FM-Track: Pushing the Limits of Contactless Multi-target Tracking using Acoustic Signals

Dong Li
University of Massachusetts Amherst
dli@cs.umass.edu

Jialin Liu
University of Massachusetts Amherst
Dalian University of Technology
jialinliu@umass.edu

Sunghoon Ivan Lee
University of Massachusetts Amherst
silee@cs.umass.edu

Jie Xiong
University of Massachusetts Amherst
jxiong@cs.umass.edu

ABSTRACT

Contactless acoustic motion tracking enables new opportunities to interact with smart devices, such as smartphones and voice-controlled smart assistants. The speakers and microphones integrated in these devices provide unique opportunities to simultaneously track multiple targets in a fine-grained manner. To this end, we propose a system, namely FM-Track, that enables contactless multi-target tracking using acoustic signals. We first introduce a signal model to characterize the location and motion status of targets by fusing the information from multiple dimensions (i.e., range, velocity, and angle of targets). Then we develop a series of techniques to separate signals reflected from multiple targets and accurately track each individual target. We implement and evaluate FM-Track on both research-purpose hardware platform (i.e., Bela) and commercial devices (i.e., smartphones and smart speakers). Extensive experiments show that FM-Track can successfully differentiate two targets with a spacing as small as 1 cm, and achieve a median tracking accuracy of 0.86 cm and 0.11 cm for absolute range and displacement estimates respectively. For multi-target tracking, FM-Track can accurately track four targets and the tracking range can be up to 3 m.

CCS CONCEPTS

• **Human-centered computing** → *Ubiquitous and mobile computing systems and tools*.

KEYWORDS

Acoustic motion tracking, Contactless tracking, Multi-target tracking, Multi-dimensional estimation

ACM Reference Format:
Dong Li, Jialin Liu, Sunghoon Ivan Lee, and Jie Xiong. 2020. FM-Track: Pushing the Limits of Contactless Multi-target Tracking using Acoustic Signals. In *The 18th ACM Conference on Embedded Networked Sensor Systems (SenSys '20), November 16–19, 2020, Virtual Event, Japan.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3384419.3430780

SenSys '20, November 16–19, 2020, Virtual Event, Japan

ACM ISBN 978-1-4503-7590-0/20/11…$15.00
https://doi.org/10.1145/3384419.3430780

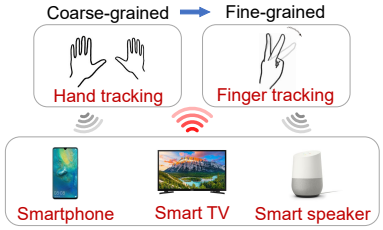


Figure 1: Application scenarios for concurrently tracking multiple hands and fingers using smart devices.

1 INTRODUCTION

Speakers and microphones are essential components in many smart devices that people interact with on a daily basis, such as smartphones, personal computers, smart TVs, and smart speakers (e.g., Amazon Alexa). Owing to the continuous advancement in processing capability, recent research has successfully demonstrated the possibility to extend their primary use from simple audio playing and voice-based interactions to multifarious applications, such as localization of a sound source [17, 29, 33], contactless motion tracking [28, 34, 45, 53] and gesture recognition [11, 31, 55], as well as monitoring of important physiological parameters in humans (e.g., fatigue detection for drivers [46, 50] and respiratory activities [5, 42, 44]). Specifically for contactless motion tracking, compared to other wireless signals such as WiFi [21, 49] and RFID [43, 51], acoustic signals have inherent superiority for sensing granularity and precision, owing to its low propagation speed (340 m/s) in the air.

Although recent efforts have pushed the granularity of acoustic tracking to millimeter level [28, 34, 45, 53] and extended the sensing range to 4.5 m [24] without requiring users to instrument a device, there still exist several fundamental challenges that hinder the widespread adoption of acoustic tracking in the real world. First, studies to date pose difficulties in tracking more than one target due to the inherent nature of contactless tracking that relies on signals reflected from the targets. Signals reflected from multiple targets are mixed at the receiver, and thus, it is difficult to separate them to obtain the context information of each individual target. This problem becomes even more challenging when targets are close to each other. Second, it is non-trivial to achieve a similar level of tracking accuracy for multiple targets comparable to prior research on single-target tracking [24, 28, 45]. More specifically, the signals

reflected from close-by targets could interfere with one another, thereby significantly degrading the tracking accuracy. Last but not least, while existing work has achieved millimeter-level accuracy in relative displacement and trajectory tracking of a single target, the estimation of the absolute distance between the target and the sensing device still offers room for improvement [6, 45].

In this paper, we propose to employ chirp-based signals to push the boundaries of acoustic tracking in three aspects: 1) enable multi-target sensing; 2) boost the multi-target tracking granularity to finger level; and 3) accurately track not only the target's relative distance change but also its absolute distance from the sensing device. As shown in Fig. 1, we believe the proposed system will support a number of HCI applications that were previously infeasible, e.g., tracking two hands to play video games with a smartphone/TV and tracking multiple close-by fingers to interact with a smart speaker.

In the literature, a lot of efforts have been devoted to enable multi-target tracking using wireless signals, such as WiFi [14, 15, 47] and radar/sonar signals [1, 4, 9, 20, 22]. However, these techniques do not directly translate to acoustic-based multi-target tracking on commodity devices, mainly due to the following reasons.

- Unlike OFDM-based WiFi signals that have multiple sub-carriers at different frequencies, the frequency of chirp-based acoustic signals changes linearly over time. This difference requires us to redesign the signal separation algorithm to achieve multi-target tracking with acoustic signals.
- Compared to WiFi or radar/sonar-based systems that usually track moving objects with a velocity in the scale of meter or decimeter per second (e.g., human walking), this work aims to achieve finer-grained velocity estimation in the scale of centimeter per second (e.g., hand/finger movements). To achieve this level of velocity estimation accuracy, data samples from a large time window are required [11]. This introduces an interesting dilemma: a larger time window is needed for more accurate estimation, but only a single velocity estimate (i.e., average velocity) can be obtained. On the other hand, many real-world objects (e.g., hands) rarely move at constant velocities and an average velocity is not able to precisely characterize the hand movement. Accurate estimation of instantaneous velocity using acoustic signals has not been addressed, which creates unique technical challenges for our goal to realize fine-grained hand/finger tracking.
- Another issue that is more prominent in acoustic tracking—when compared to WiFi or radar/sonar-based tracking—is the range-Doppler effect that yields a relatively large range estimation deviation [24]. For example, a chirp-based acoustic signal with a start frequency of $f_0 = 16\ kHz$, a bandwidth of $B = 4\ kHz$, and a sweep time of $T = 0.04\ s$, would induce a 47.1 Hz (i.e., $\Delta f = \frac{2v}{v_s} f_0$, where v_s is the speed of sound in the air) Doppler shift for a target moving at a velocity of $v = 0.5\ m/s$. This Doppler shift will cause a deviation of 8 cm (i.e., $\Delta d = \frac{v f_0 T}{B}$) when estimating the distance between the target and the sensing device. This is unacceptably large for millimeter-level tracking that we envision in our study.

To realize our vision, we develop the first contactless **F**ine-grained **M**ulti-target **Track**ing system using acoustic signals, namely FM-Track. We believe the proposed method could be generalized and applied to other chirp-based signals (e.g., LoRa signals) to enable fine-grained, multi-target sensing.

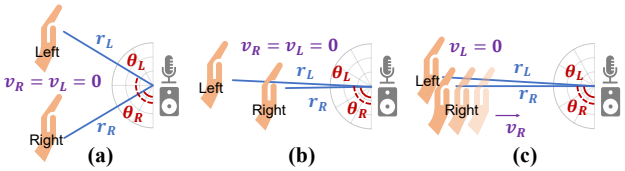


Figure 2: Separate targets with information from multiple dimensions (range r, angle θ and velocity v): (a) targets with similar r can be separated by different θ; (b) targets with similar θ can be separated by different r; (c) targets with similar r and θ can be separated by different v.

We first propose a chirp-based signal model to fuse the range (time domain), velocity (frequency domain), and angle (spatial domain) information of targets from reflected signals that can be collectively used to characterize the location and motion status of the targets, as shown in Fig. 2. Based on the signal model, we introduce an algorithm that can efficiently resolve each individual signal reflected from multiple targets based on joint parameter (range, angle, velocity and attenuation) estimation.

To address the instantaneous velocity and range-Doppler effect issues, we propose a novel method based on the fact that the information estimated from multiple dimensions are not equally accurate. In acoustic tracking, range estimation is much more accurate than the other dimensions (i.e., angle and velocity) owing to the low propagation speed. Thus, we can employ the range estimates obtained from two adjacent chirps to compute the instantaneous velocity. Once this instantaneous velocity is obtained, the Doppler shift-caused range deviation can be estimated and removed to refine the range estimate.

To track multiple targets, we need to continuously estimate the parameters of signals reflected by the targets. It is non-trivial to match the parameters from two consecutive estimates for the same target. Furthermore, there exists an intrinsic ambiguity issue in estimating the velocity of a moving target using chirp signals [30]. To address the above issues, we employ the fact that the target movement is continuous in the spatial, time, and frequency domains.

We implement FM-Track on the Bela platform [36], connected with one speaker and an array of four MEMS microphones [8] for extensive experiments. Experiment results show that, for two-target tracking, our system is able to achieve a median accuracy of 0.86 cm and 0.11 cm for absolute range and displacement estimates respectively. FM-Track can successfully separate two targets with a spacing as small as 1 cm, outperforming the state-of-the-arts. We show that FM-Track can simultaneously track *four* targets within a range of 3 m. For single-target tracking, the sensing range can be up to 5 m. Finally, we implement FM-Track on the commercial off-the-shelf (COTS) devices, i.e., a smartphone (iPhone 5c [3]) and a smart speaker prototype (MiniDSP UMA-8-SP USB mic array [26]), to demonstrate the feasibility and reliability of FM-Track with several real-life interaction applications, including multi-hand tracking and two-finger tracking.

2 PRELIMINARIES

Prior studies have focused mainly on single-target tracking, whereas the proposed work herein aims to enable fine-grained multi-target

tracking. In this section, before we present our design details, we introduce the preliminaries related to our design.

2.1 Device-free vs. Device-based Tracking

Acoustic tracking can be broadly grouped into two categories: device-based and device-free (contactless) tracking. While device-based tracking tracks a device to indirectly track a target (e.g., hand), device-free tracking employs signal reflections from the target to directly track it. Take hand tracking as the example. For device-based tracking [23, 41], we can track a smartphone hold in hand to track the hand movement. On the other hand, for device-free tracking [28, 45, 53], a smartphone can be placed on a table and the built-in speaker and microphone are leveraged to transmit and receive acoustic signals. As the user moves the hand near the smartphone, the hand movements can be directly tracked by analyzing the acoustic signals reflected from the hand. Compared to device-based tracking, the tracking range of device-free tracking is usually smaller because the reflection signal is weaker. Furthermore, multi-target tracking is particularly challenging for device-free tracking as signals reflected from different targets—that are superimposed at the receiver—are difficult to be separated to track each individual target.

2.2 Multidimensional Information

Each target can be exclusively characterized by its location and motion status with information from three dimensions: time (range), space (angle), and frequency (velocity).

Range information (r): The range—i.e., the distance between the target and the sensing device—can be obtained by measuring the propagation time of the reflected acoustic signal and multiplying it with the signal propagation speed. Fundamentally, the resolution of the range estimation is determined by the signal frequency bandwidth [41, 47]. A larger bandwidth yields a more accurate estimation of signal propagation time and thus more accurate range estimation.

Angle information (θ): The Angle-of-Arrival (AoA) information represents the direction of acoustic signals arriving at the microphone array. The number of microphones determines the resolution of the AoA estimates [48]. More microphones could generate a narrower AoA beam width, resulting in a higher angular resolution.

Velocity information (v): When a target is moving, the signal reflected from the target experiences a frequency shift, which is termed as Doppler shift. This Doppler shift value can be used to calculate the velocity of the moving target. The resolution of Doppler shift is related to the length of the observation time window [11]. A larger window size leads to a finer resolution of the Doppler shift, and thus more accurate velocity estimation.

We can exploit information from multiple dimensions to distinguish multiple targets, as illustrated in Fig. 2. When two targets have similar range values, the AoA information can be leveraged to distinguish them (Fig. 2a). Similarly, two targets that share similar angles can be distinguished by their range information (Fig. 2b). Even when two targets share both similar range and AoA values, they could still be distinguished by tracing their velocities [47] (Fig. 2c).

2.3 Tracking Accuracy vs. Resolvability

The majority of prior efforts in contactless acoustic tracking have focused on improving the tracking accuracy for a single target [28, 34, 45, 53], whereas relatively little attention has been paid to improve the resolvability to distinguish multiple targets. For applications involving a single target, the tracking accuracy should serve as the primary evaluation metric. However, for application involving multiple targets, both the tracking accuracy and resolvability (i.e., the capability to resolve multiple targets) must be considered. Without loss of generality, we consider a bandwidth of $B = 4\ kHz$, which yields a range resolution of $\Delta r = \frac{v_s}{2B} = \frac{340\ m/s}{2\times 4\ kHz} = 4.25\ cm$, where the factor 2 accounts for the fact that the reflected signal traverses the path back and forth. Consider an example where two targets are separated with a distance of D as shown in Fig. 3. When the distance (D=8 cm) between two targets is larger than the resolution ($\Delta r = 4.25\ cm$), we can clearly see two peaks, indicating that these two targets are resolvable. However, when the distance between two targets decreases to 2 cm, which is smaller than Δr, two signal peaks merge into one, and the two targets become unresolvable. This example clearly demonstrates that a coarse resolvability could negatively affect the tracking accuracy, and thus, both the tracking accuracy and resolvability are important metrics in multi-target tracking.

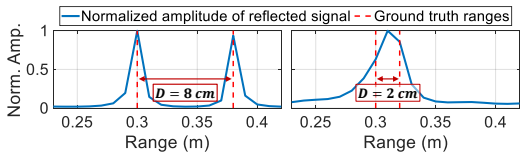


Figure 3: An illustrative example demonstrating that two targets become unresolvable when the range difference between them D is smaller than the range resolution ($\Delta r = 4.25\ cm$).

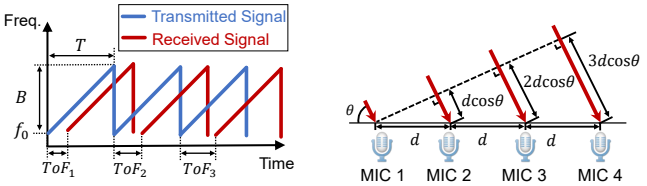
2.4 Relative Tracking vs. Absolute Tracking

Recent efforts have pushed the granularity of contactless acoustic tracking to millimeter level [28, 34, 45, 53]. The underlying mechanism of these tracking methods is to measure the fine-grained phase change. For instance, a phase change of 40° in the acoustic signal with a wavelength of 2 cm (16 kHz) corresponds to a range change of $\frac{2\ cm\times 40^\circ}{2\times 360^\circ} = 1.1\ mm$. Unfortunately, this millimeter-level accuracy is only possible when estimating changes of the target position (i.e., relative tracking) without accounting for where exactly the target is located with respect to the sensing device. The estimate of the absolute distance between the target and the sensing device (i.e., absolute tracking) is much coarser, because its accuracy depends on the bandwidth of the acoustic signals. In theory, with a 4 kHz bandwidth, the absolute tracking accuracy is around 4.25 cm. Wang *et al.* have reported a relative tracking accuracy of 0.35 cm and an absolute tracking accuracy of 3.57 cm [45], matching our analysis herein.

3 A MATHEMATICAL SIGNAL MODEL

This section describes a detailed mathematical derivation of our chirp-based signal modeling that is fundamental to our algorithm designs to track multiple targets. This signal model enables our algorithms to simultaneously extract the range, velocity, and angle information of multiple targets from the reflected signals, which allows us to fully characterize the targets' location and motion status.

As shown in Fig. 4, the core idea behind chirp-based acoustic tracking is to compute the Time-of-Flight (ToF) of the chirp signal



(a) Multiple chirps. (b) Microphones Array.

Figure 4: The range information can be computed from one ToF. (a) The velocity information can be estimated by ToF differences over chirps; (b) the angle information can be estimated by ToF differences over microphones.

transmitted from a speaker by comparing it with its delayed signal reflected by the target. The range information (i.e., the distance between the sensing device and the target) can be estimated by multiplying half of the ToF with the sound speed in the air. The velocity information of target movement can be estimated by measuring the ToF differences across multiple chirps (Fig. 4a). In addition, we can estimate the angle information (i.e., angle of the target location with respect to the sensing device) by measuring the ToF differences across multiple microphones (Fig. 4b). Without loss of generality, we present our signal model by assuming a uniform linear microphone array. The proposed methods can be applied to other microphone arrangements, such as a uniform circular array used in smart speakers [39]. In the evaluation (Sec. 7), we present results for both linear and circular microphone arrays.

3.1 Signal Model for a Single Target

We first explain how the transmitted and reflected signals are processed to derive the ToF of a single target. As shown in Fig. 4a, the speaker transmits a sequence of chirp signals and each transmitted chirp can be represented as

$$S^T(t) = \cos\left(2\pi\left(f_0 t + \frac{B}{2T}t^2\right)\right), \tag{1}$$

where f_0, B, and T represent the start frequency, frequency bandwidth, and duration of the chirp, respectively. For a chirp signal, the signal reflected from a target to the receiver (i.e., a microphone) is a delayed version of the transmitted signal that can be represented as

$$S^R(t) = \alpha\cos\left(2\pi\left(f_0(t-\tau) + \frac{B}{2T}(t-\tau)^2\right)\right) + W(t), \tag{2}$$

where α is the signal amplitude attenuation factor, τ is the ToF and $W(t)$ is the Gaussian white noise. For simplicity, we omit the Gaussian white noise in the following equations.

Fig. 5 summarizes the process to compute the ToF from the received signal. The received signal is multiplied by the transmitted signal $S^T(t)$ and its 90-degree phase-shifted version $S^{T'}(t) = \sin\left(2\pi(f_0 t + \frac{B}{2T}t^2)\right)$ to derive the In-Phase (I) and Quadrature (Q) parts of the mixed signal respectively. Specifically, after applying the product-to-sum identity (i.e., $\cos A \cdot \cos B = \frac{1}{2}\big(\cos(A-B) + \cos(A+B)\big)$) and a low-pass filter, the In-Phase part of the mixed

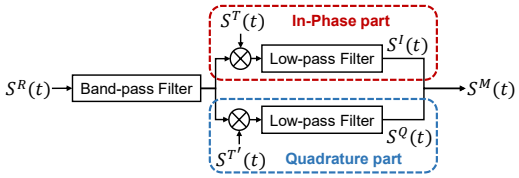


Figure 5: The mixed signal $S^M(t)$ can be constructed from the In-phase part $S^I(t)$ and Quadrature part $S^Q(t)$.

signal becomes

$$\begin{aligned} S^I(t) &= \frac{1}{2}\alpha\cos\left(2\pi\left(f_0\tau + \frac{B}{T}t\tau - \frac{B}{2T}\tau^2\right)\right) \\ &\approx \frac{1}{2}\alpha\cos\left(2\pi\left(f_0 + \frac{B}{T}t\right)\tau\right). \end{aligned} \tag{3}$$

The approximation above is based on the fact that $\frac{B}{2T}\tau^2$ is two orders of magnitude smaller than $f_0\tau$ due to a very small τ value.[1] Similarly, the Quadrature part can be approximated as

$$S^Q(t) \approx \frac{1}{2}\alpha\sin\left(2\pi\left(f_0 + \frac{B}{T}t\right)\tau\right). \tag{4}$$

By combining the obtained I and Q components, we obtain the mixed signal as

$$S^M(t) = S^I(t) + jS^Q(t) = \frac{1}{2}\alpha e^{j2\pi(f_0 + \frac{B}{T}t)\tau}. \tag{5}$$

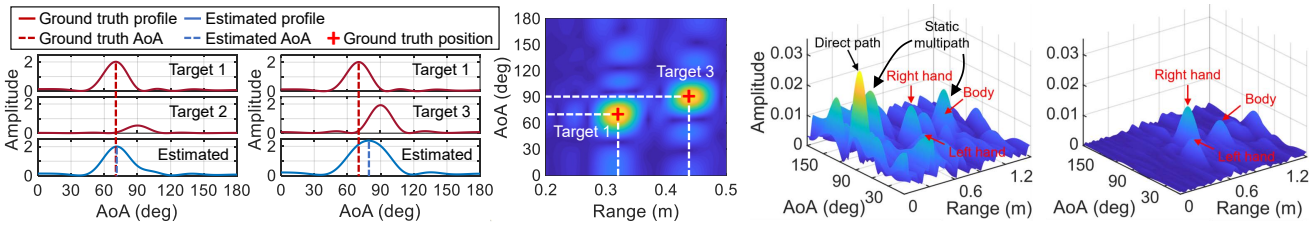
The obtained ToF information could be analyzed to extract the range, velocity, and angle information of the target. Consider a target whose distance with respect to the first microphone of the array is denoted as r. The ToF of the signal received by this microphone can be computed as the round-trip distance divided by the signal speed in air v_s, i.e., $\frac{2r}{v_s}$. Suppose that the target is moving at a radial velocity of v, where the velocity is positive when the target moves away from the microphone array and negative when the target moves closer. During the time period from the first chirp to the c^{th} chirp, the target moves an extra distance of $(c-1)Tv$, and this extra amount of movement would cause an additional round-trip time of $\frac{2(c-1)Tv}{v_s}$. For the k^{th} microphone, as shown in Fig. 4b, the ToF of the received signal would experience an extra propagation time of $\frac{(k-1)d\cos\theta}{v_s}$ compared to the first microphone, where d and θ are the distance between two adjacent microphones and the signal Angle-of-Arrival (AoA),[2] respectively. Therefore, the ToF $\tau_{c,k}$ of the signal received at the k^{th} microphone for the c^{th} chirp can be computed as

$$\tau_{c,k} = \frac{2r}{v_s} + \frac{2(c-1)Tv}{v_s} + \frac{(k-1)d\cos\theta}{v_s}. \tag{6}$$

Note that the velocity v represents the average target movement velocity over multiple chirps. By substituting Equation (6) into Equation (5), our model for the mixed signal can be represented as

$$\begin{aligned} S^M(t_i, c, k) &= \frac{1}{2}\alpha e^{j\varphi(t_i,c,k)} \\ &= \frac{1}{2}\alpha e^{j2\pi(f_0 + \frac{B}{T}t_i)(\frac{2r}{v_s} + \frac{2(c-1)Tv}{v_s} + \frac{(k-1)d\cos\theta}{v_s})}, \end{aligned} \tag{7}$$

[1] Even for a large distance of 3 m, the time-of-flight τ is just 0.0176 s.
[2] The target does not have to be in the same horizontal plane as the speaker/microphones.



(a) The AoA profile. (b) The AoA profile. (c) The AoA-range profile. (a) Before Back-Subtract. (b) After Back-Subtract.

Figure 6: Comparison between pseudo-joint and joint estimation. (a) Pseudo-joint estimation works well when two signals have different strengths; (b) pseudo-joint fails for signals with comparable strength; (c) joint estimation works well in this scenario.

Figure 7: The AoA-range profiles before and after background subtraction (Back-Subtract): the reflections from two hands and body stand out after background subtraction as shown in (b).

where t_i is the i^{th} sampling timestamp and $\varphi(t_i, c, k)$ is the phase change induced by the k^{th} microphone in the c^{th} chirp at the i^{th} sampling timestamp. Equation (7) contains information relevant to range, angle, and velocity of the target, which can be rearranged to simplify the notation as

$$S^M(t_i, c, k; \boldsymbol{p}) = \frac{1}{2}\alpha \cdot R \cdot \Theta \cdot V, \tag{8}$$

where R corresponds to the component related to range, Θ corresponds to the component related to angle, and V corresponds to the component related to velocity. These three components are represented as

$$\begin{aligned} R &= e^{j\varphi_r} = e^{j2\pi(f_0+\frac{B}{T}t_i)\frac{2r}{v_s}} \\ \Theta &= e^{j\varphi_\theta} = e^{j2\pi(f_0+\frac{B}{T}t_i)\frac{(k-1)d\cos\theta}{v_s}} \\ V &= e^{j\varphi_v} = e^{j2\pi(f_0+\frac{B}{T}t_i)\frac{2(c-1)Tv}{v_s}}. \end{aligned} \tag{9}$$

In Equation (8), we have four unknown parameters, which are the angle θ, range r, velocity v, and the signal attenuation α. We denote these parameters as a parameter vector $\boldsymbol{p} = [\theta, r, v, \alpha]$, which characterizes the location and motion status of one target. R, Θ and V are the steering vectors for range, angle, and velocity, respectively [16].

The above-stated chirp-based acoustic signal model in Equation (8) is, to the best of our knowledge, the first attempt to simultaneously extract the target's range, angle, and velocity information. This model uniquely supports the chirp-based signals, as opposed to the state-of-the-art model proposed for WiFi signals [47]. As we can observe from Equation (9), all three steering vectors in the signal model are time-varying, whereas those in WiFi-based models are time-invariant, which makes it difficult to apply the techniques used in WiFi tracking to acoustic tracking. Specifically, in WiFi-based models, the range information can be estimated from a single time-domain sample, since one time-domain OFDM WiFi sample contains information of the entire frequency band (64 sub-carriers). On the other hand, the chirp-based model requires a series of samples across one chirp, because one time-domain sample only contains information of a single frequency.

3.2 Signal Model for Multiple Targets

The signal model for a single target in Equation (8) can be extended to multiple targets. In the presence of L targets, the mixed signal at the k^{th} microphone for the c^{th} chirp can be viewed as a superposition of signals from L targets

$$S(t_i, c, k) = \sum_{l=1}^{L} S_l^M(t_i, c, k; \boldsymbol{p}_l). \tag{10}$$

The ultimate goal of our tracking algorithm—as we will present in Sec. 4—is to separate the mixed signals reflected from multiple targets and estimate the corresponding parameters $\boldsymbol{p}_l = [\theta_l, r_l, v_l, \alpha_l]$ for each individual signal. From the parameters, we can obtain the location and motion information of each target.

4 RESOLVING MULTIPLE TARGETS

This section describes how FM-Track enables multi-target tracking. The proposed algorithms provide an unprecedented opportunity to improve the signal resolvability (i.e., the minimum distance between two closely located targets which the reflected signals can still be resolved) by leveraging a novel approach to enable joint estimation of the range, angle and velocity. Before presenting our algorithms in-depth, we briefly discuss why the joint estimation algorithms designed for WiFi-based [47] and radar-based multi-target tracking [4, 10] cannot be effectively used for acoustic signals.

4.1 Discussion on Joint Estimation

One major issue rooted in multi-dimensional joint estimation is the high computational complexity [10, 47]. The computational cost increases exponentially with the increasing number of information dimensions. To avoid a high computational cost, mD-Track [47] develops a pseudo-joint estimation, which estimates the parameters on each dimension *sequentially*. It is noteworthy that, if the previously estimated parameter is inaccurate, the accuracy of the current parameter estimation is also affected. The basic idea of mD-Track is to iterate the process of estimating the strongest signal while considering the rest of the signals as noise. A key assumption for this approach to work well is that, in each iteration, there exists one signal with the prominent signal strength compared to other signals, which does not hold true when two targets are very close to each other (e.g., two hands or fingers). To illustrate this issue, we conduct the following experiment. We first place two finger-sized cardboards (Target 1 and 2) at $[0.32\ m, 70°]$ and $[0.44\ m, 90°]$, respectively. Because they have different ranges from the sensing device, the signal reflected from Target 1 is much stronger than that from Target 2. In this scenario, mD-Track would provide an

accurate result, as shown in Fig. 6a. Now we replace Target 2 with a larger cardboard Target 3. Due to its larger size, the strength of its reflected signal is now comparable to that of Target 1, and the above-mentioned assumption does not hold anymore. We can now observe the estimated angle significantly deviates from the ground truth in Fig. 6b.

Fortunately, joint estimation can effectively address the issue of interference among signals with comparable strengths. As shown in Fig. 6c, with an additional dimension (i.e., range), two signals can be clearly distinguished even though the angles are close to each other. Traditional radar-based tracking systems employ the MUSIC [32] algorithm to perform joint estimation, which contains two computationally expensive steps: eigenvalue decomposition [25] and joint optimal parameter searching [10]. In this work, we design a lightweight joint estimation algorithm (Sec. 4.2) integrated with several computation-saving techniques (Sec. 4.3). From our experiments, the median run time of the proposed 3D (i.e., range, velocity, and angle) joint estimation algorithm is only 0.0255 s for a single target, while that of the traditional MUSIC 3D is around 195 s [10].

4.2 Separating Signals from Multiple Targets

This section presents a detailed description of our estimation algorithm, which is designed based on the characteristics of chirp-based signals. We first present how to jointly estimate the parameters for a single target and then expand the algorithm for multiple targets.

4.2.1 Background Multipath Subtraction. Before estimating the multi-dimensional parameters, we first remove the background multipath, including the direct path from the speaker to microphones and reflections from static surroundings. As shown in Fig. 7a, the reflections from a human target (body and hands) are overwhelmed by the background multipath, which would significantly decrease the tracking accuracy. Therefore, we first measure the background signal when there is no target and remove it later [24, 27]. After subtracting the background multipath, the reflected signals from the hands and the body clearly stand out, as shown in Fig. 7b.

4.2.2 Single Path Estimation. Assuming there is only one signal, the estimation process can be decomposed into three steps: 1) constructing the joint estimator; 2) searching the optimal parameters for AoA, range, and velocity; and 3) computing the signal attenuation.

Overview of the joint estimator for all parameters. For each signal sample received by the k^{th} microphone in the c^{th} chirp at the i^{th} sampling timestamp, the signal is modeled by the attenuation factor α and the phase change $\varphi(t_i, c, k) = \varphi_r + \varphi_\theta + \varphi_v$ induced by the range r, AoA θ, and velocity v respectively, as shown in Equation (7). For simplicity, we generically use a symbol φ to represent $\varphi(t_i, c, k)$ hereafter. The key idea for our joint estimator is that, if θ, r, and v are correctly estimated, the phase change computed based on the estimated parameters (i.e., $\hat{\varphi}$) and the measured value of the actual phase change (i.e., φ_m) will be approximately equal

$$\varphi_m = \hat{\varphi_r} + \hat{\varphi_\theta} + \hat{\varphi_v}. \tag{11}$$

If we remove these accurately estimated phase changes $(\hat{\varphi_r}, \hat{\varphi_\theta}, \hat{\varphi_v})$ from the measured phase across each sample, a signal with a phase value of 0 (i.e., $\frac{1}{2}\alpha e^{j0}$) is resulted. This implies that the phase-removed signals will be in-phase and combined constructively, and thus, the strength of the superposed signal is maximized.

Constructing the joint estimator. Suppose that a chirp contains N samples, and a total number of C chirps is included for one round of estimate. At each sample index i, we have signal samples from K microphones. Then, we can represent the signal samples as $\Sigma = [\Sigma_1\ \Sigma_2\ \cdots\ \Sigma_N]$, where each Σ_i, $i \in [1, N]$ is a matrix of size $K \times C$. For each possible AoA θ, velocity v, and range r, the phase change induced by them could be removed from Σ_i by multiplying the conjugates of their steering vectors defined in Equation (9):

$$\mathbf{E}_i(\theta, v, r) = \Theta_i^*(\theta)\Sigma_i V_i^*(v) R_i^*(r), \tag{12}$$

where $(\cdot)^*$ is the conjugate operation, $\Theta_i(\theta)$ is a $1\times K$ vector, $V_i(v)$ is a $C\times 1$ vector, and $R_i(r)$ is a scalar. $\Theta_i(\theta)$, $V_i(v)$ and $R_i(r)$ correspond to the steering vectors of $\Theta(\theta)$, $V(v)$, and $R(r)$ respectively, for the i^{th} sample. After eliminating the phase change on each dimension for each sample, we obtain the joint estimator by summing all these phase-removed samples together

$$\mathbf{E}(\theta, v, r) = \sum_{i=1}^{N} \mathbf{E}_i(\theta, v, r). \tag{13}$$

Searching for the optimal parameters. The output of our joint estimator $\mathbf{E}(\theta, v, r)$ will be maximized at the optimal parameters $\hat{\theta}$, $\hat{v}$, and $\hat{r}$, because all signals are in-phase and combined constructively. Hence, the optimization problem can be formulated as

$$(\hat{\theta}, \hat{v}, \hat{r}) = \arg\max_{\theta, v, r} \|\mathbf{E}(\theta, v, r)\|^2. \tag{14}$$

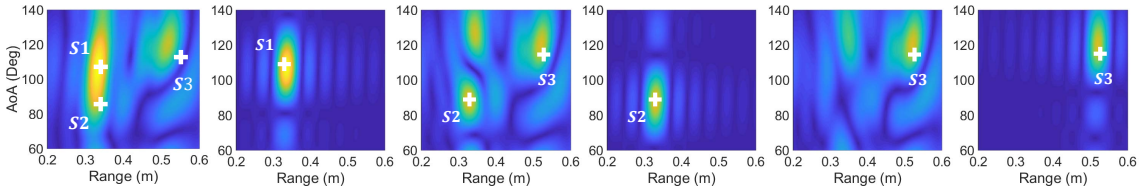
The search range of the parameters can be defined based on the application. For hand/finger motion tracking, we can define the search range of these parameters according to the physical constraints of hand/finger movement. Then, we perform a search over the three dimensions to find the optimal parameters. We present our schemes to reduce the computational cost in Sec. 4.3.

Computing the attenuation factor $\hat{\alpha}$. After obtaining the estimates for all three parameters $(\hat{\theta}, \hat{v}, \hat{r})$, we then calculate $\hat{\alpha}$ by substituting those estimates into Equation (7)

$$\hat{\alpha} = \frac{2}{C \cdot K \cdot N} \mathbf{E}(\hat{\theta}, \hat{v}, \hat{r}). \tag{15}$$

The final outcomes of the above-mentioned algorithm represent all dimensions of path parameters associated with one target. It is noteworthy that the proposed algorithm can be flexibly configured to support lower dimensions. For example, although a smartphone with only one microphone cannot support the spatial dimension (i.e., angle information), the proposed algorithm can be configured to leverage the velocity and range dimensions for multi-target tracking.

4.2.3 Multiple Signal Estimation. The proposed algorithm to estimate parameters for one signal could be extended to work for multiple signals. The algorithm is first applied to estimate the parameters $\hat{\theta}$, $\hat{v}$, $\hat{r}$, and $\hat{\alpha}$ for the strongest signal among the mixed signals. Then, we reconstruct the strongest signal based on Equation (7) and subtract this estimated signal from the mixed signal. Then, the algorithm is again applied to estimate the next strongest signal until the power of the residual signal is smaller than a pre-defined threshold. Fig. 8 demonstrates the process of estimating parameters for three signals reflected from three targets, namely $S1$, $S2$, and $S3$.



(a) Original signal. (b) Reconstructed $S1$. (c) $S1$ removed. (d) Reconstructed $S2$. (e) $S2$ removed. (f) Reconstructed $S3$.

Figure 8: An illustrative example of reconstruction and subtraction for signals reflected from three targets.

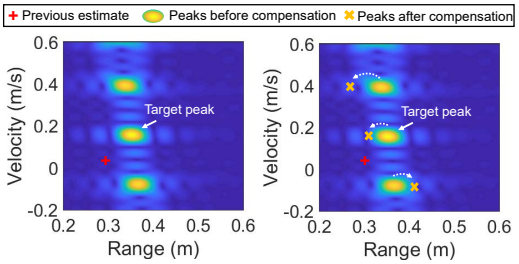
4.2.4 More Iterations. The estimated parameters could be further improved by iterating the above-mentioned algorithms. The major source of the error emerges from the inaccurate determination of background noise when the algorithm sequentially estimates the parameters for each target. For example, when the algorithm first identifies the strongest signal in the mixed signals, it considers the residual signals (i.e., mixed signals subtracted by the strongest signal) as the background noise. In contrast, the residual signals are actually composed of the weaker reflected signals from other targets and the (actual) background noise. Fortunately, the ultimate residual signal after all the targets are identified could more accurately represent the actual background noise. Thus, we iterate the entire signal estimation algorithm by adopting the estimated noise from the previous iteration to represent the background noise in the current iteration. This iterative process is repeated until the identified parameters for all targets in the current iteration do not differ from the previous iteration by a pre-defined threshold. The average number of iterations in our experiment is quite small (i.e., 2.3 iterations).

4.3 Reducing Computational Cost

The high computational cost associated with joint estimation makes it challenging to support real-time tracking. Considering K microphones, C chirps, and N samples within each chirp, the computational cost of a single execution of the estimator is $O(\eta) = O(n_\theta \cdot n_v \cdot n_r \cdot K \cdot C \cdot N)$, where n_θ, n_v and n_r are the numbers of searching steps for the three parameters, respectively. If we further assume we have L paths and N_{iter} iterations, the computational cost becomes $O(N_{iter} \cdot L \cdot n_\theta \cdot n_v \cdot n_r \cdot K \cdot C \cdot N)$.

Reducing the size of search window. From the above analysis, the number of search steps is a key factor affecting the computational cost and is determined by the size of search window and the search step size. Due to the relatively slow speed of human hand/finger movements, small search windows for all three dimensions suffice. Take hand tracking as an example, we empirically choose a search window of 60 cm for range and 60° for AoA. The search window for velocity is from -60 cm/s to 60 cm/s. For finger tracking, we can use even smaller search windows. The search step sizes can be chosen to balance the desired accuracy and computational efficiency.

Subsampling signal samples in time domain. The number of samples (N) in each chirp is two orders of magnitude larger than the number of microphones (K) and the number of chirps (C). Therefore, to improve the computational efficiency, we choose to subsample the mixed signal S^M defined in Equation (7) in the time domain by a factor of D. There is a trade-off between reducing the computational cost and maintaining a high accuracy. That is, choosing a large D can reduce the computation cost, but also decreases the number of samples in estimating parameters, thus degrading the accuracy. We empirically set $D = 40$ based on the trade-off analysis.



(a) Multiple peaks. (b) Picking out the target peak.

Figure 9: Velocity ambiguity: (a) there exists one target peak and multiple ambiguous peaks; (b) identifying the target peak by matching the previous estimate (red plus) with the peaks after compensation (yellow crosses).

Accurate starting position estimate. An accurate starting position estimate can significantly reduce the computational cost. The range estimate of the starting position is particularly accurate for acoustic signals due to the relatively slow propagation speed in air. Also at the beginning, targets usually initiate movements from a static position (i.e., velocity is 0). In this case, we can perform estimation with information from two dimensions (i.e., range and AoA) rather than three dimensions to quickly obtain the starting positions of targets.[3] Then we immediately switch to 3-dimensional estimation and limit the search space by applying the continuity property of the target movement. Note that users can perform a simple yet unique initial gesture (e.g., push the hand twice) to differentiate the targets from the interferers [24].

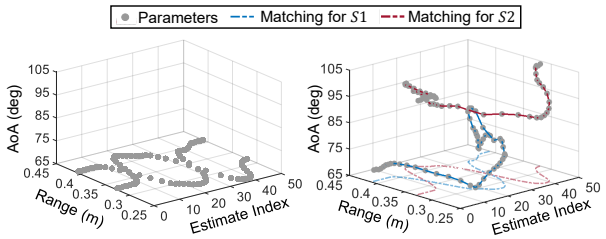
5 MULTI-TARGET TRACKING

This section first introduces how to deal with the velocity ambiguity issue, followed by matching two consecutive estimates for the same target. At last, we propose a novel method to compute the instantaneous velocity and compensate the range-Doppler effect.

5.1 Dealing with Velocity Ambiguity

There is an intrinsic ambiguity issue in estimating the velocity of a moving target using chirp-based signals [30]. Specifically, there exist multiple peaks, including the target peak and several ambiguous peaks, as shown in Fig. 9a. Due to noise and multipath, the peak with the largest amplitude may not correspond to the target peak,

[3]Two-dimensional estimation takes just a few milliseconds as shown in Sec. 7.4.



(a) Matching using 1D parameters. (b) Matching using 2D parameters.

Figure 10: Signal-target matching for signals $S1$ and $S2$: (a) it is hard to match signals using only range information, (b) but much easier using both range and angle information.

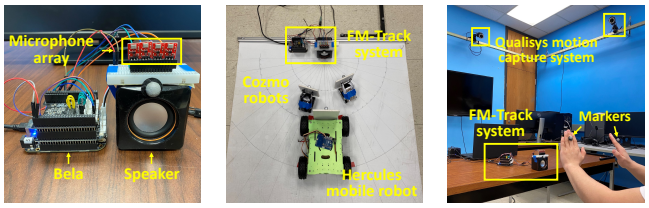
making it difficult to identify the target peak by amplitude. We propose to apply the continuity property among adjacent estimates to identify the target peak. The basic idea is that within a short period of time (one chirp period is 40 ms), the velocity and range variations are small. Thus we can include the estimate from the previous round to help identify the target peak in the current round by choosing the peak that has the smallest distance from the previous estimate. However, this approach may not work effectively due to the range-Doppler deviation induced by the target movement. We thus first compensate the raw peak estimates to remove the range deviation before we apply the continuity property. In Fig. 9a, without compensation, two peaks have similar distances to the previous round estimate (marked as a red plus). After compensating the range-Doppler effect (detailed in Sec. 5.3), the compensated target peak (marked as a yellow cross) is much closer to the previous estimate compared to the other two ambiguity peaks, as shown in Fig. 9b.

5.2 Matching Signals with Targets

To track moving targets, we need to continuously estimate the parameters of signals from these targets. At each timestamp, our joint estimator outputs a collection of parameters for multiple targets. Then, one challenge naturally appears: how to associate the parameter estimates for the same target at different timestamps? The problem becomes particularly challenging when multiple targets have overlapping trajectories. The key idea of our signal-to-target matching solution is that the target movement is continuous (particularly for human movements), and thus, the parameters between two consecutive estimates from the same target should not differ by a large magnitude. However, when two targets have overlapping trajectories, the parameters of the two targets become similar, causing confusion in tracking, as shown in Fig. 10a. We propose to employ parameters from more dimensions to address this issue, because it is unlikely multiple targets have similar parameters in all dimensions at a given timestamp. For example, Fig. 10b shows that the ambiguity in Fig. 10a could be resolved in higher dimensional space (i.e., 2D space of range and AoA). To quantify the overlap in the trajectories of different targets, we use a weighted L1-norm

$$\sum_{i=1}^{4} \boldsymbol{\beta}[i] \cdot \left| \boldsymbol{p}^t[i] - \boldsymbol{p}^{t-1}[i] \right|, \tag{16}$$

where $\boldsymbol{p}^t = [\theta^t, r^t, v^t, \alpha^t]$ represents the parameter estimate at time t, and $i \in [1, 4]$ represents the index for the four parameters. $\boldsymbol{\beta}$ is a scale vector that normalizes the value ranges of the four parameters into the same scale, which can be determined in advance. The pair of two consecutive estimates with minimum L1 distance is chosen as the best match for each signal. However, this approach requires a computation cost of $O(n!)$ for n signals to search the optimal pairs. We thus employ the Hungarian algorithm [14, 19] to reduce the computational complexity from $O(n!)$ to polynomial time.



(a) FM-Track system. (b) Robot scenarios. (c) Real-life scenarios.

Figure 11: Experiment setup.

5.3 Refining Parameter Estimates

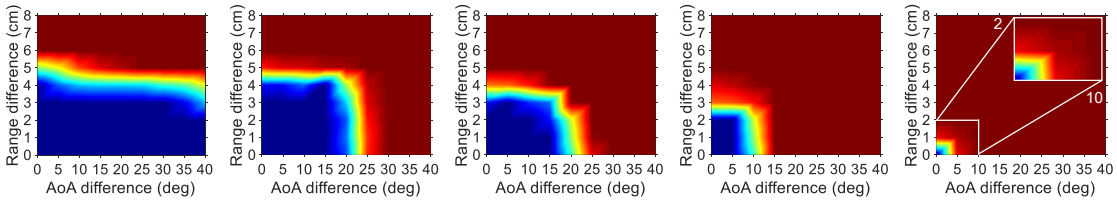
It is well-known that the chirp-based methods have difficulties in obtaining accurate instantaneous velocity [11]. If the instantaneous velocity of a moving target could be estimated for each chirp, we can achieve a more fine-grained estimation of the motion status of a moving target. In this section, we introduce a novel method to obtain instantaneous velocity by exploiting the fact that the information estimated from multiple dimensions are not equally accurate. Specifically, we capitalize on the range estimates, which is more accurate compared to other estimates due to the low propagation speed of acoustic signals, to estimate the instantaneous velocity. Once the instantaneous velocity is obtained, we can refine the range deviation with a more accurate instantaneous velocity.

When the target is stationary, the estimated range r_1 has no deviation. At the next timestamp when the target starts to move, the target's location is changed and thus, the estimated range r_2 is deviated due to the range-Doppler effect [24]. Compared with the range estimate r_1 in the previous timestamp, the new range estimate r_2 can be expressed by adding the target displacement vT and also the deviation caused by the non-zero target velocity $\frac{vf_0T}{B}$ to r_1: $r_2 = r_1 + vT + \frac{vf_0T}{B}$, where v is the velocity of the moving target. Since both r_1 and r_2 can be obtained from range estimates, we can derive the velocity v from the above equation. Because velocity v is obtained using two range estimates from two adjacent chirps (i.e., 40 ms apart), we can consider it as the instantaneous velocity. With the instantaneous velocity, we can calculate the movement-caused range deviation and remove the deviation from the range estimate.

6 IMPLEMENTATION

We implement FM-Track on both the research-purpose Bela platform [36] and COTS devices. The signals are analyzed in the MATLAB environment using a laptop equipped with an Intel i7 processor.

Bela platform: The Bela platform [36] is widely used for research involving acoustic signals owing to its ability to flexibly support different numbers and locations of microphones and speakers. The Bela hardware is connected with one general-purpose EARISE AL-202 speaker [7] to transmit acoustic signals and an array of



(a) MUSIC 1D (r). (b) MUSIC 2D (r+θ). (c) FM-Track 2D (r+θ). (d) mD-Track 3D (r+θ+v). (e) FM-Track 3D (r+θ+v).

Figure 12: Resolvability comparison among different approaches. 1D, 2D, and 3D represent the number of information dimensions adopted in parameter estimation. The color indicates the probability of resolving two targets: red indicates *fully resolvable* and blue means *non-resolvable*. The algorithm with the smaller blue area can achieve better resolvability.

up-to four SparkFun ADMP401 MEMS microphones [8] to receive acoustic signals. The device components are shown in Fig. 11a.

COTS devices: Without loss of generality, we adopt iPhone 5c [3] to verify the effectiveness of FM-Track. To simplify the process of development, the system is implemented based on the existing framework LibAS [40], which allows us to develop our sensing algorithm using MATLAB on the laptop without considering the smartphone-specific details. We further implement our system on a smart speaker prototype (i.e., MiniDSP UMA-8-SP USB mic array [26]), which is equipped with seven Knowles SPH1668LM4H microphones. This prototype has the identical layout and the same sensitivity as Amazon Echo Dot [2], a widely used smart speaker.

Acoustic signals: The default chirp signals adopted in our implementation lie in the inaudible frequency band from 18 kHz to 22 kHz with a chirp duration of 40 ms. The sampling rate of the Bela platform and COTS devices are 44.1 kHz and 48 kHz, respectively.

Robots: To enable controlled experiments with a target moving at a specific velocity and following a pre-defined trajectory, we mount cardboards on two different types of robots as shown in Fig. 11b: 1) Cozmo [37] that can be controlled at a speed granularity of 1 mm/s with a maximum speed of 20 cm/s and 2) 4WD Hercules Robot [38] with a speed granularity of 2 cm/s and a maximum speed of 2 m/s.

Ground truth measurements: As shown in Fig. 11c, we employ an optoelectronic motion capture system (i.e., Qualisys [13]) that supports sub-mm-level multi-target motion tracking with a frame rate of 250 Hz to obtain the ground truths of the target movements.

7 EVALUATION

We comprehensively evaluate the performance of FM-Track with the Bela platform by varying the parameters under different conditions. For experiments involving movements, we mount the targets on mobile robots to precisely control their movements. We further conduct a series of field studies by tracking hands and fingers using the Bela platform to demonstrate the feasibility of FM-Track in real-life scenarios. In addition, we showcase two interaction applications using a smartphone and a smart speaker. Unless otherwise specified, we apply three-dimensional information (i.e., FM-Track 3D) to track two targets using four microphones and four chirps with a bandwidth of 4 kHz. Each experiment is repeated 20 times.

Evaluation metrics: We report both resolvability and tracking accuracy in our evaluation. For resolvability, we compute the probability that two targets are resolvable in a total of 20 experiments. On the other hand, we quantify the tracking accuracy using four different metrics: absolute range error, absolute AoA error, relative range (displacement) error, and position error. The absolute range error is the difference between the estimated and ground truth distance between the target and the sensing device. Similarly, we define the absolute AoA error. Unless otherwise specified, we employ the absolute range and AoA error to quantify the tracking performance. We also present the relative range error, which is defined as the difference between the estimated and ground truth displacements. In the end, we convert the range-AoA position into the Cartesian (X-Y) coordinate and define a new metric, i.e., position error. The position error is the Euclidean distance between the estimated position and ground truth measured by the motion capture system.

7.1 Overall Performance

In this section, to manifest the performance of FM-Track, we compare the capability of resolving multiple targets, as well as the tracking accuracy between FM-Track and the state-of-the-art approaches, such as mD-Track [47] and MUSIC [18, 23, 32]. The system proposed by Mao *et al.* leveraged 2D MUSIC to estimate the range and AoA, and compensated the range-Doppler effect using a Recurrent Neural Network (RNN) [24]. Because we do not have their data set to train the RNN, we emulate their performance by performing estimation using 2D-MUSIC and compensating the range-Doppler effect using the ground-truth instantaneous velocity captured by the motion capture system for a fair comparison.

Capability of resolving multiple targets. The key to enabling multi-target tracking is the capability of separating signals reflected from different targets. We demonstrate that FM-Track outperforms prior studies in resolving signals from two close-by targets. We perform experiments by varying the range, AoA, and velocity difference between two finger-sized cardboards. The range difference is varied from 0 to 8 cm at a step size of 1 cm, AoA from 0° to 40° at a step size of 5°, and velocity from 0 to 5 cm/s at a step size of 1 cm/s. Specifically, we vary the starting positions of the two targets to provide different ranges and AoAs. To have different velocities, we keep one target static and control the velocity of the other target with the help of the robots. As presented in Sec. 2.3, when two signals are too close to each other, they will not be resolved as two separate signals but only one merged signal. Fig. 12 depicts the resolvability comparison among different approaches. Specifically, cooler (blue) colors indicate "non-resolvable" and warmer (red) colors denote "resolvable". Thus, smaller blue regions indicate better resolvability performance. We observe that, with information from

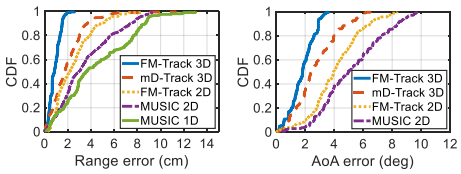


Figure 13: Absolute range and AoA error.

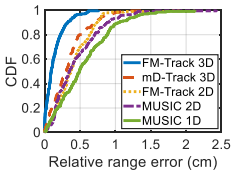


Figure 14: Relative range error.

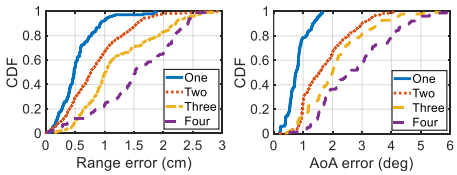


Figure 15: Impact of target numbers.

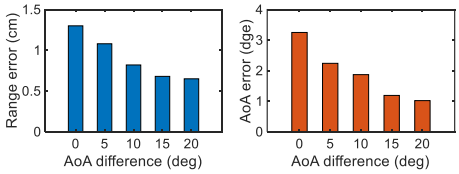


Figure 16: Impact of angle difference.

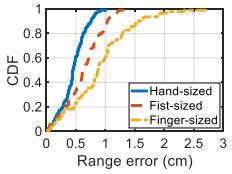


Figure 17: Impact of target size.

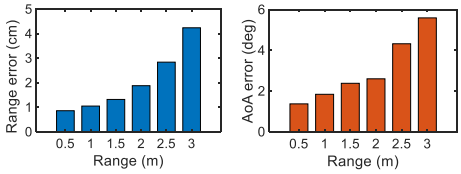


Figure 18: Impact of different ranges.

three dimensions (i.e., range, velocity and AoA), FM-Track 3D can resolve signals reflected by two targets as close as 1 cm, outperforming the state-of-the-art mD-Track 3D by 200%. This granularity is fine enough to track two close-by fingers.

Tracking accuracy. We compare the tracking accuracy of FM-Track with two state-of-the-art systems. We mount two pieces of finger-sized cardboard on two robots, one of which is kept stationary at position $[0.30\ m, 70°]$. We make the second robot move towards the sensing device from the initial position at $[0.38\ m, 70°]$. We also vary the initial position of the second target by changing the angle from 70° to 90° at a step size of 5°. For each movement, we vary the robot's maximum speed from 5 cm/s to 20 cm/s at a step size of 5 cm/s. For a fair comparison, we compensate the range deviation for other approaches using the ground truth instantaneous velocity.

Absolute range and AoA accuracy. Fig. 13 shows the absolute tracking errors for range and AoA, respectively. The median range error for FM-Track 3D[4] is 0.86 cm, outperforming mD-Track 3D by approximately 100%. The median AoA errors achieved by FM-Track 3D and mD-Track 3D are 1.82° and 2.45°, respectively. The proposed system outperforms mD-Track 3D mainly due to its capability of resolving signals with comparable strengths. These results indicate that the increased number of dimensionality can improve not only the resolvability but also the accuracy of estimation in each individual dimension.

Relative range (displacement) accuracy. We compare the relative range errors of different systems in Fig. 14. The median relative range error for FM-Track 3D is as small as 0.11 cm, outperforming mD-Track 3D and MUSIC 2D by 160% and 220%, respectively. Note that the relative range error is much smaller than the absolute range error depicted in Fig. 13, which supports our discussion of relative vs. absolute ranges in Sec. 2.4. Our system is able to achieve millimeter tracking accuracy for both relative and absolute ranges.

7.2 Factors Affecting the Performance

In this section, we evaluate different factors affecting the capability to resolve multiple targets and the overall tracking accuracy.

Impact of number of microphones, number of chirps, and bandwidth size. We increase the number of microphones from two to four. With more microphones, we can improve the resolution of AoA estimation, thereby achieving higher overall resolvability. Owning to high dimensionality, even with just two microphones, the achieved range resolvability is as small as 3 cm, which means two close-by targets separated by just 3 cm could still be resolved and accurately tracked. With four microphones, the resolvability is further improved to 1 cm. Similarly, with more chirps, we can improve the resolution of velocity and obtain higher overall resolvability. However, when we increase the number of chirps beyond four, the improvement is only marginal. We further study the impact by varying the bandwidth size from 2 kHz to 6 kHz. With a larger bandwidth, we can improve the resolution of range and thus, achieve better resolvability. Interestingly, the improvement is marginal when we increase the bandwidth beyond 4 kHz. The above results collectively show that we can improve the resolvability of multiple targets by either improving the resolution of each single dimension or increasing the dimensionality (number of dimensions) of the signals. Due to protocol limit (e.g., a fixed bandwidth size), it is more convenient to increase the number of dimensions.

Impact of number of targets. We employ finger-sized cardboards as targets and mount each cardboard on a Cozmo robot. We increase the number of targets from one to four. For more than two targets, one robot is kept stationary and the rest move at different velocities. Fig. 15 depicts the absolute range and AoA errors for different numbers of targets. We observe that increasing the number of targets also increases the errors. However, even with four targets, FM-Track can still achieve an accuracy of 1.52 cm and 2.5° for range and AoA estimation, respectively.

Impact of angle difference between two targets. We evaluate the performance of FM-Track when the angle difference between two targets becomes small. As shown in Fig. 16, when the angles of the two targets become similar, both the range and AoA errors increase mainly because the interference between them becomes more severe. However, even when the angle difference between two targets is approaching 0°, which means the two small targets are located side by side, the median range error and AoA error are still acceptably small (i.e., 1.3 cm and 3.2°).

Impact of target size. We evaluate the impact of target size by adopting hand-sized, fist-sized and finger-sized cardboards as targets respectively. As we can observe from Fig. 17, the tracking accuracy

[4] 3D indicates that information from 3 dimensions is utilized for tracking.

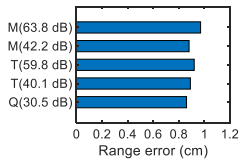


Figure 19: Impact of ambient noises.

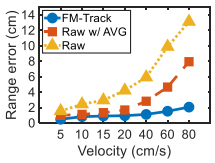


Figure 20: Parameter refinement.

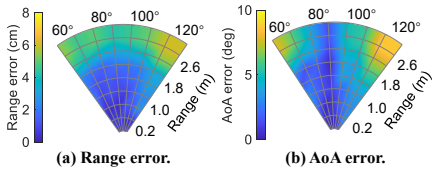


Figure 21: Starting position error.

Table 1: Processing time for each component (unit: *ms*).

Pre-processing	Parameter estimation	Signal-target matching	Parameter refinement	Total
13.4	25.5	1.8	0.2	40.9

is higher when the target size gets larger. The reason is that stronger signals are reflected from targets of larger sizes and result in higher Signal-to-Noise Ratio (SNR). Therefore, higher tracking accuracy can be achieved. However, we do notice that when the target size reaches a threshold, the performance improvement gets saturated. This threshold varies with target material and target-device distance.

Impact of different ranges. To enable long-range multi-target tracking, we adopt the idea of middle chirps proposed in RTrack [24]. Fig. 18 shows the performance of tracking two hand-sized targets at varying distances. We can observe that the performance degrades as the distance between targets and sensing device increases, which is expected due to the lower SNR. When the target is 3 m away, the range error increases to around 4 cm. Note that the displacement error is much smaller which is 0.82 cm. In fact, we can still track the target when the target is 5 m away, but the median tracking accuracy increases to around 8 cm.

Impact of ambient noises. We evaluate the tracking performance under different ambient noises. Two types of noises (i.e., talk and music) are considered, and two different volume levels are tested for each noise. The noise source is placed 0.5 m away from the sensing device. Fig. 19 shows the tracking accuracies for five different scenarios where Q represents sensing with no created noise, T represents sensing with people talking, and M represents sensing with MUSIC playing. We found that the achieved accuracies are very similar which means the ambient noise has little effect on the sensing performance. It is not surprising since the frequency we adopt in FM-Track is much higher than that of the ambient noises.

Effectiveness of parameter refinement. We demonstrate the benefit of the parameter refinement method presented in Sec. 5.3. We compare the results obtained with and without parameter refinement. More specifically, we have two refinement schemes: 1) refinement with the instantaneous velocity proposed in FM-Track and 2) refinement with the traditional average velocity. As shown in Fig. 20, we can clearly observe that FM-Track achieves much better performance compared to the other two methods.

7.3 Starting Position Estimation

An accurate estimation of the starting position of the target is critical for the tracking afterwards. Many existing studies only track the relative motion of the target without knowing its starting position. In this section, we evaluate the accuracy of the starting position estimate. We assume the target has zero velocity at the start position. We evaluate the performance by varying the range and AoA of the target (a hand-sized cardboard) with respect to the sensing device. The range is increased from 0.2 m to 3 m at a step size of 0.4 m, and the AoA is varied from 60° to 120° at a step size of 10°. Fig. 21 shows the starting position estimation accuracy at different positions after background subtraction. The median accuracy for the range and AoA are 1.48 cm and 2.91°, respectively. We observe that the performance degrades when the target moves out of the above-mentioned sector region. The sector width depends on the frequency-dependent radiation pattern of the speaker [12].

Table 2: Comparison of parameter estimation time (unit: *ms*).

Dimensions	MUSIC	mD-Track	**FM-Track**
1D	17.5	1.1	**1.2**
2D	520	2.3	**4.7**
3D	195000.2	3.9	**25.5**

7.4 System Latency

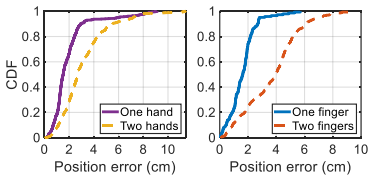
Table 1 shows the processing time of each component in FM-Track for tracking one target. The end-to-end latency is around 40.9 ms (i.e., a maximum update rate of 24.4 Hz), which is similar to the latest work [24]. Even for four targets, the total end-to-end latency is 130 ms, which is sufficiently short to support real-time tracking.

Comparison among different approaches. We set the search step size for range, AoA, and velocity as 1 cm, 1°, and 2 cm/s respectively for all the approaches and evaluate the processing time of parameter estimation. As shown in Table 2, FM-Track runs much faster than MUSIC. Compared with mD-Track which adopts pseudo-joint estimation, our processing time is slightly higher but still good enough for real-time tracking.

7.5 Hand/Finger Tracking with Bela

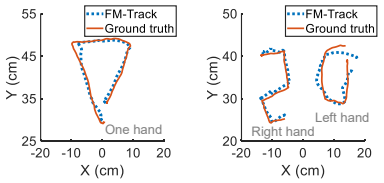
Hand Tracking. We evaluate the performance of FM-Track in tracking hands. A volunteer was asked to sit 0.8 m away from the acoustic device. The volunteer was asked to draw different shapes and Arabic numbers with her hands. Fig. 22a shows the range and position errors for both one hand (single target) and two hands (multiple targets). We observe that the position errors for one hand and two hands are 1.51 cm and 2.65 cm, respectively. To visualize the performance, we show two drawing samples using one hand and two hands in Fig. 23.

Finger Tracking. We also evaluate the performance of FM-Track in finger tracking. Similar to hand tracking, a volunteer was asked to draw shapes with the index finger on the horizontal plane. Then, the volunteer was asked to use the two index fingers to simultaneously draw two shapes. We can see from Fig. 22b, with a very small reflection area, the accuracies are still as high as 1.63 cm and 3.94 cm for one finger and two fingers, respectively. We show two drawing samples using one finger and two fingers in Fig. 24.



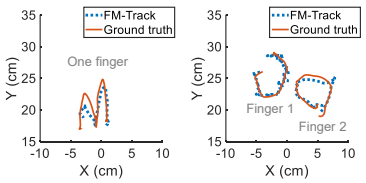
(a) Hand tracking. (b) Finger tracking.

Figure 22: Real-life performance.



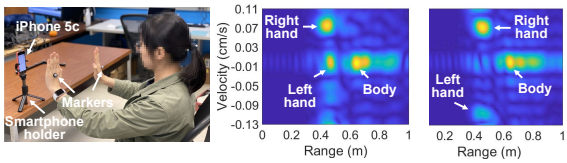
(a) Triangle. (b) Numbers "20".

Figure 23: Hand drawing samples.



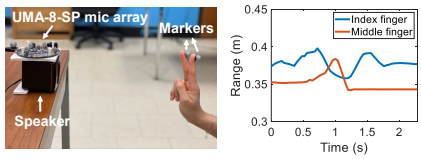
(a) Letter "M". (b) Circles.

Figure 24: Finger drawing samples.



(a) Setup. (b) Snapshot one. (c) Snapshot two.

Figure 25: Two-hand tracking using the smartphone.



(a) Setup. (b) Finger trajectory.

Figure 26: Two-finger tracking using the smart speaker.

7.6 Hand/Finger Tracking with COTS Devices

Tracking two hands using a smartphone. We implement FM-Track on iPhone 5c [3] with one pair of speaker/microphone to transmit and receive acoustic signals for sensing. Note that, because only one microphone is used here, spatial domain information is not available, and thus, we only employ information from the other two dimensions for sensing. This is a good example to show that the proposed system can still track multiple targets without having an array of microphones. A volunteer was asked to sit in front of the smartphone and move both her hands. Fig. 25 shows two snapshots of the obtained range-velocity profiles. With just one speaker and one microphone, we can clearly track the two hands. Another interesting observation is that even our objective is to track the two hands, the human body is also clearly tracked as shown in Fig. 25.

Tracking two fingers using a smart speaker. We also implement FM-Track on a smart speaker prototype, i.e., UMA-8-SP USB mic array [26], which is equipped with seven microphones. With more microphones, smart speakers present a higher resolution in spatial domain, and accordingly better overall sensing performance. Fig. 26 shows the trajectories of the index finger and middle finger. We can see that the two fingers can be clearly resolved and they can be accurately tracked at an accuracy of 0.93 *cm*. We believe this finger-level tracking granularity and multi-finger tracking capability could enable new applications on commodity smart speakers.

8 RELATED WORK

Device-based acoustic tracking. Owing to its low propagation speed, acoustic signals have been widely employed for motion tracking, where users are required to hold a device. AAMouse [52] tracks a mobile device based on Doppler shift at multiple frequencies. CAT [23] and Rabbit [25] employ distributed FMCW devices to estimate the distance between the transmitter and receiver. SoundTrak [54] extracts the phase information from the sinusoidal waves to track a customized finger ring with respect to a smartwatch in 3D space. MilliSonic [41] leverages the phase of the FMCW signal to enable concurrent tracking of multiple mobile devices. All these systems infer human motions by tracking the movement of devices.

Device-free (contactless) acoustic tracking. Many systems have been developed to perform contactless tracking without requiring users to carry a device. Efforts have pushed the granularity of tracking to a millimeter level. FingerIO [28] exploits the auto-correlation properties of OFDM symbols to achieve *mm*-level tracking, while LLAP [45], Strata [53] and the system proposed by Sun *et al.* [34] track fine-grained movements by capturing the phase change of signals. VSkin [35] tracks finger movements on the back of a smartphone by extracting the amplitude and phase information from both structure-borne and air-borne acoustic signals. A recent work RTrack [24] enables room-scale hand motion tracking by combining a microphone array with a series of signal processing techniques and an RNN. These studies have focused mainly on single-target, whereas FM-Track aims to enable fine-grained multi-target tracking.

Contactless multi-target tracking. There have been a few attempts to contactlessly track multiple targets using wireless signals [1, 9, 14, 20, 47]. For WiFi signals, WiDeo [14] and mD-Track [47] propose to leverage information from multiple dimensions including angle, ToF and Doppler shift, to isolate the superposed signals reflected from multiple targets. For radar/sonar signals, previous work [1, 9, 20] leverage their inherent large bandwidths and antenna arrays to support multi-target tracking. The above-mentioned multi-target tracking systems focus on body-level human motion tracking, whereas FM-Track targets at finer-grained hand and finger tracking.

9 CONCLUSION

This paper presents FM-Track, a novel acoustic-based system that pushes the limits of contactless multi-target tracking. We demonstrate, for the first time, an approach to accurately track multiple finger-sized targets using acoustic signals and showcase its feasibility in real-life applications. To achieve this, we propose a chirp-based signal model to fuse the angle, range, and velocity information of signals reflected from multiple targets. We also propose solutions to address the unique issues associated with chirp-based tracking, such as velocity ambiguity and unavailability of instantaneous velocity. We believe the proposed methods can benefit other work on chirp-based tracking, such as LoRa and Radar.

ACKNOWLEDGMENT

This work was partially supported by National Institutes of Health (NIH) under Award Number 5R01MH122371-02.

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] Amazon. 2020. *Amazon Echo Dot 2nd Generation*. https://www.amazon.com/All-New-Amazon-Echo-Dot-Add-Alexa-To-Any-Room/dp/B01DFKC2SO

[3] Apple. 2020. *iPhone 5c*. https://support.apple.com/kb/sp684?locale=en_US

[4] Francesco Belfiori, Wim van Rossum, and Peter Hoogeboom. 2012. 2D-MUSIC technique applied to a coherent FMCW MIMO radar. (2012).

[5] 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 Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 278–291.

[6] 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.

[7] EARISE. 2020. *AL-202 High-fidelity USB Acoustics System*. https://www.sears.com/gemini-doctor-al-202-high-fidelity-usb-acoustics/p-SPM11706412330

[8] SparkFun Electronics. 2020. *ADMP401 MEMS microphones*. https://www.sparkfun.com/products/9868

[9] Doug Grimmett and Cherry Wakayama. 2013. Multistatic tracking for continous active sonar using Doppler-bearing measurements. In *Proceedings of the 16th International Conference on Information Fusion*. IEEE, 258–265.

[10] Yalin Guercan and Alexander Yarovoy. 2017. Super-resolution algorithm for joint range-azimuth-Doppler estimation in automotive radars. In *2017 European Radar Conference (EURAD)*. IEEE, 73–76.

[11] 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*. ACM, 1911–1914.

[12] Yaxiong Huang, Simon C Busbridge, and Deshinder S Gill. 2001. Distortion and directivity in a digital transducer array loudspeaker. *Journal of the Audio Engineering Society* 49, 5 (2001), 337–352.

[13] Qualisys Inc. 2020. *Qualisys motion capture systems*. https://www.qualisys.com/hardware/miqus/

[14] Kiran Joshi, Dinesh Bharadia, Manikanta Kotaru, and Sachin Katti. 2015. WiDeo: Fine-grained Device-free Motion Tracing using RF Backscatter. In *12th USENIX Symposium on Networked Systems Design and Implementation (NSDI 15)*. 189–204.

[15] Chitra R Karanam, Belal Korany, and Yasamin Mostofi. 2019. Tracking from one side: multi-person passive tracking with WiFi magnitude measurements. In *Proceedings of the 18th International Conference on Information Processing in Sensor Networks*. 181–192.

[16] Vladimir Katkovnik, Moon-Sik Lee, and Yong-Hoon Kim. 2002. High-resolution signal processing for a switch antenna array FMCW radar with a single channel receiver. In *Sensor Array and Multichannel Signal Processing Workshop Proceedings, 2002*. IEEE, 543–547.

[17] Hyosu Kim, Anish Byanjankar, Yunxin Liu, Yuanchao Shu, and Insik Shin. 2018. UbiTap: Leveraging acoustic dispersion for ubiquitous touch interface on solid surfaces. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*. 211–223.

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

[19] Harold W Kuhn. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly* 2, 1-2 (1955), 83–97.

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

[21] Xiang Li, Daqing Zhang, Qin Lv, Jie Xiong, Shengjie Li, Yue Zhang, and Hong Mei. 2017. IndoTrack: Device-free indoor human tracking with commodity Wi-Fi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–22.

[22] Jaime Lien, Nicholas Gillian, M Emre Karagozler, Patrick Amihood, Carsten Schwesig, Erik Olson, Hakim Raja, and Ivan Poupyrev. 2016. Soli: Ubiquitous gesture sensing with millimeter wave radar. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 1–19.

[23] 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*. ACM, 69–81.

[24] Wenguang Mao, Mei Wang, Wei Sun, Lili Qiu, Swadhin Pradhan, and Yi-Chao Chen. 2019. RNN-Based Room Scale Hand Motion Tracking. In *The 25th Annual International Conference on Mobile Computing and Networking*. ACM, 38.

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

[26] MiniDSP. 2020. *UMA-8-SP USB mic array*. https://www.minidsp.com/products/usb-audio-interface/uma-8-sp-detail

[27] Rajalakshmi Nandakumar, Krishna Kant Chintalapudi, Venkat Padmanabhan, and Ramarathnam Venkatesan. 2013. Dhwani: secure peer-to-peer acoustic NFC. *ACM SIGCOMM Computer Communication Review* 43, 4 (2013), 63–74.

[28] 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*. ACM, 1515–1525.

[29] Boaz Rafaely. 2004. Analysis and design of spherical microphone arrays. *IEEE Transactions on speech and audio processing* 13, 1 (2004), 135–143.

[30] Hermann Rohling and M-M Meinecke. 2001. Waveform design principles for automotive radar systems. In *2001 CIE International Conference on Radar Proceedings (Cat No. 01TH8559)*. IEEE, 1–4.

[31] Wenjie Ruan, Quan Z Sheng, Lei Yang, Tao Gu, Peipei Xu, and Longfei Shangguan. 2016. AudioGest: enabling fine-grained hand gesture detection by decoding echo signal. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 474–485.

[32] Ralph Schmidt. 1986. Multiple emitter location and signal parameter estimation. *IEEE transactions on antennas and propagation* 34, 3 (1986), 276–280.

[33] Sheng Shen, Daguan Chen, Yu-Lin Wei, Zhijian Yang, and Romit Roy Choudhury. 2020. Voice Localization Using Nearby Wall Reflections. In *The 25th Annual International Conference on Mobile Computing and Networking*. ACM. Preprint for MobiCom 2020.

[34] Ke Sun, Wei Wang, Alex X Liu, and Haipeng Dai. 2018. Depth aware finger tapping on virtual displays. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 283–295.

[35] Ke Sun, Ting Zhao, Wei Wang, and Lei Xie. 2018. Vskin: Sensing touch gestures on surfaces of mobile devices using acoustic signals. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. ACM, 591–605.

[36] Bela Team. 2020. *Bela Platform*. https://bela.io

[37] Cozmo Team. 2020. *Cozmo Smart Robot*. https://anki.com

[38] Hercules Team. 2020. *4WD Hercules Mobile Robotic Platform*. http://wiki.seeedstudio.com/Skeleton_Bot-4WD_hercules_mobile_robotic_platform

[39] Elisabet Tiana-Roig, Finn Jacobsen, and Efrén Fernández Grande. 2010. Beamforming with a circular microphone array for localization of environmental noise sources. *The Journal of the Acoustical Society of America* 128, 6 (2010), 3535–3542.

[40] Yu-Chih Tung, Duc Bui, and Kang G Shin. 2018. Cross-platform support for rapid development of mobile acoustic sensing applications. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. 455–467.

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

[42] Anran Wang, Jacob E Sunshine, and Shyamnath Gollakota. 2019. Contactless infant monitoring using white noise. In *The 25th Annual International Conference on Mobile Computing and Networking*. 1–16.

[43] Ju Wang, Jie Xiong, Hongbo Jiang, Xiaojiang Chen, and Dingyi Fang. 2017. D-Watch: Embracing "Bad" multipaths for device-free localization with COTS RFID devices. *IEEE/ACM Transactions on Networking* 25, 6 (2017), 3559–3572.

[44] Tianben Wang, Daqing Zhang, Yuanqing Zheng, Tao Gu, Xingshe Zhou, and Bernadette Dorizzi. 2018. C-FMCW based contactless respiration detection using acoustic signal. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 170.

[45] 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*. ACM, 82–94.

[46] Yadong Xie, Fan Li, Yue Wu, Song Yang, and Yu Wang. 2019. D3-Guard: Acoustic-based Drowsy Driving Detection Using Smartphones. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 1225–1233.

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

[48] 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.

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

[50] Xiangyu Xu, Jiadi Yu, Yingying Chen, Yanmin Zhu, Linghe Kong, and Minglu Li. 2019. BreathListener: Fine-grained Breathing Monitoring in Driving Environments Utilizing Acoustic Signals. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 54–66.

[51] Lei Yang, Yekui Chen, Xiang-Yang Li, Chaowei Xiao, Mo Li, and Yunhao Liu. 2014. Tagoram: Real-time tracking of mobile RFID tags to high precision using COTS devices. In *Proceedings of the 20th annual international conference on Mobile computing and networking*. 237–248.

[52] 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.

[53] 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*. ACM, 15–28.

[54] 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.

[55] Maotian Zhang, Qian Dai, Panlong Yang, Jie Xiong, Chang Tian, and Chaocan Xiang. 2018. idial: Enabling a virtual dial plate on the hand back for around-device interaction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 55.