, from aforementioned steps, we should have spotted the location (x, y) of a person. However, there may be some potential errors due to the wrong selection of peak values from the residual multipath reflection. Considering the characteristic of a moving person, a series of his relative positions at different time points can be traced to further correct our calculation. Here, we leverage the Kalman filter [57] by taking into account two consecutive measurements for correcting the current location. In Kalman filter, we consider two consecutive measurements and use the previous calculated location to predict the one at the next time point, which can be modeled as follows:

$$\hat{\boldsymbol{x}}_t = \boldsymbol{F}\hat{\boldsymbol{x}}_{t-1} + \boldsymbol{B}a_t, \tag{12}$$

where $\hat{\boldsymbol{x}}_t = [\hat{p}_t, \hat{v}_t]^T$ and $\hat{p}_t$ represents the predicted (x, y) and $\hat{v}_t$ indicates the predicated velocity at time t. $\hat{x}_{t-1}$ represents the prediction at the time point $t - 1$. a_t is the acceleration speed of the moving person. We assume the motion of a human contains both the acceleration and deceleration phases, so acceleration shall follow a Gaussian distribution with the mean of 0. [23] suggests the maximum acceleration speed of a walking person is 0.2g to 0.3g. As the human would have a relatively low speed in indoor environment, we choose 0.6 as the

variance σ_a roughly. Then the a_t could be represented as $a_t \sim \mathcal{N}\left(0, \sigma_a^2\right)$. $\boldsymbol{F}$ is the transmission matrix and $\boldsymbol{B}$ is the control matrix, which can be expressed as:

$$\boldsymbol{F} = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix}, \boldsymbol{B} = \begin{bmatrix} \frac{\Delta t^2}{2} \\ \Delta t \end{bmatrix}, \tag{13}$$

where Δt is the elapsed time between two consecutive measurements. Denoting $\boldsymbol{P}_t$ as the prediction covariance matrix at time t and $\boldsymbol{Q}$ as the covariance matrix of the acceleration a_t, we have:

$$\boldsymbol{P}_t = \boldsymbol{F}\boldsymbol{P}_{t-1}\boldsymbol{F}^T + \boldsymbol{Q}. \tag{14}$$

Notably, $\boldsymbol{P}_t$ is determined by a_t. According to Kalman filter, we have

$$\begin{aligned} \boldsymbol{z_t} &= \boldsymbol{H}\boldsymbol{x}_t + \boldsymbol{s}, \\ \boldsymbol{R}_t &= \boldsymbol{H}\boldsymbol{P}_t\boldsymbol{H}^T, \end{aligned} \tag{15}$$

where $\boldsymbol{z_t} = (x, y)$, calculated in Section 3.4, and $\boldsymbol{H}$ is defined as $\boldsymbol{H} = \begin{pmatrix} 1 & 0 \end{pmatrix}$, as we do not aim to calculate the velocity. $\boldsymbol{x}_t$ is defined as the actual location which is the hidden state. $\boldsymbol{s}$ is the observation noise of the system that could not be directly measured. $\boldsymbol{R}_t$ is the covariance matrix corresponding to (x, y).

Since $\boldsymbol{z_t}$ is known, we could update the estimation by

$$\hat{\boldsymbol{x}}_t' = \hat{\boldsymbol{x}}_t + \boldsymbol{K}'\left(\boldsymbol{z}_t - \boldsymbol{H}\hat{\boldsymbol{x}}_t\right), \tag{16}$$

where $\hat{\boldsymbol{x}}_t'$ is denoted as the updated estimation of the x, y. $\boldsymbol{K}'$ can be expressed as $\boldsymbol{K}' = \boldsymbol{P}_t\boldsymbol{H}^T\left(\boldsymbol{H}\boldsymbol{P}_t\boldsymbol{H}^T + \boldsymbol{R}_t\right)^{-1}$. Through this way, we can get a better estimation of the current location. This process can be continued so that we can have a series of new estimation of a person location at different time.

4 PERFORMANCE EVALUATION

In this section, we implement the EchoSpot system and deploy it in the real home environment for experiments. A set of experiments has been conducted and analyzed to show the performance of the EchoSpot in terms of localization accuracy for both static and moving person at different positions.

4.1 Experiments

We use one pair of commodity speaker (Edifier R1280DB) and microphone (SAMSON MeteorMic, 16 bit, 48 KHz), and bind them together to work as our experimental devices. The output power of the speaker is set to around 80% of the speaker's maximum power. We measure the sound pressure level at 1 meter from the speaker, which is 45dB. The speaker is programmed as the signal transceiver to transmit FMCW signals with the carrier frequency sweeping from 18kHz to 23kHz. The microphone works as the signal receiver to record the reflected signals at a 48kHz sampling rate. It is connected to a laptop and uploads the reflected signals to this laptop for processing. Our localization algorithm is implemented on this laptop with Matlab.

For performance validation, we mark some reference points and draw one trajectory with known positions on the floor to serve as the ground truth. They will be used to measure the performance of EchoSpot in terms of spotting the static locations and continuous locations, respectively. For static localization, we ask the target person to stand on each reference point for 10s. For continuous localization, the target person will be asked to walk along this predefined trajectory. We consider the following performance measurement metrics:

- **Localization Error** is the distance between the measured position and the ground truth reference point.
- **Trajectory Error** is the vertical distance between the walking trajectory and ground truth trajectory.

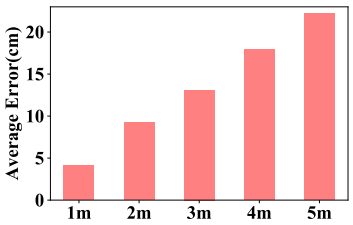

Fig. 11. Averaged localization errors at $1m$, $2m$, $3m$, $4m$ and $5m$.

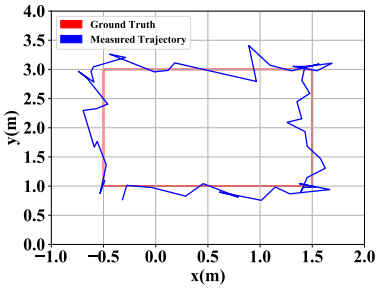

Fig. 12. Tracking the moving trajectory of a person, where the red lines indicate the predefined rectangle trajectory, serving as the ground truth, while the irregular blue curve represents the measured trajectory by our EchoSpot.

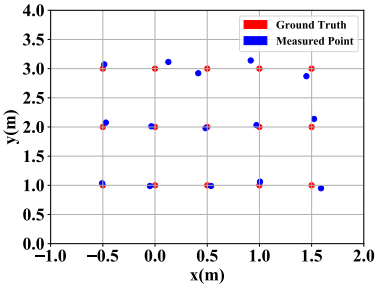

Fig. 13. Locating a static person at different reference points, where the red points indicate the reference points and the blue points represent the spotted location by our EchoSpot.

4.2 Performance of Localization

We place the speaker to be $100cm$ away from the wall and $110cm$ height above the ground in a room.

Locating a static person. We consider the distance ranges of $1m$, $2m$, $3m$, $4m$ and $5m$ away from the speaker, and at each distance range, we take 10 different reference points. We ask a person to stand at the 10 different reference points corresponding to each distance range and use EchoSpot to spot his position. At each reference point, this person will stand $10s$, so we can collect 50 location data at this point, giving that EchoSpot sends one chirp within each $200ms$. We calculate these 50 positions and average them to be used as the located position of this person. Figure 11 shows the averaged localization error over the 10 reference points at each distance. Specifically, EchoSpot achieves the averaged localization errors of $4.1cm$, $9.2cm$, $13.1cm$, $17.9cm$, $22.2cm$, respectively, at the distance of $1m$, $2m$, $3m$, $4m$ and $5m$. The median error is $12.4cm$, which is comparable to the state-of-the-art Wi-Fi and mmWave-based localization systems [21, 60] with median error over $23cm$. We observe that the localization error grows with the increase of distance range. This is due to the signal attenuation: the longer range will result in weak peak, making the peak selection more difficult.

Locating a moving person. We continue to show the performance of EchoSpot for localizing a moving person. We ask this person to walk in his natural speed and pattern along a predefined rectangle trajectory, indicated by the red line in Figure 12. The microphone and speaker are placed at the coordinate of $(0, 0)$ and the wall is at $x = -0.5$. The irregular blue curve represents the measured trajectory from EchoSpot by spotting a series of positions of the walking person. In this measured trajectory, the averaged trajectory error is $21.9cm$. As shown in the figure, when the person is far away from the wall, the trajectory error will become large. The reason is that the wall reflection would be weaker than that from other objects, which may result in wrong peak selection corresponding to the wall reflection. However, we also observe when the person walks along the wall, the trajectory error is also relatively large. The reason is that when a person walks along the wall, the error is caused by the arm or leg swing, resulting in the strongest peak appearing at the wrong place, leading to wrong peak selection and affecting the distance estimation.

We further compare its performance to the case that the static person stands on a series of reference points on the rectangle trajectory. In Figure 13, the red points indicate the reference points on the rectangle trajectory, and the blue points represent the spotted locations from the EchoSpot. Figure 14 shows the CDF of the trajectory

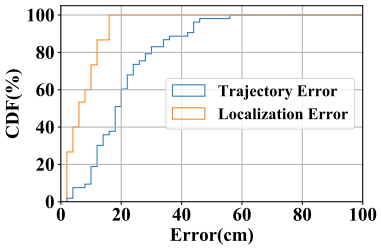

Fig. 14. CDF for Trajectory Error and Localization Error.

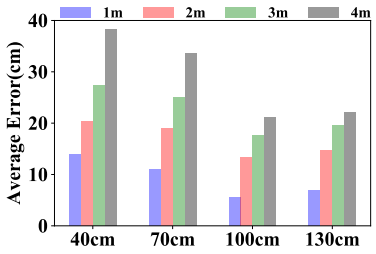

Fig. 15. Impact of the speaker separation.

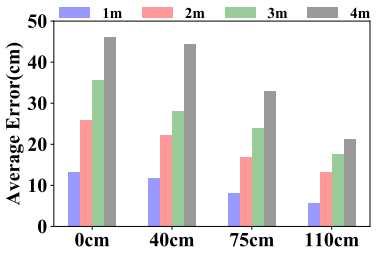

Fig. 16. Impact of the device heights.

error and the localization error, corresponding to Figure 12 and Figure 13, respectively. From this figure, we can see the overall performance of EchoSpot for tracking a moving person is promising, even though it performs a little worse than that in the static status. Specifically, the median errors are 19.8cm and 8.5cm, respectively, in the walking and static status. This is due to the fact that, when walking, localization data collected by EchoSpot at each point is much less than that at the static status. Thus, the calculation error from one location data will significantly impact the performance. However, in the static status, EchoSpot can collect 50 location data and average them, mitigating the errors from some location data. On the other hand, the movement of the leg and the arm will cause strong reflection, which will lead to the wrong peak selection for wall reflection.

4.3 Impact of Device Placement

We next conduct experiments to show the impact of device placement on the performance of EchoSpot. In a home environment, the devices may be placed differently, such as on the floor, on the coffee table or on a desk, with different distances to the wall and having different heights. We will vary the separation distance between the wall and the speaker, and the placement height to show EchoSpot's localization errors.

First, we place the speaker at a 100cm desk and examine the impact of the distance between wall and speaker. We let a person stand on four reference points on the straight line with 1m, 2m, 3m, and 4m away from the speaker. At each point, the person will stay for 10s, so that EchoSpot will get 50 position data and average them. Figure 15 plots the localization errors with different separation distance between the speaker and the wall. From this figure, we can see the localization errors decrease first and then grow up at each distance range (i.e., 1m, 2m, 3m, and 4m). The lowest localization errors are all achieved at the separation distance of 100cm. We notice that when the separation distance is small, the error is relatively high. The reason is that, according to Eqn. (9), a small distance (i.e., w) between wall and speaker will enlarge the distance estimation error of D, which thus will result in a large x. On the other hand, we observe that the localization error will slightly increase when the separation distance is more than 100cm. The reason is that when the separation distance increases, the signal path with respect to the wall reflection will become longer, leading to more attenuation of the reflection signals. The peak value from the wall reflection will become smaller, making the peak selection incorrect.

Next, we use the same setting as the above experiment, except for fixing the distance between the wall and the speaker as 100cm and varying the height of the speaker from 0cm to 100cm. Figure 16 shows the localization errors at the distance ranges of 1m, 2m, 3m, and 4m with respect to different heights of the speaker. From this figure, we can see the localization errors drop with the increase of height. Specifically, when placing the speaker on the ground, the shortest reflection is from the leg or feet. From Section 3.4, we know the duration of the virtual signal is determined by the height of person or height of the device. When the device's height is larger than the half of the person's height, the duration is determined by the height of device and could be calculated through

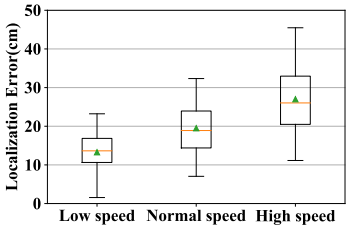

Fig. 17. Impact of the moving speed.

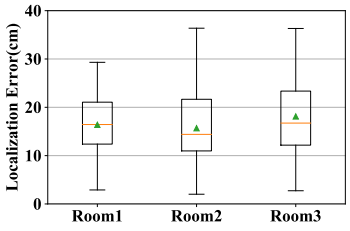

Fig. 18. Impact of different environments.

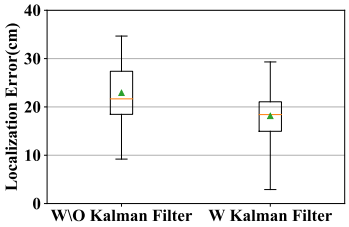

Fig. 19. Impact of the Kalman filter.

the device's height and shortest path. Otherwise, we could not accurately measure the duration, leading to the wrong selection of wall reflection.

4.4 Impact of Moving Speed

For the moving person setting, we further examine the impact of the walking speed to the localization accuracy of EchoSpot. We consider a person walking at a very low speed, at his normal speed, and at a high speed, respectively, toward the speaker from $4m$ to $1m$. Figure 17 shows the quartiles figure under three walking speeds. The green point is the mean error and the orange horizontal line is the median error. The maximum point represents 90th percentile error, and the minimum point represents the minimum error. The upper quartile represents 75th percentile error, and the lower quartile represents 25th percentile error. From this figure, we can see EchoSpot has the averaged error of $13.1cm$ at a very low speed, which increases to $19.2cm$ and $26.6cm$ at the normal and high speeds, respectively. The median errors are $13.8cm$, $18.5cm$, $25.7cm$ and the 75th percentile errors are $17.1cm$, $23.2cm$, $32.8cm$, corresponding to the three walking speeds. Obviously, the error of EchoSpot increases with the high walking speed, because when a person moves faster, the arm swing or leg swing will have more impact on the peak selection.

4.5 Performance at Different Room Layout

We also deploy EchoSpot in three rooms with different layouts to show its performance. In all three rooms, the device is placed at $110cm$ height and at $100cm$ away from the wall. A person moves from $4m$ to $1m$ toward the speaker. Figure 18 shows the localization errors at each room. At three rooms, the mean errors are $15.5cm$, $14.8cm$, $18.9cm$, the median errors are $15.5cm$, $13.8cm$, $15.7cm$, and the 75th percentile errors are $21.5cm$, $21.9cm$, $23.2cm$, respectively, without significant change. This set of experiments demonstrate the robustness of EchoSpot in different layout environment.

4.6 Impact of Different Devices

We evaluate the Echospot on various speakers, represented by Speaker 1 (Edifier R1280DB), Speaker 2 (Logitech z200), and Speaker 3 (Amazon Echo), to show that Echospot does not rely on specific hardware. We tune the volume of these speakers so that they have similar transmission power. The Amazon Echo is connected via Bluetooth pairing, which may occur a certain latency. But notably, our Starting Error Cancellation module in Section 3 could eliminate such a latency. For each speaker, we use the same frequency, i.e., $18 - 23khz$. We draw the CDFs of each device as shown in Figure 24. The median errors for speaker 1, speaker 2, and speaker 3 are $16.5cm$, $17.9cm$, and $20.4cm$, respectively. We could observe a similar trend on these CDFs. Notably, our system on Amazon Echo achieves slightly worse performance than that on Edifier R1280DB and Logitech z200. The reason

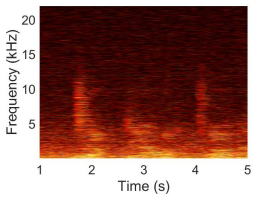

Fig. 20. Spectrum in the human talking environment.

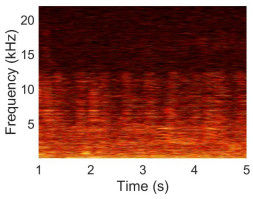

Fig. 21. Spectrum in the environment playing normal music.

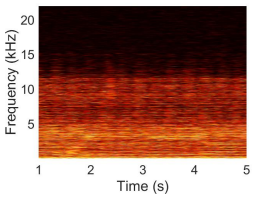

Fig. 22. Spectrum in the environment playing rock music.

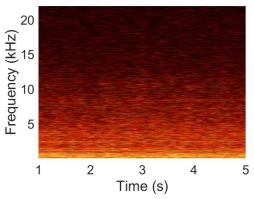

Fig. 23. Spectrum in the environment washing the utensils.

is that Amazon Echo generates the ultrasound to all directions, so the received signals will mix the reflection signals from different directions. This results in a relatively low SNR for the target signals, deteriorating the localization accuracy. Also, we could see the maximum tracking error is about 40cm. This is close to the width of the human body, so such an error is acceptable when tracking a person. The similar CDFs on different devices implies that our system is robust against device heterogeneity.

4.7 Impact of Noise

We continue to conduct experiments to evaluate the performance of Echospot under different types of noise: human talking and playing music. The source of noise is 0.5m away from the system. Figure 25 shows the errors under the silent, human talking and playing music environments. From the figure, we observe EchoSpot performs similarly in the three environments, having the 75th percentile errors of 21.5cm, 23.4cm, 22.6cm, respectively. The mean errors are 18.1cm, 19.5cm, 18.9cm, respectively. The median errors are 18.5cm, 19.2cm, 18.7cm, respectively. Thus, we can conclude that these audible noises can only slightly affect EchoSpot. The reason is that EchoSpot's working frequency is higher than the noise level. To validate this point, we generate the spectrum under the scenarios of human talking and playing music when no human exists. Figures 20 and 21 show the respective spectrum, which clearly show that all generated noises are much lower than 20khz.

We continue to consider more real scenarios when playing the rock music and washing the utensils in the apartment. Figures 22 and 23 show their respective spectrum. From the two figures, we could see sounds generated from these noise sources are still much lower than our working frequency, which shall have no overlapping with our Doppler shift frequency and can be easily subtracted via EchoSpot. This further demonstrates that our EchoSpot is robust to the environmental noises.

4.8 Impact of Location Correction Module

We conduct experiments to validate the importance of *Location Correction (Kalman filter)* module in our system by comparing EchoSpot's performance with and without this module. We ask a person to walk at his normal speeds towards the speaker from 4m to 1m. Figure 19 compares the localization errors of EchoSpot without and with the Kalman filter. From this figure, we can observe the 75th percentile errors of the two cases are 27.3cm and 21.5cm, respectively. The mean errors are 22.9cm and 18.1cm, respectively. The median errors are 21.6cm and 18.4cm, respectively. Such performance results indicate that this module is important to improve our system performance.

5 DISCUSSIONS

EchoSpot aims to demonstrate the feasibility of only using one speaker and one microphone to perform acoustic sensing for precise human localization, with the effective working range to be several meters. Notably, it is

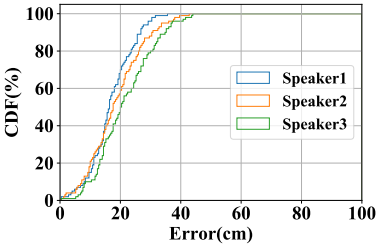

Fig. 24. Impact of the device.

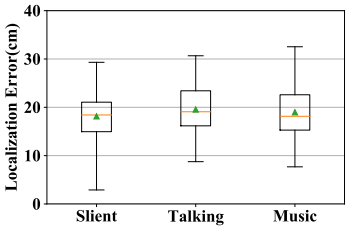

Fig. 25. Impact of the noise.

not necessary to keep EchoSpot always on. In practical deployment, we can let it continuously play the low energy ultrasound sound to sense the TOF profile of the reflected signals and then compare the ToF profile to the prerecorded environment ToF profile. Only once the difference between them reaches to a certain level, indicating that a moving object is detected, EchoSpot is triggered to generate the normal power ultrasound signals.

The current design of EchoSpot still has some limitations in the practical deployment. We briefly discuss some of them and leave the exploration in our future work. First, EchoSpot is sensitive to the complex indoor environment, whereas the multi-path reflection from different furniture may significantly degrade the localization performance, requiring to have the strong interference cancellation technologies for processing. Besides, the device placement also plays a key factor to impact EchoSpot's performance. As shown in Section 4.3, when the separation distance between EchoSpot and the wall is large, or the height of EchoSpot is low, the performance of EchoSpot will drop. To be robust to device placement, one plausible solution is to modulate the signals to contain the correlation after a long propagation, for example, generating OFDM symbol as the transmitted signal. This allows us to keep detecting the clear wall reflection, even when there is a large separation between the wall and EchoSpot. Another solution is to increase the signal's volume to enhance the signal strength. To eliminate the influence of device placement height, we can get a human's height information and obtain the reflection from head for better estimating the duration of the virtual sending signals.

Second, the performance of EchoSpot is affected by the target speed. As shown in Section 4.4, when the target has a high moving speed, the performance of EchoSpot will drop. To address this issue, one plausible solution is to send more FMCW signals per second so as to obtain more sample points of the walking trajectory, which can help reduce the calculation error to a certain extent. However, a short time interval between consecutive FMCW signals would increase the influence of multipath effect, i.e., the multipath effect from the previous FMCW signal would severely affect the ToF profile corresponding to the current FMCW signal. This requires us to develop a new interference cancellation method to eliminate multipath interference.

Third, the current design of EchoSpot can only locate one person each time, while the multi-person localization remains open and challenging. If there are multiple people in the environment, it is challenging to pinpoint the reflection peak corresponding to each person. One possible solution is to apply the machine learning method. Due to different heights and contours among people, there exist notable differences between different person's virtual signals. The wall reflection is caused by the reflection of the virtual signal, thus the difference among wall reflections should be similar to the corresponding virtual signals. We could learn features to evaluate the similarity of the wall reflections and virtual signal to identify the corresponding peaks.

Fourth, our system is based on the ultrasonic signal. A concern is that ultrasonic signals may affect brain activity, which is called hypersonic effect [37]. The hypersonic effect sometimes would relax the person, and the side effect of ultrasound is still questionable. Existing works [8, 25] also provided different safety guidelines of the ultrasound. Among them, 70 dB is the most strict guideline. On the other hand, [58] pointed out that the COTs

devices have limited ability to play the ultrasound. According to these existing studies, we believe our sound pressure level 45dB (complying with the safety guidelines) played by the COTS devices, has no harm to humans.

Lastly, EchoSpot may be affected by the temperature, considering the slight change of the speed of sound. According to [9], the relationship between temperature and sound speed can be modeled by $v = 331 + 0.6^{*}\mathrm{T}$, where T is the celsius temperature. From this equation, we can see a 5-degree temperature variation will lead to the $3m/s$ sound speed change, which can impact the performance of our system to a certain extent if we cannot adapt to the new speed. But we would like to argue that, our system targets at the indoor localization, which typically has the A/C to maintain the stable temperature. Hence, the temperature variation can be ignored. On the other hand, since most smart devices have the built-in thermometers and hygrometers, we can incorporate their sensing into our system for adaptively adjusting the sound speed being used.

6 CONCLUSION

This paper proposes a novel device-free indoor localization system, EchoSpot, that leverages only one speaker and one microphone for localization via acoustic sensing. While controlling the speaker to emit the FMCW signals and using the co-located microphone to record the reflected signals, we develop a series of solutions for effectively processing to identify the ToF profile corresponding to the body reflection. Meanwhile, we propose a new approach by considering the body as sending virtual signals to identify the ToF profile corresponding to the wall reflection. In the end, we calculate the position based on the two reflections and then further correct it using the Kalman filter. Through implementing EchoSpot in the commercial device and conducting extensive experiments, we exhibit that EchoSpot can achieve promising localization accuracy, comparable to the state-of-the-art RF-based methods. In addition, it possesses salient advantages, including but are not limited to: 1) it is a software-based system that can be self-installed in COST audio devices; 2) it emits inaudible acoustic signals, which does not generate noise in a house; 3) it does not occupy the communication channels which thus has no interference to home Wi-Fi devices; 4) it cannot be affected by the lighting condition; 5) it does not require multiple receivers or transceivers.

ACKNOWLEDGMENTS

This work was supported in part by NSF under Grants 1763620, 1948374, 2019511, and in part by BoRSF under the contract LEQSF(2019-22)-RD-A-21. Any opinion and findings expressed in the paper are those of the authors and do not necessarily reflect the view of funding agency.

REFERENCES

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

[2] Sean L Bowman, Nikolay Atanasov, Kostas Daniilidis, and George J Pappas. 2017. Probabilistic data association for semantic slam. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. 1722–1729.

[3] Matthias Brugger, Tonia Christ, Ferdinand Kemeth, Sandor Nagy, Matthias Schaefer, and Michal M Pietrzyk. 2010. The FMCW technology-based indoor localization system. In *2010 Ubiquitous Positioning Indoor Navigation and Location Based Service*. IEEE, 1–6.

[4] Jagmohan Chauhan, Yining Hu, Suranga Seneviratne, Archan Misra, Aruna Seneviratne, and Youngki Lee. 2017. BreathPrint: Breathing acoustics-based user authentication. In *Proceedings of the 15th ACM Annual International Conference on Mobile Systems, Applications, and Services*. 278–291.

[5] Huijie Chen, Fan Li, and Yu Wang. 2017. EchoTrack: Acoustic device-free hand tracking on smart phones. In *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*. IEEE, 1–9.

[6] 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 Conference on Computer Communications (INFOCOM)*. 307–316.

[7] Linsong Cheng, Zhao Wang, Yunting Zhang, Weiyi Wang, Weimin Xu, and Jiliang Wang. 2020. AcouRadar: Towards Single Source based Acoustic Localization. In *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)*. 1848–1856.

[8] Mark D Fletcher, Sian Lloyd Jones, Paul R White, Craig N Dolder, Timothy G Leighton, and Benjamin Lineton. 2018. Effects of very high-frequency sound and ultrasound on humans. Part II: A double-blind randomized provocation study of inaudible 20-kHz ultrasound. *The Journal of the Acoustical Society of America* 144, 4 (2018), 2521–2531.
[9] Iowa State University Center for Nondestructive Evaluation. [n.d.]. Temperature and the Speed of Sound. https://www.nde-ed.org/Physics/Sound/tempandspeed.xhtml
[10] Hassen Fourati. 2014. Heterogeneous data fusion algorithm for pedestrian navigation via foot-mounted inertial measurement unit and complementary filter. *IEEE Transactions on Instrumentation and Measurement* 64, 1 (2014), 221–229.
[11] Ruipeng Gao, Yang Tian, Fan Ye, Guojie Luo, Kaigui Bian, Yizhou Wang, Tao Wang, and Xiaoming Li. 2015. Sextant: Towards ubiquitous indoor localization service by photo-taking of the environment. *IEEE Transactions on Mobile Computing* 15, 2 (2015), 460–474.
[12] Jon Gjengset, Jie Xiong, Graeme McPhillips, and Kyle Jamieson. 2014. Phaser: Enabling phased array signal processing on commodity WiFi access points. In *Proceedings of the 20th ACM Annual International Conference on Mobile Computing and Networking (MobiCom)*. 153–164.
[13] Daniel Graham, George Simmons, David T Nguyen, and Gang Zhou. 2015. A software-based sonar ranging sensor for smart phones. *IEEE Internet of Things Journal* 2, 6 (2015), 479–489.
[14] 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.
[15] Sidhant Gupta, Daniel Morris, Shwetak Patel, and Desney Tan. 2012. Soundwave: using the doppler effect to sense gestures. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. 1911–1914.
[16] Stefan L Hahn. 1996. *Hilbert transforms in signal processing*. Vol. 2. Artech House Boston.
[17] Yiqing Hu, Yan Xiong, Wenchao Huang, Xiang-Yang Li, Panlong Yang, Yanan Zhang, and Xufei Mao. 2018. Lightitude: Indoor positioning using uneven light intensity distribution. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–25.
[18] IndustryArc. 2017. Indoor Positioning and Navigation Market - Forecast(2020 - 2025). https://www.industryarc.com/Report/43/global-indoor-positioning-navigation-market.html
[19] Guang-yao Jin, Xiao-yi Lu, and Myong-Soon Park. 2006. An indoor localization mechanism using active RFID tag. In *IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing (SUTC'06)*, Vol. 1. IEEE, 4–pp.
[20] 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 (NSDI)*. 189–204.
[21] Manikanta Kotaru, Kiran Joshi, Dinesh Bharadia, and Sachin Katti. 2015. Spotfi: Decimeter level localization using WiFi. In *Proceedings of the ACM Conference on Special Interest Group on Data Communication*. 269–282.
[22] Ye-Sheng Kuo, Pat Pannuto, Ko-Jen Hsiao, and Prabal Dutta. 2014. Luxapose: Indoor positioning with mobile phones and visible light. In *Proceedings of the 20th annual international conference on Mobile computing and networking*. 447–458.
[23] Taras I Lakoba, David J Kaup, and Neal M Finkelstein. 2005. Modifications of the Helbing-Molnar-Farkas-Vicsek social force model for pedestrian evolution. *Simulation* 81, 5 (2005), 339–352.
[24] Christopher Langlois, Saideep Tiku, and Sudeep Pasricha. 2017. Indoor localization with smartphones: Harnessing the sensor suite in your pocket. *IEEE Consumer Electronics Magazine* 6, 4 (2017), 70–80.
[25] TG Leighton. 2016. Are some people suffering as a result of increasing mass exposure of the public to ultrasound in air? *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 472, 2185 (2016), 20150624.
[26] Qiongzheng Lin, Zhenlin An, and Lei Yang. 2019. Rebooting ultrasonic positioning systems for ultrasound-incapable smart devices. In *Proceedings of the 25th ACM Annual International Conference on Mobile Computing and Networking (MobiCom)*. 1–16.
[27] Kaikai Liu, Xinxin Liu, and Xiaolin Li. 2015. Guoguo: Enabling fine-grained smartphone localization via acoustic anchors. *IEEE transactions on mobile computing* 15, 5 (2015), 1144–1156.
[28] Song Liu and Tian He. 2017. Smartlight: Light-weight 3d indoor localization using a single led lamp. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*. 1–14.
[29] Li Lu, Jiadi Yu, Yingying Chen, Hongbo Liu, Yanmin Zhu, Yunfei Liu, and Minglu Li. 2018. LipPass: Lip reading-based user authentication on smartphones leveraging acoustic signals. In *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)*. 1466–1474.
[30] 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.
[31] 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 ACM Annual International Conference on Mobile Computing and Networking (MobiCom)*. 1–16.
[32] 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.
[33] Mostafa Mirshekari, Shijia Pan, Jonathon Fagert, Eve M Schooler, Pei Zhang, and Hae Young Noh. 2018. Occupant localization using footstep-induced structural vibration. *Mechanical Systems and Signal Processing* 112 (2018), 77–97.

[34] 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.

[35] Rajalakshmi Nandakumar, Alex Takakuwa, Tadayoshi Kohno, and Shyamnath Gollakota. 2017. Covertband: Activity information leakage using music. *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 87.

[36] Lionel M Ni, Yunhao Liu, Yiu Cho Lau, and Abhishek P Patil. 2003. LANDMARC: indoor location sensing using active RFID. In *Proceedings of the First IEEE International Conference on Pervasive Computing and Communications, 2003.(PerCom 2003)*. IEEE, 407–415.

[37] Tsutomu Oohashi, Emi Nishina, Manabu Honda, Yoshiharu Yonekura, Yoshitaka Fuwamoto, Norie Kawai, Tadao Maekawa, Satoshi Nakamura, Hidenao Fukuyama, and Hiroshi Shibasaki. 2000. Inaudible high-frequency sounds affect brain activity: hypersonic effect. *Journal of neurophysiology* (2000).

[38] Robert J Orr and Gregory D Abowd. 2000. The smart floor: A mechanism for natural user identification and tracking. In *Proceedings of the CHI Extended Abstracts on Human Factors in Computing Systems*. 275–276.

[39] Jacopo Pegoraro, Francesca Meneghello, and Michele Rossi. 2020. Multi-Person Continuous Tracking and Identification from mm-Wave micro-Doppler Signatures. *arXiv preprint arXiv:2003.03571* (2020).

[40] Chunyi Peng, Guobin Shen, and Yongguang Zhang. 2012. BeepBeep: A high-accuracy acoustic-based system for ranging and localization using COTS devices. *ACM Transactions on Embedded Computing Systems (TECS)* 11, 1 (2012), 1–29.

[41] Kun Qian, Chenshu Wu, Yi Zhang, Guidong Zhang, Zheng Yang, and Yunhao Liu. 2018. Widar2. 0: Passive human tracking with a single Wi-Fi link. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. 350–361.

[42] Christoph Schroeder and Hermann Rohling. 2010. X-band FMCW radar system with variable chirp duration. In *2010 IEEE Radar Conference*. IEEE, 1255–1259.

[43] Claude Elwood Shannon. 1949. Communication in the presence of noise. *Proceedings of the IRE* 37, 1 (1949), 10–21.

[44] 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.

[45] Elahe Soltanaghaei, Avinash Kalyanaraman, and Kamin Whitehouse. 2018. Multipath triangulation: Decimeter-level wifi localization and orientation with a single unaided receiver. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. 376–388.

[46] Roberto Sorrentino, Elisa Sbarra, Laura Urbani, Simone Montori, Roberto Vincenti Gatti, and Luca Marcaccioli. 2012. Accurate FMCW radar-based indoor localization system. In *2012 IEEE International Conference on RFID-Technologies and Applications (RFID-TA)*. IEEE, 362–368.

[47] Takafumi Taketomi, Hideaki Uchiyama, and Sei Ikeda. 2017. Visual SLAM algorithms: a survey from 2010 to 2016. *IPSJ Transactions on Computer Vision and Applications* 9, 1 (2017), 1–11.

[48] Dominik Van Opdenbosch, Georg Schroth, Robert Huitl, Sebastian Hilsenbeck, Adrian Garcea, and Eckehard Steinbach. 2014. Camera-based indoor positioning using scalable streaming of compressed binary image signatures. In *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2804–2808.

[49] Deepak Vasisht, Anubhav Jain, Chen-Yu Hsu, Zachary Kabelac, and Dina Katabi. 2018. Duet: Estimating user position and identity in smart homes using intermittent and incomplete RF-data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–21.

[50] Deepak Vasisht, Swarun Kumar, and Dina Katabi. 2016. Decimeter-level localization with a single WiFi access point. In *Proceedings of the 13th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*. 165–178.

[51] 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 & networking*. 3–14.

[52] Ju Wang, Hongbo Jiang, Jie Xiong, Kyle Jamieson, Xiaojiang Chen, Dingyi Fang, and Binbin Xie. 2016. LIFS: low human-effort, device-free localization with fine-grained subcarrier information. In *Proceedings of the 22nd ACM Annual International Conference on Mobile Computing and Networking (MobiCom)*. 243–256.

[53] 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.

[54] 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.

[55] Yanwen Wang, Jiaxing Shen, and Yuanqing Zheng. 2020. Push the Limit of Acoustic Gesture Recognition. In *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)*. 566–575.

[56] Yu-Lin Wei, Chang-Jung Huang, Hsin-Mu Tsai, and Kate Ching-Ju Lin. 2017. Celli: Indoor positioning using polarized sweeping light beams. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. 136–147.

[57] Greg Welch, Gary Bishop, et al. 1995. An introduction to the Kalman filter.

[58] Matt Wixey, Emiliano De Cristofaro, and Shane D Johnson. 2020. On the Feasibility of Acoustic Attacks Using Commodity Smart Devices. In *2020 IEEE Security and Privacy Workshops (SPW)*. IEEE, 88–97.

[59] Bo Xie, Kongyang Chen, Guang Tan, Mingming Lu, Yunhuai Liu, Jie Wu, and Tian He. 2016. LIPS: A light intensity–based positioning system for indoor environments. *ACM Transactions on Sensor Networks (TOSN)* 12, 4 (2016), 1–27.

[60] 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.

[61] Jie Xiong, Karthikeyan Sundaresan, and Kyle Jamieson. 2015. Tonetrack: Leveraging frequency-agile radios for time-based indoor wireless localization. In *Proceedings of the 21st ACM Annual International Conference on Mobile Computing and Networking (MobiCom)*. 537–549.

[62] Jingao Xu, Hao Cao, Danyang Li, Kehong Huang, Chen Qian, Longfei Shangguan, and Zheng Yang. 2020. Edge Assisted Mobile Semantic Visual SLAM. In *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)*. 1828–1837.

[63] 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 Conference on Computer Communications (INFOCOM)*. 297–306.

[64] Qilong Yuan and I-Ming Chen. 2014. Localization and velocity tracking of human via 3 IMU sensors. *Sensors and Actuators A: Physical* 212 (2014), 25–33.

[65] 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.

[66] Chi Zhang and Xinyu Zhang. 2016. LiTell: Robust indoor localization using unmodified light fixtures. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. 230–242.

[67] Linghan Zhang, Sheng Tan, and Jie Yang. 2017. Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 57–71.

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

[69] 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 *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, 33–40.

[70] Bing Zhou, Mohammed Elbadry, Ruipeng Gao, and Fan Ye. 2017. BatTracker: High precision infrastructure-free mobile device tracking in indoor environments. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*. 1–14.

[71] Shilin Zhu and Xinyu Zhang. 2017. Enabling high-precision visible light localization in today's buildings. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. 96–108.