ReflecTrack: Enabling 3D Acoustic Position Tracking Using Commodity Dual-Microphone Smartphones

Yuzhou Zhuang
Department of Computer Science and Technology, Global Innovation Exchange Institute, Tsinghua University
Beijing, China
zhuangyz19@mails.tsinghua.edu.cn

Yuntao Wang*
Department of Computer Science and Technology, Key Laboratory of Pervasive Computing, Ministry of Education, Tsinghua University
Beijing, China
yuntaowang@tsinghua.edu.cn

Yukang Yan
Department of Computer Science and Technology, Key Laboratory of Pervasive Computing, Ministry of Education, Tsinghua University
Beijing, China
yyk@mail.tsinghua.edu.cn

Xuhai Xu
Information School, University of Washington
Seattle, Washington, USA
xuhaixu@uw.edu

Yuanchun Shi
Department of Computer Science and Technology, Key Laboratory of Pervasive Computing, Ministry of Education, Tsinghua University
Beijing, China
shiyc@tsinghua.edu.cn

ABSTRACT

3D position tracking on smartphones has the potential to unlock a variety of novel applications, but has not been made widely available due to limitations in smartphone sensors. In this paper, we propose ReflecTrack, a novel 3D acoustic position tracking method for commodity dual-microphone smartphones. A ubiquitous speaker (e.g., smartwatch or earbud) generates inaudible Frequency Modulated Continuous Wave (FMCW) acoustic signals that are picked up by both smartphone microphones. To enable 3D tracking with two microphones, we introduce a reflective surface that can be easily found in everyday objects near the smartphone. Thus, the microphones can receive sound from the speaker and echoes from the surface for FMCW-based acoustic ranging. To simultaneously estimate the distances from the direct and reflective paths, we propose the echo-aware FMCW technique with a new signal pattern and target detection process. Our user study shows that ReflecTrack achieves a median error of 28.4 mm in the $60cm \times 60cm \times 60cm$ space and 22.1 mm in the $30cm \times 30cm \times 30cm$ space for 3D positioning. We demonstrate the easy accessibility of ReflecTrack using everyday surfaces and objects with several typical applications of 3D position tracking, including 3D input for smartphones, fine-grained gesture recognition, and motion tracking in smartphone-based VR systems.

*denotes the corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
UIST '21, October 10–14, 2021, Virtual Event, USA
© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8635-7/21/10…$15.00
https://doi.org/10.1145/3472749.3474805

CCS CONCEPTS

• **Human-centered computing → Sound-based input / output**; **Interaction techniques**.

KEYWORDS

Acoustic tracking, sound reflection, FMCW, smartphones

ACM Reference Format:
Yuzhou Zhuang, Yuntao Wang, Yukang Yan, Xuhai Xu, and Yuanchun Shi. 2021. ReflecTrack: Enabling 3D Acoustic Position Tracking Using Commodity Dual-Microphone Smartphones. In *The 34th Annual ACM Symposium on User Interface Software and Technology (UIST '21), October 10–14, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3472749.3474805

1 INTRODUCTION

3D position tracking with smartphones could potentially unlock a wide range of applications in 3D input, spatial awareness and VR/AR tracking, which can not only expand the input bandwidth, but also facilitate user experience [8, 24, 30]. However, due to limitations in existing smartphone sensors, it has not been widely implemented in modern day devices. Previous 3D tracking systems for smartphones relied heavily on external sensors or dedicated hardware [14, 24, 31], which are not available in commodity smartphones.

We propose ReflecTrack, a novel 3D acoustic position tracking method leveraging the built-in dual microphones that exist in almost all modern smartphones [13] and a ubiquitous speaker. The system tracks the 3D position of an active speaker device (e.g. an earbud or smartwatch) relative to the smartphone using acoustic ranging techniques. In order to enable 3D tracking with two microphones, we introduce a reflective surface near the device such that the microphones can not only receive sound from the direct path, but also echoes from the reflective path. The reflective surface can be easily found among everyday objects as long as it has a flat and smooth surface. Potential reflective surfaces include tabletops,

(a) (b) (c)

Figure 1: ReflecTrack enables 3D acoustic positioning for commodity dual-microphone smartphones by introducing a nearby reflective surface and developing echo-aware Frequency-Modulated-Continuous-Wave (FMCW). Applications include: (a) tracking the smartwatch movement with the desktop as the reflective surface; (b) tracking the earbud movements with the desk partition panel as the reflective surface; (c) tracking the smartwatch as an input device for smartphone-based VR headset by attaching a reflective surface to it. In the illustrations, the blue circle indicates the speaker, the solid yellow circles indicate the microphones, the dotted yellow circles indicate the mirror microphones, and the blue lines indicate different signal propagation paths.

pieces of cardboard, walls, or desk partition panels, providing favorable generalizability and scalability. ReflecTrack works as follows: The speaker device sends out a triangular chirp based Frequency Modulated Continuous Wave (FMCW) in inaudible frequencies. The smartphone is placed at an appropriate distance from the reflective surface and receives sound in both microphones. The acoustic ranging algorithm is able to recover distances of both the direct and the reflective path, effectively creating virtual microphones mirrored from the reflective surface according to the well-known image-source model [2]. These four distances (two for each microphone) can then be used to estimate the 3D position of the sound source using the triangulation algorithm. Therefore, the 3D position obtained from ReflecTrack is relative to the smartphone and the reflective surface.

ReflecTrack has two possible setups. One setup is placing the smartphone near an everyday reflective surface, turning the smartphone into a fixed 3D positioning proxy. The other setup is fabricating a smartphone add-on, allowing 3D tracking in motion. Figure 1 illustrates how ReflecTrack works in different scenarios. Without requiring dedicated hardware, ReflecTrack makes 3D position tracking more accessible to end users.

In order to model the reflective path, ReflecTrack assumes the reflective surface to be the closest object that generates the earliest echo. In our evaluation study, a reasonable surface-to-microphone separation of 16 cm makes sure the assumption is satisfied in most real-world scenarios. In case there is a closer surface to the smartphone (e.g., the supportive table in scenario b of Figure 1), one may cover it using easily found sound-absorbent materials (e.g., cloth) in order to mitigate irrelevant reflection. Furthermore, echoes from other environmental factors (e.g., user's body and hand) are usually not strong enough to form an earlier detection, thus not significantly affecting the overall tracking performance.

To overcome the challenges in simultaneous estimation of direct and reflective distances, we propose the echo-aware FMCW approach that improves the FMCW acoustic ranging technique proposed in CAT [14] with a new signal pattern and target detection process. In echo-aware FMCW, we leverage the triangular chirp signal to mitigate the Doppler effect as well as to enhance the signal-to-noise ratio (SNR), and apply the Constant False Alarm Rate (CFAR) peak detection process. This allows us to robustly identify the two target peaks corresponding to the direct path and the reflective path.

We also conducted extensive experimental evaluation studies and a user study to investigate the effect of both internal and external factors on the 3D tracking performance of ReflecTrack. The internal factors include the surface material, surface placement, surface-to-microphone distance, and the device used as the speaker. The external factors include motion speed, environmental noise, and user behavior. These results demonstrate that ReflecTrack is able to achieve accurate 3D position tracking with flexible configurations. The results also provide implementation guidelines for setting up the system in future use cases.

In summary, this paper makes the following major contributions:

(1) We present ReflecTrack as a general scheme for enabling relative 3D position tracking using commodity dual-microphone smartphones. We propose the echo-aware FMCW approach to implement the system.

(2) We present evaluation results that show ReflecTrack achieves a median error of 28.4 mm in the $60cm \times 60cm \times 60cm$ space and 22.1 mm in the $30cm \times 30cm \times 30cm$ space for 3D positioning. We also provide implementation guidelines for setting up the system based on our evaluation.

(3) We demonstrate that ReflecTrack is easily accessible through several demonstrative applications (3D object manipulation, fine-grained gesture recognition, and motion tracking for mobile virtual reality) using everyday surfaces and objects.

2 RELATED WORK

We first review the related work in near-device position tracking systems. Then, we summarize the acoustic-based methods in detail as they are more related to our work. Finally, we pay special attention to the echo-based position tracking systems.

2.1 Near-device Position Tracking

Researchers have been developing various near-device position tracking systems to extend the capabilities of user-device interaction. These systems can be classified based on the type of signal used for tracking and localization.

Infrared based systems use either infrared sensors or infrared cameras to measure the proximity of surrounding objects. Such systems can achieve coarse position awareness, but are limited by line-of-sight and do not provide fine-grained tracking. Furthermore, they usually require instrumenting the device or the target (e.g. the finger) with additional infrared sensors. iRing [17] uses an infrared reflection sensor to recognize finger gestures. LightRing [10] tracks the 2D location of a fingertip on any surface with an infrared proximity sensor and a 1-axis gyroscope. Both iRing and LightRing requires instrumenting the finger with a ring-form wearable device. SideSight [4] uses an array of infrared proximity sensors along the edges of the device to support multi-"touch" interactions around the device. Digits [11] recovers the full 3D pose of the user's hand to enable freehand 3D interactions using a wrist-worn infrared camera.

RF (Radio Frequency) signals have also been widely used in localization and tracking systems. WiTrack [1] leverages the RF-based FMCW technique to accurately localize multiple people to centimeter-scale in indoor multipath-rich environments. WiDraw [22] uses the angle-of-arrival values of incoming WiFi signals at the mobile device to track the user's hand trajectory, and achieves a median error lower than $5cm$. mTrack [26] uses highly-directional 60 GHz millimeter wave radios to achieve sub-centimeter tracking precision. These systems, however, require dedicated hardware setups and are not readily available in daily life. In addition, RF signal processing is usually computationally expensive for mobile devices due to its extremely high operating frequencies.

Recently, acoustic signals have been studied extensively for near-device tracking and localization. Compared to RF signals, sound travels at a significantly lower speed, thus providing same-level accuracy with much lower hardware requirements and computational costs. Furthermore, audio devices (i.e. microphones and speakers) are already widely available on modern mobile devices such as smartphones, smartwatches and earbuds. Currently the most advanced acoustic tracking systems (e.g. CAT [14], MilliSonic [24]) are able to achieve mm-level tracking accuracy on commodity devices. We will detail the different techniques of acoustic position tracking implemented in previous systems in Section 2.2.

Researchers have also explored the use of other signals and sensors. IMU sensors have been integrated in some systems to further improve the tracking accuracy. [11, 14] IMUs alone, however, cannot measure accurate absolute distance (>20 cm error long-run with calibration [28]) due to error accumulation in the double integration process [29]. uTrack [5] uses magnetic sensors for continuous finger tracking. It achieves an average tracking accuracy of 4.84mm in a limited 3D area. In addition, vision techniques have also been explored. [21] uses the built-in camera of a mobile device to recognize in-air gestures.

2.2 Techniques for Acoustic Position Tracking

Most acoustic position tracking systems rely on accurate distance measurements to perform triangulation in order to obtain the position. This usually requires some kinds of acoustic ranging techniques, which is essentially the core difference between acoustic position tracking systems.

Doppler effect based methods calculate the frequency shift to infer speed, and thus distance, of a moving object. Ultrasonic Doppler sonar has been used to recognize one-handed gestures. [9] Similarly, SoundWave [7] leverages the frequency shift of the reflection of an inaudible tone to sense in-air gestures around the device using sensors on existing commodity devices. AAMouse [29] uses the frequency shifts of transmitted signals to enable accurate device tracking and achieves a median error of around $1.4cm$. The range accuracy of the Doppler effect is both limited by the frequency resolution and the error accumulation in the integration process.

Correlation based methods perform autocorrelation or cross-correlation to find the best match between the received signal and the reference signal in order to determine the time-of-arrival (ToA), and thus the distance, from the speaker to the microphone. BeepBeep [18] leverages the linear chirp signal to perform autocorrelation and the two-way sensing technique to enable device-to-device ranging with 1-2 cm accuracy. Tracko [8] builds on the algorithm proposed in BeepBeep and fuses BLE, sound and IMUs to achieve 3D spatial awareness across devices. SwordFight [32] combines autocorrelation with cross-correlation to fundamentally reduce computational complexity while preserving accuracy. In practice, it is often hard to reliably locate the correlation peak even if the signal has good autocorrelation properties due to environmental noise and imperfect correlation, limiting the range accuracy and robustness.

Phase based methods treat received signals as phase modulated signals and extract distance information from phase shift. LLAP [25] measures the phase changes in the reflected sine wave signals caused by hand/finger movements and converts them into the distance of the movement, achieving a 2D tracking accuracy of 4.6 mm. SoundTrak [30] uses a similar method but extends the technique to 3D space. The system consists of a finger-worn speaker-embedded ring and a smartwatch with an array of microphones, and achieves an average accuracy of 1.3 cm. FingerIO [16] leverages the phase shift of the echoed OFDM signal to track the moving finger and achieves an average accuracy of 8 mm in 2D. However, phase methods could suffer from error accumulation because of the 2π phase ambiguity, especially for fast movement in echo-rich environment.

Frequency Modulated Continuous Wave (FMCW) has been widely used in radar ranging systems. However, the same techniques can also be applied to acoustic signals, making them the most accurate

acoustic ranging technique so far. The linear chirp based FMCW is a signal whose frequency changes linearly with time and has been used extensively in such systems. CAT [14] uses a distributed FMCW to accurately estimate the absolute distance between a transmitter and a receiver. It further combines FMCW estimation with the Doppler shifts and IMU measurements to achieve 8-9 mm 3D tracking accuracy. MilliSonic [24] utilizes the phase information of the FMCW signals to compute distances and further improves the tracking accuracy to 2.6 mm in 3D. Both CAT and MilliSonic use a four-speaker or four-microphone prototype to achieve 3D tracking. While most accurate, FMCW still requires handling of the multi-path effect and the motion effect to further improve robustness.

2.3 Echo-based Acoustic Position Tracking

In acoustic tracking systems, the microphone can not only receive the signal transmitted directly from the speaker, but also echoes of the original signal from surrounding objects. This results in the multi-path effect. Usually we want to leave only the direct path and minimize other irrelevant indirect paths because multi-path can cause significant distortion in the received signal, thus limiting tracking accuracy [24]. In some cases, however, echoes can be leveraged to facilitate acoustic position tracking.

In some systems, echoes are used to achieve device-free tracking, behaving like an active sonar system. The position or movement of the target is tracked by detecting the echo signals from it. For example, SoundWave [7] senses in-air gestures around the device by measuring the frequency shift of the inaudible tone when it reflects off moving objects such as the hand. ApneaApp [15] tracks the minute breathing movements by monitoring the reflected FMCW signals corresponding to the estimated distance value. FingerIO [16] transmits inaudible OFDM signals and tracks the echoes of the finger at its microphones. It realizes 2D finger tracking with an average accuracy of 8 mm and works even in the presence of occlusions between the finger and the device.

It has also been shown that knowledge of acoustic echoes could be used to reconstruct the geometry of an audio scene and benefit sound source localization (SSL) [6]. [3] estimate the 3D sound source position in indoor environments by generating and tracing direct and reflective acoustic paths. MIRAGE [6] introduces a reflective surface and uses a simple echo model to allow 2D SSL with only two microphones. This work is most similar to ours in concept but is limited to simulated data. Leveraging echo signals is significantly more difficult than direct signals and requires more sophisticated acoustic ranging techniques.

3 METHOD

We introduce the technical details of ReflecTrack in this section. ReflecTrack improved the linear FMCW approach to leverage echoes to enable 3D position tracking with only two microphones.

3.1 FMCW Background

Frequency Modulated Continuous Wave (FMCW) is widely used in radar systems to estimate the distance from the transmitter to the receiver. Recent research [14, 24] has shown that the technique could also be leveraged in acoustic tracking systems to achieve mm-level tracking accuracy.

In linear chirp based FMCW [14], we can denote the transmitted signal as $x_t(t) = A_0 cos(2\pi f_0 t + \pi \frac{B}{T} t^2)$, and the received signal as $x_r(t) = A_1 cos(2\pi f_0 (t - t_d) + \pi \frac{B}{T}(t - t_d)^2)$, where $B = f_1 - f_0$ is the chirp bandwidth, T is the chirp period and t_d is the delay time traveled by the signal. We then mix the transmitted signal x_t and the received signal x_r to obtain (after low-pass filter)

$$y(t) = \frac{A_0 A_1}{2} cos(2\pi f_0 t_d + \pi \frac{B}{T}(2tt_d - t_d^2)) \tag{1}$$

The delay time can be extracted from the peak frequency f_d in the FFT profile of $y(t)$: $t_d = \frac{T}{B} f_d$. Given the speed of sound in air c, the distance R between the transmitter and the receiver can then be calculated as

$$R = ct_d = \frac{cT}{B} f_d \tag{2}$$

FMCW is chosen over other acoustic ranging techniques due to its high tracking precision and robustness against noise. More importantly, it can be effectively extended to estimate multiple acoustic propagation paths. However, the traditional FMCW approach still faces challenges in handling the multi-path effect and the motion effect, which are not readily incorporated in the range equation 2. The multi-path effect could lower the signal-to-noise ratio and distort the FFT profile, while the motion effect could shift the FFT peak. These challenges become even more severe for ReflecTrack since we intentionally create and leverage echoes.

3.2 Echo-aware FMCW

Traditionally, FMCW uses the first significant FFT peak to estimate range corresponding to the shortest path, which is always the direct path [14]. In echo-aware FMCW, we leverage the multi-path effect and identify multiple peaks corresponding to not only the direct path but also different reflective paths as long as they are different in distance. In particular, when we add one reflective surface, we will observe one more significant peak corresponding to the reflective path from that surface. Here, we only consider the specular reflections from the surface since they carry relatively high energy [3] and use the image-source method [2] to model the reflective path for the physical microphone as the direct path for the mirror microphone.

There are two major advantages of echo-aware FMCW: 1) Enhanced sensing capability. Since we can easily create mirror microphones from a single physical microphone, this provides an accessible yet effective way to increase the number of microphones to enable previously infeasible tasks on existing devices due to sensor limitation. 2) Reduced processing time. Unlike standalone microphones, a pair of physical and mirror microphones are coupled together in the FFT computation. Therefore, we only need to do the signal processing steps once to extract two distances, which costs the majority of time in the traditional FMCW approach. Furthermore, we do not need to calibrate the two microphones separately since they are perfectly synchronized by nature. In order to identify multiple peaks in the more complex sound field due to echo interactions, we introduce some novel techniques to facilitate echo-aware FMCW.

Figure 2: Frequency pattern (top) and waveform (bottom) of triangular chirp

3.2.1 Triangular Chirp Based FMCW. In this section we describe the intuition as well as the formulation for triangular chirp based FMCW. Previously, we use the linear chirp to derive Equation 2 with the assumption that both the transmitter and the receiver are stationary. If we take velocity into account and suppose the transmitter is moving away from the receiver at a speed of v, the delay time then becomes $t_d = \frac{R+vt}{c}$. Substitute this into equation 1, we have

$$y(t) = \frac{A_0 A_1}{2} cos(2\pi f_0 \frac{R+vt}{c} + 2\pi \frac{Bt(R+vt)}{cT} - \pi \frac{B(R+vt)^2}{c^2 T}) \quad (3)$$

Ignoring the last term with $\frac{1}{c^2}$ and taking the mean value $t = \frac{T}{2}$ [14], the new peak frequency becomes

$$f_d = \frac{BR}{cT} + \frac{(f_0 + B)v}{c} \quad (4)$$

Compared to previous formulation, the error caused by motion is $\frac{(f_0+B)v}{c}$.

In order to mitigate the effect of motion on FFT, we introduce the triangular chirp based FMCW used in some FMCW radar systems [19], which concatenate two reversed linear FMCWs in one period: the first one sweeps from f_0 to f_1, and the second one sweeps from f_1 to f_0. Figure 2 demonstrates the frequency pattern as well as the waveform for triangular chirp. On receiving, we separately apply signal mixing and FFT (pulse compression) on each linear chirp and then average the two spectrums (pulse integration) to obtain the final FFT profile. Thus, the average peak frequency is

$$f_d = \frac{1}{2}[(\frac{BR}{cT} + \frac{(f_0 + B)v}{c}) + (\frac{BR}{cT} + \frac{(-f_1 + B)v}{c})] = \frac{BR}{cT} + \frac{Bv}{2c} \quad (5)$$

For chirps in the ultrasonic frequency band, the error term has reduced by one magnitude from $\frac{(f_0+B)v}{c}$ to $\frac{Bv}{2c}$, which is less than $1Hz$ for bandwidth $B = 6000Hz$ and motion speed $v = 0.1m/s$.

3.2.2 CFAR Peak Detection Process. Another challenge for echo-aware FMCW is to extract two peaks from a single FFT result. The interaction between the direct signal and echoes can not only lower the signal-to-noise ratio (SNR) significantly, but also result in ghost peaks and missed peaks. [12] In order to accurately identify the correct peaks, we adopt the Constant False Alarm Rate (CFAR) process used in typical FMCW radars. CFAR estimates the noise level from adjacent samples and achieves constant false alarm rate. [27] In particular, we use the cell-averaging CFAR with smallest of selection logic (SO-CFAR) in order to improve resolution of closely spaced peaks. [23]

The first two peaks that pass the CFAR adaptive threshold are selected corresponding to the direct and the reflective paths. Figure 3 shows the CFAR adaptive thresholds along with the first two peak detections on the FFT profile. Here, we plot the thresholds for all points for visualization. However, we actually only need to calculate them at each FFT peak. We observe that the SO-CFAR procedure is more robust than heuristic-based peak detection algorithms in echo-rich environments. Figure 4 demonstrates the FFT spectrogram along with the direct (lower curve) and the reflective (higher curve) distance as identified by the algorithm (in black) and obtained from ground truth (in red). We can see that echo-aware FMCW accurately and consistently finds the peaks for the direct and the reflective path.

Figure 3: CFAR adaptive thresholding for detecting multiple peaks

Figure 4: FFT spectrogram with direct and reflective distances

3.3 Calibration

In the traditional FMCW approach, the transmitter and the receiver are synchronized, meaning t_d represents the exact time traveled by the signal. In reality, however, the transmitter and the receiver can be separate devices with a previously unknown clock difference. In order to achieve accurate range estimation for distributed systems, CAT proposes the distributed FMCW approach [14]. This approach introduces a calibration phase to estimate the constant clock difference. In our experiments, we observe that the clock difference can also drift over time due to imperfect clocks. In order to cancel any

(a) (b) (c)

Figure 5: The 3D trajectory obtained from ReflecTrack versus the ground truth for a) a circle, b) a rectangle, and c) a spiral.

drift effect, we adopt the linear curve fitting technique used in both CAT [14] and MilliSonic [24], which models the clock difference as a linear function of time. To estimate the initial clock difference t_0 and the clock drift rate k, we touch the speaker device to one microphone of the smartphone during the short calibration phase assuming a reference distance of 0. Once calibrated, we can obtain the real distance R_{real} from the raw FMCW distance R_{FMCW}:

$$R_{real}(t) = R_{FMCW}(t) + c(t_0 + kt) \tag{6}$$

Since the dual microphones share the same internal clock, we only need to calibrate once at the beginning of each tracking session.

3.4 3D Position Estimation

In echo-aware FMCW, we can derive two distances from one physical microphone, corresponding to the direct path and the reflective path. For the reflective path, we can imagine a virtual microphone located at the mirror position from the reflective surface. The reflective path for the physical microphone is then the direct path for the mirror microphone. In the dual-microphone smartphone setting, we have two physical microphones and two virtual microphones. Thus, providing four distances to infer the 3D position of the sound source using triangulation [30]. Since the direct distances are more accurate than the reflective distances, we perform triangulation twice with two physical microphones and each of the mirror microphone, and simply average the inferred 3D positions. Figure 5 demonstrates a few 3D shapes we drew with ReflecTrack. We can see that the trajectories match well with the ground truth.

4 IMPLEMENTATION

In this section, we introduce both the hardware and software implementation details of ReflecTrack.

4.1 Hardware

The hardware of ReflecTrack consists of a dual-microphone smartphone, a speaker device, a PC, and a reflective surface. The smartphone is placed at an appropriate distance from the reflective surface and runs a customized Android app to record and transmit dual-channel audio data to the PC through Android Debug Bridge (ADB) using USB connection. The PC runs the echo-aware FMCW and triangulation algorithm on the received audio data to perform real-time 3D position tracking of the speaker device. Only commodity devices were used to implement the system. For the smartphone, we used a Samsung Galaxy S10+ and recorded dual-channel audio at a 48000Hz sampling rate and 16 bits sample width. For the PC, we used a Dell XPS 15 7590 laptop (CPU: Intel Core i7-9750H, 6 cores, 2.60GHz; RAM: 16GB). We chose two types of devices as the speaker: the Sony WF-1000XM3 wireless earbuds [1], and the Xiaomi Mi Watch (Standard Edition) [2]. Since wireless earbuds currently do not support high frequency audio transmission over Bluetooth, we wired the earbud with a 3.5mm audio cable and connected it to the laptop to play audio. The smartwatch has a single speaker on the left side of the screen and can play audio locally.

4.1.1 Everyday objects as a reflective surface. The reflective surface can be any flat and smooth surface that is large enough to reflect sound from the intended tracking space. Everyday objects that can be used as the reflective surface include walls, desk partition panels, and tabletops. We can also use raw materials such as acrylic, glass, wood, and cardboard to conveniently fabricate reflective surfaces that suit our applications. In addition, we can use sound absorbent materials found in daily life to mitigate echoes from irrelevant surfaces. We demonstrate a few of these objects and materials in Figure 6. Three reflective surfaces of size $60cm \times 40cm$ in different materials (acrylic, cardboard, wood) were tested in our implementation.

(a) (b) (c)

Figure 6: (a) Everyday objects for the reflective surface; (b) Fabricating the reflective surface; (c) Everyday objects for sound absorption

4.2 Signal

The speaker continuously transmits a triangular chirp signal with frequency sweeping between 17.5kHz and 23.5kHz, which is inaudible to most people [20]. Every period of the triangular chirp covers 2048 samples, with the first half sweeping linearly from 23.5kHz to 17.5kHz and the second half from 17.5kHz to 23.5kHz. We also apply the Tukey window with $\alpha = 0.1$ on each period to envelope the signal and avoid sudden changes in amplitude, which could cause clicking noises in speakers. The signal is pre-generated and saved in wav format for replay.

[1] https://www.sony.com/electronics/truly-wireless/wf-1000xm3
[2] https://www.mi.com/miwatch

4.3 Software

The echo-aware FMCW and triangulation software is implemented in Python and utilizes multi-threading to enable real-time processing. One thread continuously receives audio data from the smartphone and prepares the data for further processing. Another thread runs the echo-aware FMCW acoustic ranging algorithm on each audio channel to extract both the direct and the reflective distances. Once all four distances are calculated and collected, triangulation is performed to infer the 3D position of the speaker. In order to improve tracking smoothness, we apply FIR filters on both the 1D distances and the 3D trajectory. This method introduces a total delay of about 0.4 seconds.

Before entering the tracking phase, we used the initial 5 seconds to calibrate for the clock difference between the smartphone and the speaker device, as described in detail in Section 3.3. According to our evaluation study, one-time calibration is adequate for the whole session of more than 10 minutes.

In our implementation, it takes 5 ms on average to process a 42.7 ms (2048 samples in 48 kHz) triangular chirp signal, making ReflecTrack suitable for real-time 3D position tracking as well as integrating other downstream applications. Although we have implemented the software on PC for faster development, it can be easily deployed on mobile platforms to support real-time local processing due to its low computational cost.

5 EVALUATION

We first evaluated ReflecTrack in a controlled lab environment to investigate how different factors might affect its position tracking performance. We then recruited 10 participants to evaluate the system in a real office room where there were environmental noises like people chatting and keyboard typing.

In order to benchmark the 3D tracking performance across sessions, we defined a fixed $60cm \times 60cm \times 60cm$ cubic space (called the tracking space) around the smartphone and designed a series of drawing tasks strategically to cover the entire tracking space. We did not limit how one held the earbud or wore the smartwatch as long as the speaker was not blocked. We used the OptiTrack motion capture system [3] to collect the ground truth positions of the devices while we move the speaker device in the 3D space. In order to obtain the precise position within the OptiTrack system, we attached 4-5 optical markers around each device to form a rigid body and let the centroid or one of the markers align with the speaker/microphone.

5.1 Effect of Reflective Surface Material and Placement

The selection and setup of the reflective surface is crucial for ReflecTrack. In order to predict the reflective path and the location of the mirror microphone, we want the reflective surface to be flat and smooth. Here we evaluate the combinatorial effect of two factors related to reflective surface: material and placement.

We tested two placements of the reflective surface. In the first placement (the vertical placement), the surface is placed vertically on the table and the smartphone is placed parallel to the surface on the table at a certain distance. We also placed a towel under the smartphone to absorb reflection from the table surface. In the second placement (the horizontal placement), the surface is placed horizontally, while the smartphone is lifted above the surface with a supportive cylinder. For each placement, we tested three different materials of the reflective surface: acrylic, cardboard, and wood. These materials are chosen for their wide existence in daily life. For all six experiments, we put the smartphone at the $16cm$ distance from the reflective surface and used the earbud as the sound source. Figure 7 shows the boxplot of 3D tracking errors for each combination of surface material and placement. While the horizontal placement has a larger error variance than the vertical placement, there is no significant difference in the median 3D tracking error ($20 - 25mm$), proving that ReflecTrack is a general scheme applicable to various surface materials and placements. We chose to use the acrylic board in the vertical placement for future evaluation.

Figure 7: 3D tracking error for different surface materials and placements

5.2 Effect of Surface-to-microphone Distance

The surface-to-microphone distance can greatly affect the 3D tracking accuracy because it controls the separation between physical and virtual microphones. If the microphone separation is too small, the two FFT peaks may interfere or even overlap with each other, resulting in incorrect distance estimation. Furthermore, triangulation solutions have lower errors when the microphones are more separated as long as the echo signals are strong enough to be detected. Since the microphone separation is twice the surface-to-microphone distance, ReflecTrack also provides an accessible yet effective way to extend microphone separation in limited space. In order to evaluate the effect of surface-to-microphone distance, we tested 10 distances evenly spaced from $10cm$ to $24cm$. The results are shown in Figure 8. We can see that tracking error drops rapidly at first from $10cm$ to $14cm$, and remains small from $14cm$ to $24cm$ with the minimum at $18cm$. We chose to use the surface-to-smartphone distance of $16cm$ in future experiments since it is reasonably small and allows great 3D tracking performance ($21.9mm$ median error).

5.3 Effect of Motion Speed

Motion speed could affect tracking accuracy in different ways. First, the range equation 2 are derived ignoring the velocity. The Doppler effect caused by motion could distort the peaks and thus the distances. Even with motion error correction in triangular chirp based

[3] https://optitrack.com/

Figure 8: 3D tracking error for different surface-to-smartphone distances

FMCW, there is still a small error term related to motion speed. Second, faster changes in distance means a shorter context window we can rely on to make future inferences. Our method uses the triangular chirp signal to mitigate the motion effect and the CFAR adaptive thresholding process to make context-free inference. To evaluate the effectiveness of triangular chirp in reducing motion error, we asked users to move the earbud in different speeds and selected 5 independent sessions from slow to fast. The average motion speed for each session was calculated from the ground truth data provided by OptiTrack. Figure 9 shows the 3D tracking error with the linear chirp (blue) and the triangular chirp (orange). We used the second half of the triangular chirp (which is a standard linear chirp) to estimate the performance with the linear chirp. As expected, triangular chirp based FMCW is significantly more accurate and robust to motion effect, resulting in an average 17.5% decrease in the median 3D tracking error.

Figure 9: 3D tracking error for different motion speeds with the linear chirp and the triangular chirp

5.4 Effect of Environmental Noise

To evaluate the robustness of ReflecTrack to environmental noise, we conduct another tracking experiment with the same setup as before while playing background noise of people talking and phone ringing from a call center. The noise recording is played through a different smartphone placed in the same office room with maximum volume. Figure 10 shows the CDF of 3D tracking errors with and without environmental noise. We can see that ReflecTrack is resilient to environmental noise, which is consistent with the results in CAT [14] and MilliSonic [24].

Figure 10: Impact of Environmental Noise

5.5 User Study

5.5.1 Participants and Apparatus. We recruited 10 participants from campus (4 females and 6 males) to participate in the user study. The average age of our participants was 23.27 (SD = 1.60). All participants used their right hand to hold the earbud or wear the smartwatch.

As described in Section 4.1, we use the Samsung Galaxy S10+ dual-microphone smartphone and the Dell XPS 15-7590 laptop for the user study. Both the Sony WF-1000XM2 wireless earbud and the Xiaomi Mi Watch are tested as the speaker device. We set up a $60cm \times 40cm$ acrylic board vertically on a lab office table, and placed the smartphone $16cm$ away from the surface. The entire scene with a $60cm \times 60cm \times 60cm$ tracking space was surrounded by the OptiTrack motion capture system with 12 Prime 41 cameras for ground truth collection.

5.5.2 Experiment Design and Procedure. This user study explores the tracking performance of ReflecTrack across individuals and in different regions in space. Participants were asked to use two types of speaker devices (holding an earbud and wearing a smartwatch) and move in the $60cm \times 60cm \times 60cm$ tracking space. In order for participants to cover the entire tracking space as evenly as possible, we marked a 3×3 grid on the table and instructed them to perform a series of tasks strategically as well as moving with free motion.

The experimenter first introduced the purpose and procedure of the user study to the participant, and helped him/her become familiar with the operations. Then the participant was asked to complete four tracking sessions: 1) holding the earbud and moving in normal speed 2) holding the earbud and moving in slow and even speed 3) wearing the smartwatch and moving in normal speed 4) wearing the smartwatch and moving in slow and even speed. The experimenter helped the participant calibrate the earbud/smartwatch before handing the speaker device to them. In each session, the participant performed the following tasks: 1) drawing a circle and a triangle in the evenly distributed $3 \times 3 \times 3 = 27$ locations in space 2) drawing a spiral from bottom to top on the $3 \times 3 = 9$ grid points 3) moving freely in the entire tracking space for a while. We reminded

the participant to face the speaker downwards or towards the smartphone during movement so that it was not occluded. Between two sessions the participant took a rest. The whole experiment lasted about 40 minutes. Each participant received a $25 USD gift card for their time and effort.

5.5.3 Results. We collected 40 sessions of data from 10 participants in total. Each session lasted 5-10 minutes with 7000-14000 sample points, depending on the participant's motion speed. After examining all the tracking results, we excluded one session (session 3 from participant 10) from our analysis because we found out that this participant occluded the smartwatch speaker with his hand during this session.

First we investigate the 1D distance error across participants. Figure 11 shows the median error for both the direct distance (in blue) and the reflective distance (in orange) of each session. A factorial ANOVA indicates a statistically significant effect on the distance error of the acoustic propagation path ($F(1, 70) = 15.78, p < .01$) and the session ($F(3, 70) = 22.33, p < .01$). We can see that the distance error of the reflective path is consistently higher than the direct path for all sessions, which is reasonable since reflected signals tend to have lower SNR. In addition, the earbud sessions have better performance than the smartwatch sessions. We suspect it is because participants have a more consistent and stable way of holding the earbud compared to wearing and moving the smartwatch.

Figure 11: 1D distance error of the direct path and the reflective path for different sessions

Next we investigate the 3D tracking performance of ReflecTrack. Figure 12 contains six heatmaps with different height ranges, showing the 3D tracking error in each region of the entire tracking space. The two microphones of the smartphone are placed at (0, 0, 0) and (155, 0, 0), and the reflective surface is placed vertically at y=-160. For each grid we averaged the median 3D tracking error of the trajectory points inside it across all participants. White grids indicate there were no available trajectory points since they were usually at the edge of the tracking space. We can see from the heatmaps that ReflecTrack generally performs better in the central area of the entire tracking space. If the speaker is too close to the reflective surface, the distance difference of the direct and reflective paths becomes small and the algorithm may fail to find the two correct peaks since they overlap too much to be distinguished. On the other hand, if the speaker is too far from the reflective surface, the echo signal could be too weak to stand out in the FFT profile. According to the heatmaps, we designate an inner $30cm \times 30cm \times 30cm$ cubic space (defined by $-100 < x < 200, 0 < y < 300, 100 < z < 400$) which represents the high-accuracy region of ReflecTrack.

Figure 12: Heatmaps of 3D tracking error within different height (z) ranges. White color indicates no data for that cell.

Figure 13 shows the 3D tracking error in both the entire $60cm \times 60cm \times 60cm$ tracking space and the $30cm \times 30cm \times 30cm$ high-accuracy region for each session across participants. A factorial ANOVA indicates that the 3D tracking error in the high-accuracy region is significantly lower than that in the entire tracking space ($F(1, 70) = 24.47, p < .01$), which can be leveraged in tasks requiring extra tracking accuracy. Similar to the 1D distance error, we observe that the sessions with slow and even speed has a lower 3D tracking error than the sessions with normal speed. We found that some participants move not only much faster but also more arbitrarily in the normal speed session, posing more challenges to the current acoustic ranging algorithm. The median 3D tracking error average across all participants for the second session is $28.4mm$ in the entire tracking space and $22.1mm$ in the high-accuracy region.

Figure 13: 3D tracking error in the entire tracking space and the high-accuracy region for different sessions

We further investigate the effect of motion speed and speaker device on the 3D tracking performance. Figure 14 shows the 3D tracking error in the high-accuracy region against the motion velocity for the earbud (blue) and the smartwatch (orange) sessions.

We can see that the 3D tracking error remains relatively stable for a long speed range and only gradually increases after motion speed exceeds $300mm/s$. Finally, We show the CDFs of the 3D tracking errors for the second session in Figure 15 , which show consistent and accurate tracking performance across all participants.

Figure 14: 3D tracking error against motion speed for all sessions

Figure 15: CDF of 3D tracking error for all participants

6 APPLICATIONS

ReflecTrack provides a general scheme for making 3D position tracking more accessible to end users using a commodity dual-microphone smartphone and a ubiquitous speaker. It has great flexibility in choosing/fabricating the reflective surface and setting up the system, enabling a number of previously infeasible usage scenarios. We present several demonstrative applications in this section using everyday surfaces and objects.

6.1 3D Input Space for Smartphones

ReflecTrack effectively creates a 3D absolute input space around the smartphone. 3D position tracking allows both 3DoF control and absolute coordinate input, which can be used to manipulate 3D objects or add another dimension of control to 2D interactions. We present an application of 3D object manipulation in office settings. As demonstrated in Figure 16, a user can transform the office desk into a 3D tracking space by simply placing a smartphone parallel to the commonly seen desk partition panel. One may also put a cloth or mouse pad under the smartphone to absorb reflection from the desk. The user can then control the viewpoint of a 3D object (which we displayed on the monitor for visualization) by moving an earbud in the air. This is achieved by deriving the elevation and azimuth angles of the object from the 3D position of the earbud relative to the smartphone, thus enabling real-time manipulation of the viewpoint. Such controls can also be used in 3D mobile games for more intuitive 3D interaction.

6.2 Fine-grained Gesture Recognition

The 3D trajectories recorded in ReflecTrack can also be used to realize fine-grained gesture recognition. Since a trajectory stores the absolute 3D position of each point, subtle differences in trajectories can be detected and quantified; adding another level of information to each gesture. We present an application of fine-grained gesture recognition in video player controls in Figure 17. A user is making dinner in the kitchen while watching a cooking video. He/She can

Figure 16: Left: ReflecTrack can control the viewpoint of a 3D object by placing a smartphone next to the desk partition panel. Right: The scene labeled with the speaker (blue dot), the microphones (yellow dots) and the acoustic propagation paths (yellow lines).

set up the tracking system by holding up the smartphone with some supportive object (e.g., a bottle) and using the table top as the reflective surface. The smartphone then tracks the 3D position of the smartwatch worn on the user's wrist. The user can not only perform simple gesture commands like play/pause/next, but also conduct accurate controls like dragging the progress bar to a specific location or turning up/down the volume to a specific value. To realize fine-grained gesture recognition, we adopt a two-stage pipeline. We first detect the gesture category of the recent trajectory using a simple rule-based classification model, and then analyze the key points in the trajectory to extract quantifiable measures of the gesture. For more complex gesture set, we can leverage machine learning to perform end-to-end inference as well.

Figure 17: Left: ReflecTrack tracks fine-grained in-air gestures to control the video player remotely. Right: The scene labeled with the speaker (blue dot), the microphones (yellow dots) and the acoustic propagation paths (yellow lines).

6.3 Motion Tracking in Smartphone VR Systems

ReflecTrack can also support 3D position tracking in motion by fabricating a smartphone add-on. One typical application is to enable 3D position tracking in smartphone-based VR systems. Lack of motion tracking in such systems have greatly limited their popularity and applicability. MilliSonic [24] enables 6DoF motion tracking capability for smartphone-based VR headsets using a four-microphone array as a beacon, which immediately reduces the accessibility of such systems. ReflecTrack, on the other hand, enables 3D position tracking by simply attaching a reflective surface to the VR headsets during assembly. We demonstrate its applicability using the Google Cardboard VR [4]. A cardboard add-on can be

[4] https://arvr.google.com/cardboard/

easily cut out and attached to the headset to create a local reflection structure. The modified cardboard VR system is able to track the 3D position of an external speaker relative to the headset. In this application (Figure 18), the user can use the earbud or wear the smartwatch as a traditional VR controller to perform 3D target selection tasks without rotating his/her head. The real world coordinates of the device is transformed into the VR world coordinates to determine whether the controller touches the target.

Figure 18: Left: ReflecTrack enables 3D motion tracking in Google Cardboard VR by attaching a cardboard surface on the top of the viewer. Right: The scene labeled with the speaker (blue dot), the microphones (yellow dots) and the acoustic propagation paths (yellow lines).

7 DISCUSSION, LIMITATION, AND FUTURE WORK

In this paper, we present ReflecTrack, a novel method that enables 3D acoustic position tracking for commodity dual-microphone smartphones. To achieve this, we introduce a reflective surface near the smartphone and propose the echo-aware FMCW approach. The 3D position obtained from ReflecTrack is thus relative to the smartphone and the reflective surface. We describe both the hardware and software implementations in detail. Our evaluation and user study investigate the effect of both internal factors (e.g. surface material, surface placement, surface-to-microphone distance, speaker device) and external factors (e.g. motion speed, environmental noise, participants) on the 3D tracking performance. These results not only demonstrate that ReflecTrack achieves relatively high 3D tracking accuracy with only commodity devices, but also provide guidelines for setting up the system in future scenarios.

While we have presented multiple promising applications of ReflecTrack, deeper understanding of sound reflection and acoustic ranging is required to inform better system design as well as achieve higher accuracy. Next we discuss a few limitations in the current implementation of ReflecTrack as well as potential future work.

7.1 Comparing with Prior Work

We compare ReflecTrack with several typical acoustic tracking systems that leverage the properties of different signal encodings. SoundTrak [30] uses sine wave signals, FingerIO [16] uses OFDM, CAT [14] uses frequency of FMCW, and MilliSonic [24] uses phase of FMCW. We develop our echo-aware FMCW based on the traditional FMCW in CAT not only because it has relatively high accuracy but also because it can be naturally extended to multiple acoustic propagation paths. Another difference of ReflecTrack to these previous systems is that ReflecTrack relies on only two microphones to perform 3D position tracking, while SoundTrak, CAT, MilliSonic requires dedicated hardware and FingerIO leaves this to future work by using a third microphone. Since ReflecTrack intentionally create and estimate echoes, it suffers from a lower tracking accuracy and requires a more careful selection and setup of the tracking scene. Specifically, the smartphone should be placed at an appropriate distance from the reflective surface and interference from closer surfaces should be avoided or mitigated. Among the aforementioned systems, FingerIO achieves device-free tracking by transforming the device into an active sonar and estimating echoes from the finger. While this technique frees the hand from holding or wearing a speaker device, it assumes that the finger is the closest moving object and may not be compatible with ReflecTrack since they use echoes in different ways.

7.2 Improving Tracking Accuracy

In the user study, ReflecTrack achieves a median 3D tracking error of 28.4 mm in the entire $60cm \times 60cm \times 60cm$ tracking space and 22.1 mm in the central $30cm \times 30cm \times 30cm$ high-accuracy region. The tracking accuracy is worse than previous FMCW based system such as CAT [14] and MilliSonic [24] because every triangulation requires both the direct distances and the reflective distances. We can see from Figure 11 that the median distance error is slightly larger than what is reported in CAT, especially for the reflective path. We argue that the drop in accuracy is due to the interference of the direct and the reflective paths, resulting in overlapping of two peaks in the FFT. Such overlapping produces wider peaks and lower SNR for both peaks. To effectively increase SNR, we can try reflective surfaces with better reflection properties, or use speakers/microphones that support higher signal gain. We can also design a new signal pattern that is more resilient to echo interference. Alternatively, we can model the overlapping of multiple peaks and recover the peak frequencies from the model rather than the FFT profile. We leave these attempts to future work.

7.3 Handling Irrelevant Multi-path Effect

Usually the presence of a close reflective surface makes the echoes from it the earliest and strongest indirect path, significantly reducing other irrelevant multi-path effect. In our experiments, we did observe that tracking performance degraded if there was another flat object nearby in addition to the reflective surface (e.g. the smooth tabletop while using a vertical reflective surface). We suspect that there exists not only one-time reflection from each surface but also multi-time reflection between the surfaces, causing more complex multi-path effect. Our current solution is to remove any nearby large object that has a flat and smooth surface or use sound absorbent material (e.g. cloth) to cover them to mitigate irrelevant multi-path effect. A more sophisticated algorithm is required to accurately recover the direct and reflective distances in these cases.

7.4 Supporting Multiple Reflective Surfaces

One immediate question for ReflecTrack is whether it can support more than one reflective surface, creating multiple mirror microphones from a single microphone and thus enabling 3D position tracking even for single-microphone systems. We believe this is possible by appropriately arranging the reflective surfaces as well

as carefully restricting the tracking space such that the reflective distances of different reflective surfaces always follow a fixed order. The CFAR peak detection algorithm can be easily extended to identify multiple peaks in one pass of the FFT. However, we should be concerned about dealing with the multi-path effect between reflective surfaces as mentioned in Section 7.3.

8 CONCLUSION

This paper presents ReflecTrack, a novel 3D acoustic position tracking method for commodity dual-microphone smartphones. The system leverages the FMCW acoustic ranging technique to track the 3D position of an external sound source relative to the smartphone. In order to enable 3D tracking with two microphones, we introduce a reflective surface near the smartphone to create mirror microphones, and propose the echo-aware FMCW approach to facilitate acoustic ranging of the reflective path. Evaluation results show that ReflecTrack achieves a median error of 28.4 mm in the $60cm \times 60cm \times 60cm$ space and 22.1 mm in the $30cm \times 30cm \times 30cm$ space for 3D positioning. We provide general guidelines for implementing the system as well as demonstrate its accessibility with several demonstrative applications. We compare ReflecTrack with prior work and discuss its advantages and limitations. In the future, we will continue improving the tracking accuracy and robustness of ReflecTrack, and exploring the design space for the reflective surface, which will further extend its usability in various application scenarios.

ACKNOWLEDGMENTS

This work is supported by the National Key R&D Program of China under Grant No. 2018YFB1005000, the Natural Science Foundation of China under Grant No. 62002198, and the China Postdoctoral Science Foundation under Grant No. 2021M691788. Our work is also supported by the Beijing Key Lab of Networked Multimedia, Beijing Academy of Artificial Intelligence (BAAI), as well as Undergraduate / Graduate Education Innovation Grants, the Institute for Guo Qiang and Institute for Artificial Intelligence, Tsinghua University (THUAI). We would like to thank Robin Yang and all participants for their time and effort.

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 ({NSDI} 15)*. 279–292.

[2] Jont B Allen and David A Berkley. 1979. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America* 65, 4 (1979), 943–950. https://doi.org/10.1121/1.382599

[3] Inkyu An, Myungbae Son, Dinesh Manocha, and Sung-Eui Yoon. 2018. Reflection-aware sound source localization. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 66–73. https://doi.org/10.1109/ICRA.2018.8461268

[4] Alex Butler, Shahram Izadi, and Steve Hodges. 2008. SideSight: multi-" touch" interaction around small devices. In *Proceedings of the 21st annual ACM symposium on User interface software and technology*. 201–204. https://doi.org/10.1145/1449715.1449746

[5] Ke-Yu Chen, Kent Lyons, Sean White, and Shwetak Patel. 2013. uTrack: 3D input using two magnetic sensors. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. 237–244. https://doi.org/10.1145/2501988.2502035

[6] Diego Di Carlo, Antoine Deleforge, and Nancy Bertin. 2019. Mirage: 2d source localization using microphone pair augmentation with echoes. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 775–779. https://doi.org/10.1109/ICASSP.2019.8683534

[7] Sidhant Gupta, Daniel Morris, Shwetak Patel, and Desney Tan. 2012. Soundwave: using the doppler effect to sense gestures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1911–1914. https://doi.org/10.1145/2207676.2208331

[8] Haojian Jin, Christian Holz, and Kasper Hornbæk. 2015. Tracko: Ad-hoc mobile 3d tracking using bluetooth low energy and inaudible signals for cross-device interaction. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. 147–156. https://doi.org/10.1145/2807442.2807475

[9] Kaustubh Kalgaonkar and Bhiksha Raj. 2009. One-handed gesture recognition using ultrasonic Doppler sonar. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1889–1892. https://doi.org/10.1109/ICASSP.2009.4959977

[10] Wolf Kienzle and Ken Hinckley. 2014. LightRing: always-available 2D input on any surface. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. 157–160. https://doi.org/10.1145/2642918.2647376

[11] David Kim, Otmar Hilliges, Shahram Izadi, Alex D Butler, Jiawen Chen, Iason Oikonomidis, and Patrick Olivier. 2012. Digits: freehand 3D interactions anywhere using a wrist-worn gloveless sensor. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. 167–176. https://doi.org/10.1145/2380116.2380139

[12] Jau-Jr Lin, Yuan-Ping Li, Wei-Chiang Hsu, and Ta-Sung Lee. 2016. Design of an FMCW radar baseband signal processing system for automotive application. *SpringerPlus* 5, 1 (2016), 1–16. https://doi.org/10.1186/s40064-015-1583-5

[13] Iván López-Espejo, Angel M Gomez, José A González, and Antonio M Peinado. 2014. Feature enhancement for robust speech recognition on smartphones with dual-microphone. In *2014 22nd European Signal Processing Conference (EUSIPCO)*. IEEE, 21–25.

[14] 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. https://doi.org/10.1145/2973750.2973755

[15] Rajalakshmi Nandakumar, Shyamnath Gollakota, and Nathaniel Watson. 2015. Contactless sleep apnea detection on smartphones. In *Proceedings of the 13th annual international conference on mobile systems, applications, and services*. 45–57. https://doi.org/10.1145/2742647.2742674

[16] 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. https://doi.org/10.1145/2858036.2858580

[17] Masa Ogata, Yuta Sugiura, Hirotaka Osawa, and Michita Imai. 2012. iRing: intelligent ring using infrared reflection. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. 131–136. https://doi.org/10.1145/2380116.2380135

[18] Chunyi Peng, Guobin Shen, Yongguang Zhang, Yanlin Li, and Kun Tan. 2007. Beepbeep: a high accuracy acoustic ranging system using cots mobile devices. In *Proceedings of the 5th international conference on Embedded networked sensor systems*. 1–14. https://doi.org/10.1145/1322263.1322265

[19] Diego F Pierrottet, Farzin Amzajerdian, Larry Petway, Bruce Barnes, George Lockard, and Manuel Rubio. 2008. Linear FMCW laser radar for precision range and vector velocity measurements. *MRS Online Proceedings Library* 1076, 1 (2008), 1–9. https://doi.org/10.1557/PROC-1076-K04-06

[20] A Rodríguez Valiente, A Trinidad, JR García Berrocal, C Górriz, and R Ramírez Camacho. 2014. Extended high-frequency (9–20 kHz) audiometry reference thresholds in 645 healthy subjects. *International journal of audiology* 53, 8 (2014), 531–545. https://doi.org/10.3109/14992027.2014.893375

[21] Jie Song, Gábor Sörös, Fabrizio Pece, Sean Ryan Fanello, Shahram Izadi, Cem Keskin, and Otmar Hilliges. 2014. In-air gestures around unmodified mobile devices. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. 319–329. https://doi.org/10.1145/2642918.2647373

[22] 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. https://doi.org/10.1145/2789168.2790129

[23] Gerard V Trunk. 1978. Range resolution of targets using automatic detectors. *IEEE Trans. Aerospace Electron. Systems* 5 (1978), 750–755. https://doi.org/10.1109/TAES.1978.308625

[24] 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. https://doi.org/10.1145/3290605.3300248

[25] 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. https://doi.org/10.1145/2973750.2973764

[26] 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. https://doi.org/10.1145/2789168.2790113

[27] M Weiss. 1982. Analysis of some modified cell-averaging CFAR processors in multiple-target situations. *IEEE Trans. Aerospace Electron. Systems* 1 (1982), 102–114. https://doi.org/10.1109/TAES.1982.309210

[28] Cheng Xu, Jie He, Yuanyuan Li, Xiaotong Zhang, Xinghang Zhou, and Shihong Duan. 2019. Optimal estimation and fundamental limits for target localization using IMU/TOA fusion method. *IEEE Access* 7 (2019), 28124–28136. https://doi.org/10.1109/ACCESS.2019.2902127

[29] 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. https://doi.org/10.1145/2742647.2742662

[30] 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. https://doi.org/10.1145/3090095

[31] 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. https://doi.org/10.1109/INFOCOM.2018.8486365

[32] Zengbin Zhang, David Chu, Xiaomeng Chen, and Thomas Moscibroda. 2012. Swordfight: Enabling a new class of phone-to-phone action games on commodity phones. In *Proceedings of the 10th international conference on Mobile systems, applications, and services*. 1–14. https://doi.org/10.1145/2307636.2307638