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

Paper #414

ABSTRACT

Contactless acoustic motion tracking enables new opportunities to interact with smart devices, such as smartphones, smart TVs, 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 introduce a signal model to characterize the location and motion status of multiple targets by fusing the information from multiple dimensions (i.e., angle, range and velocity of targets). With the proposed model, we can distinguish signals reflected from multiple targets and accurately track each individual signal. We implement FM-Track on an embedded system (i.e., Bela) connected with a speaker and an array of four microphones. Experiments show that FM-Track can successfully differentiate two targets with a spacing as small as 1 cm and achieve a median accuracy of 0.86 cm. We also demonstrate the feasibility and reliability of real-world applications by tracking hands and fingers.

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 findings have 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 [21, 24], contactless motion tracking [19, 25, 33, 41] and gesture recognition [7, 23, 43], as well as monitoring of important physiological parameters in humans (e.g., fatigue detection for drivers [34, 38] and respiratory activities [5, 32]). Specifically for contactless motion tracking, compared to other wireless signals such as WiFi [15, 37] and RFID [31, 39] which have been employed to enable similar applications, 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 without requiring users to instrument a device [19, 25, 33, 41] and extended the sensing range to 4 m [17], 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 passive sensing that it relies on signals reflected from the targets. Signals reflected from multiple targets are mixed at the receiver, and it is difficult to separate them to obtain the context information of each individual target. The problem becomes even more challenging when targets are close to each other. Second, while prior studies have achieved a high accuracy for tracking targets with relatively large sizes (e.g., body [17, 20] and hand [6]), the performance of tracking small targets such as fingers significantly degrades when the target is far away from the sensing device. This is mainly because the signal strength dramatically diminishes when the reflection area is small. Last but not least, while existing work has achieved mm-level accuracy in tracking the displacement and trajectory of one target, the estimation of the distance between the target and the sensing device still offers room for improvement.

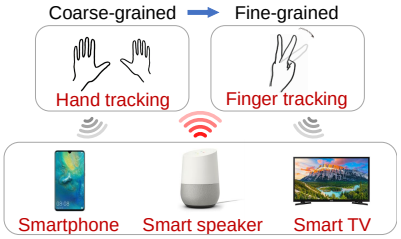


Figure 1: Application scenarios for contactless multi-target tracking using acoustic signals.

In this paper, we propose to employ chirp-based acoustic signals to push the boundaries of acoustic tracking in three aspects: 1) enable multi-target sensing, 2) increase the tracking granularity for finger-sized targets, and 3) accurately track not just the target's displacement but also the absolute distance from the sensing device. We demonstrate the application of fine-grained contactless interactions between the user and smart devices by tracking their hands and fingers in Figure 1.

In literature, there have been several attempts to address multi-target tracking using other types of wireless signals, such as WiFi [11, 35] and radar signals [3]. However, these techniques do not successfully translate to acoustic-based multi-target tracking that we consider in this paper mainly due to the following reasons.

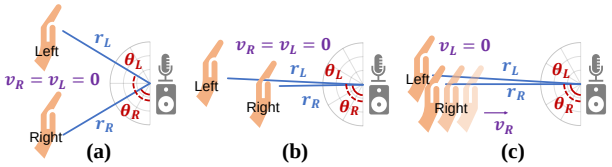


Figure 2: Separate targets with information of multiple dimensions (range r, angle θ and velocity v):(a) targets with similar r can be separated by different θ; (b) targets with similar θ can be separated by different r; (c) targets with similar r and θ can be separated by different v.

- Unlike OFDM-based WiFi signals, the frequency of chirp-based acoustic signals changes linearly over time. This difference requires to completely redesign the signal separation algorithm to achieve multi-target tracking with acoustic signals. On the other hand, compared to radar devices whose bandwidth is in the scale of GHz, the bandwidth of inaudible frequencies in acoustic devices is limited to several KHz.
- Previous studies that leverage WiFi signals to track multiple targets rely on one important assumption that the strengths of the signals reflected from multiple targets are different [35]. While this assumption is valid for large-size targets, it is not true for closely-located small targets, such as two fingers, because the strengths of reflected signals could be very similar.
- Compared to WiFi or radar-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 a finer-grained velocity estimation in the scale of centimeter per second (e.g., finger movements). To achieve accurate velocity estimation, data samples from a large time window are needed [7]. 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 speeds. Thus, accurate estimation of instantaneous velocity using acoustic signals remains unaddressed.
- Another issue that is more prominent in acoustic-based tracking—when compared to WiFi or radar-based tracking—is the range-Doppler effect that yields a relatively large range deviation [17]. For example, a chirp-based acoustic signal with a start frequency of $f_0 = 16\ KHz$, bandwidth of $B = 4\ KHz$, and sweep time of $T = 0.04\ s$, would induce a $47.1\ Hz$ Doppler shift for a moving target at a velocity of $v = 0.5\ m/s$. This Doppler shift will cause a deviation of $8\ cm$ ($\Delta d = \frac{v f_0 T}{B}$) when estimating the distance between the sensing device and the target. This is unacceptably large for millimeter-level acoustic 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 is general and can be applied to other chirp-based signals (e.g., LoRa signals) to enable fine-grained multi-target sensing.

We first propose a chirp-based signal model to fuse the angle (spatial domain), range (time domain) and velocity (frequency domain) information of multiple targets from reflected signals, which can characterize the location and motion information for targets as shown in Figure 2. Based on the signal model, we propose our joint parameter (range, angle, velocity and attenuation) estimation algorithm to resolve each individual signal reflected from multiple targets. Different from the pseudo-joint estimation proposed in mD-Track [35], our joint estimation does not require the signal strengths to be dramatically different from each other.

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 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 those 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 [22]. To address the above issues, we employ the fact that the movement of targets is continuous in spatial, time as well as frequency domain.

We implement FM-Track on the Bela platform [27], connected with one speaker and an array of four MEMS microphones [2]. 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 relative range estimates respectively. FM-Track can successfully separate two targets with a spacing as small as $1\ cm$, outperforming the state-of-the-arts. FM-Track can simultaneously track up to *four* targets at high accuracies. We also demonstrate the high accuracies of FM-Track with several real-life applications, including two-hand tracking and multi-finger tracking.

2 PRELIMINARIES

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

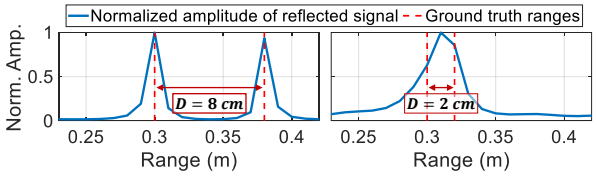
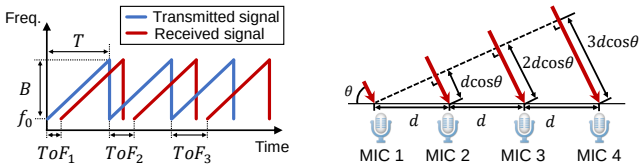


Figure 3: Two targets become unresolvable when the distance between them D is smaller than the range resolution ($r = 4.25\ cm$ in this example).



(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.

2.1 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., distance between the target and the sensing device—can be obtained by measuring signal propagation time of the reflected acoustic signal and multiplying by signal propagation speed. Fundamentally, the resolution of the range estimation is determined by the frequency bandwidth of the signal [30, 35]. A larger bandwidth yields more accurate estimation of the signal propagation delay, which leads to 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 [36].

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. The larger the window size, the finer the resolution and thus, more accurate velocity estimation.

We can exploit the above-mentioned multi-dimensional information to distinguish multiple targets, as illustrated in Figure 2. When two targets have similar range values, the AoA information can be leveraged to distinguish them (Figure 2a). Similarly, two targets that share similar angles can be distinguished by their range information (Figure 2b). Even when two targets share both similar range and AoA values, they could still be distinguished by their velocities (Figure 2c).

2.2 Tracking Accuracy vs. Resolvability

The majority of prior efforts in acoustic-based tracking systems has focused on improving the tracking accuracy for a single target [19, 25, 33, 41], whereas relatively little attention has been paid to improving the resolvability to distinguish multiple targets. For multi-target tracking, both the tracking accuracy and resolvability should be considered. Consider an example where two targets are separated with different range differences of $D = 8\ cm$ and $2\ cm$ in Figure 3. The dashed red line represents the ground truth range values. We consider a signal bandwidth of $B = 4\ KHz$, which yields a range resolution of $r = 4.25\ cm$. As shown in Figure 3, when the range difference ($D = 8\ cm$) between two targets is greater than the r, we can clearly see two peaks and thus, the two targets are resolvable. When the range difference decreases to $2\ cm$ which is smaller than r, the two signal peaks merge into one and the two targets become unresolvable. This example demonstrates that for multi-target tracking, both accuracy and resolvability are important metrics to characterize the performance.

2.3 Relative Tracking vs. Absolute Tracking

Recent efforts have pushed the granularity of contactless acoustic tracking to millimeter level [19, 25, 33, 41]. The underlying fundamentals of these tracking methods are fine-grained phase change. For instance, a phase change of 40° of acoustic signals 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 chirp signals. In theory, with a $4\ KHz$ bandwidth, the absolute tracking accuracy is around $\frac{34000\ cm/s}{2 \times 4000\ Hz} = 4.25\ cm$. Wang *et al.* have reported a relative tracking accuracy of $3.5\ mm$ and an absolute tracking accuracy of $3.57\ cm$ [33], matching our analysis here.

3 A MATHEMATICAL SIGNAL MODEL

This section describes a detailed mathematical derivation of our chirp-based signal modeling that is fundamental to our algorithms to identify and track multiple targets.

As shown in Figure 4, the core idea behind chirp-based acoustic tracking is to compute the Time-of-Flight (ToF) of the chirp signal 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 by the sound speed in the air. The velocity information can be estimated by measuring the ToF differences across multiple chirps (Figure 4a). In addition, we can estimate the angle information (i.e., angle of the target location with respect to the sensing device's orientation) by measuring the ToF differences across multiple microphones (Figure 4b).

3.1 Signal Model for a Single Target

We first explain how the transmitted and reflected signals are processed to derive the ToF of one target. As shown in Figure 4a, the speaker transmits a sequence of chirp signals where the frequency in each chirp varies linearly with time. 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 the 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.

Figure 5 illustratively 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 signal $S^{T'}(t) = \sin\left(2\pi(f_0 t + \frac{B}{2T}t^2)\right)$ 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., $\cos A \cdot \cos B = \frac{1}{2}(\cos(A-B) + \cos(A+B))$ and a low-pass filter, the In-Phase part of the mixed 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. Similarly, the Q 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 construct 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, angle, and velocity 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

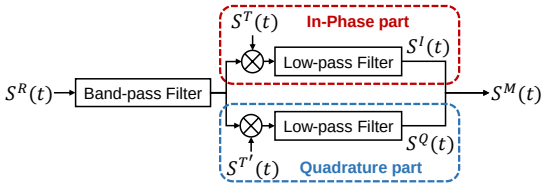


Figure 5: The mixed signal $S^M(t)$ can be constructed from the In-phase part $S^I(t)$ and Quadrature part $S^Q(t)$.

received by this microphone can be computed by 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. 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 Figure 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), 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 derived as

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

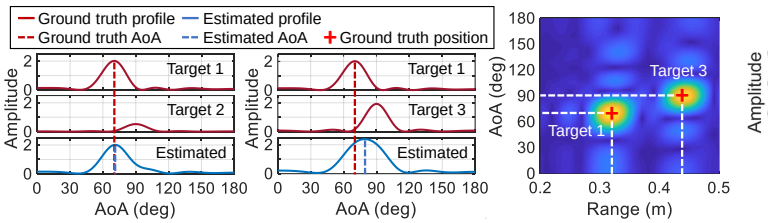
where t_i is the i^{th} sampling timestamp within the c^{th} chirp and $\varphi(t_i, c, k)$ is the phase change induced by the k^{th} microphone in the c^{th} chirp at the i^{th} sampling timestamp. Equation (7) contains information relevant to range, angle, and velocity of the target, which can be rearranged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

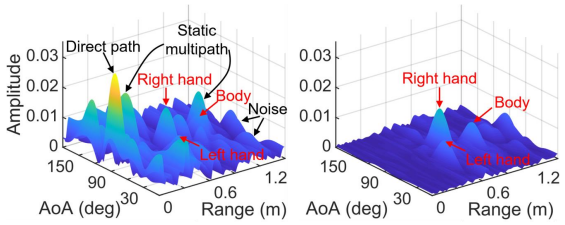
$$\begin{aligned} R &= e^{j\varphi_r} = e^{j2\pi(f_0 + \frac{B}{T}t_i)\frac{2r}{v_s}} \\ \Theta &= e^{j\varphi_\theta} = e^{j2\pi(f_0 + \frac{B}{T}t_i)\frac{(k-1)d\cos\theta}{v_s}} \\ V &= e^{j\varphi_v} = e^{j2\pi(f_0 + \frac{B}{T}t_i)\frac{2(c-1)Tv}{v_s}}. \end{aligned} \tag{9}$$

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 fully characterize the location and motion



(a) The AoA profile. (b) The AoA profile. (c) The AoA-range profile.

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.



(a) Before Back-Subtract. (b) After Back-Subtract.

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).

status of one target. R, Θ and V are the steering vectors for each dimension [12], i.e., range, angle, and velocity.

By now, we have developed the sensing model for one single target. It is the first chirp-based acoustic model which can simultaneously extract the target's range, angle, and velocity information. Note that the proposed chirp-based signal model is significantly different from the state-of-the-art models proposed for WiFi [35]. 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 on acoustic tracking. Specifically, in WiFi-based models, the range information can be estimated from a single time domain sample. 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. Furthermore, in WiFi signals, the velocity information of a moving target is obtained from the frequency-domain Doppler shift. In contrast, in our chirp-based model, the velocity information is calculated from an extra time delay.

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 Section 4—is to separate the mixed signals, and estimate the corresponding parameters $\boldsymbol{p}_l = [\theta_l, r_l, v_l, \alpha_l]$ for each individual signal.

4 RESOLVING MULTIPLE TARGETS

In this section, we describe how FM-Track enables multi-target tracking. We believe that the proposed algorithms provide an unprecedented opportunity to improve the signal resolvability (i.e., the ability of distinguishing reflected signals from two closely located targets), when compared to existing chirp-based acoustic tracking solutions [16–18, 30], using a novel approach to enable joint estimation of the range, angle and velocity. Prior to presenting our algorithms in depth, we briefly discuss why the state-of-the-art algorithm (mD-Track) designed for WiFi-based multi-target tracking [35] cannot be effectively translated to work with acoustic signals.

4.1 Joint vs. Pseudo-joint Estimation

Pseudo-joint estimation is adopted in mD-Track, which estimates the parameters on each dimension sequentially to avoid high computational cost. The basic idea of mD-Track is to iterate the process of estimating the strongest signal while considering the rest signals as noise. A fundamental assumption for this approach is that, in each iteration, there exists one signal with prominent strength compared to other signals, which does not hold true when two targets are placed close to each other (e.g., hands or fingers, as considered in this study).

To illustrate this issue, we conduct the following experiment. We first place two finger-sized cardboards Target 1 and 2 at position $[0.32\ m, 70°]$ 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 works well as shown in Figure 6a. Now we replace Target 2 with a larger size cardboard Target 3. Due to a larger size, the strength of the reflected signal is now comparable to that from Target 1. Now the assumption does not hold any more and we can see the estimated angle is deviated from the ground truth in Figure 6b.

To address the issue of strong interference among signals with comparable strengths, we propose to jointly consider parameters from all the dimensions. This true joint estimation can effectively separate multiple signals since the increased dimensionality can augment the uniqueness with information from multiple dimensions, thus reducing the effect of interference from one single dimension. As shown in Figure 6c, with the second dimension, i.e., range, two signals can be clearly separated even though the angles are close to each other. We want to point out that the computational cost is not

an issue for our joint estimation. The first reason is that due to the low propagation speed, the start point estimate can be quite accurate. The second reason is that we adopt a small window (several chirps) to estimate signal parameters. Within such a small time window, the hand and finger movements are limited and thus we only need to search a small space. From our experiments, the median run time of our algorithm is only 39.9 *ms* for a single target and 56.6 *ms* for two targets.

4.2 Resolving 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 Noise and Multipath Subtraction. Before estimating the multi-dimensional parameters, we need to first remove the background noise, including the direct path from the speaker to microphones and reflections from surroundings. As shown in Figure 7a, the reflections from a human target (body and hands) are overwhelmed in the background noise, which would significantly decrease the tracking accuracy. Therefore, we measure the background signal when there is no target and remove it later to improve the tacking accuracy. After background subtraction, the reflections from hands and body clearly stand out as shown in Figure 7b.

4.2.2 Single Path Estimation. Assuming there is only one target to track, the estimation process can be decomposed into three steps: 1) constructing the joint estimator, 2) searching the optimal parameters for the 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 (9). Note that, 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 (i.e., $\hat{\varphi}$) based on the estimated parameters 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 now we remove these accurately estimated phase changes $(\hat{\varphi_r}, \hat{\varphi_\theta}, \hat{\varphi_v})$ from the measured phase, a signal (i.e., $\frac{1}{2}\alpha e^{j0}$) is resulted. This implies that the phase-removed signals will be in-phase and combine 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 corresponding 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 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 will be in-phase. Hence, the optimization problem can be formulated as

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

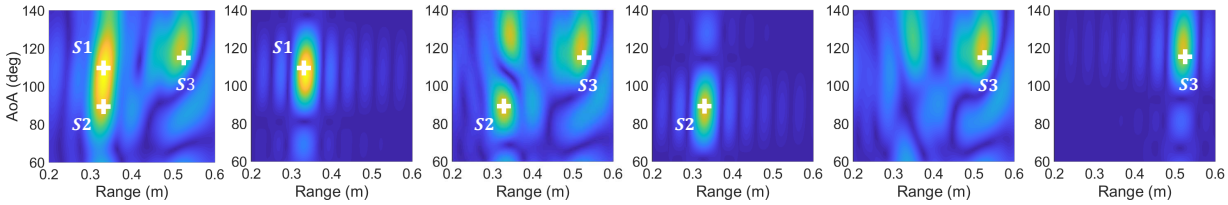
The search range of the parameters can be defined based on the application. In our application of hand motion tracking, we can define the search range of these parameters according to the physical constraints of hand movement. Then, we perform a search over the three dimensions to find the optimal parameters.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

Computing the attenuation factor ($\hat{\alpha}$). After obtaining the estimates for all the 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 algorithm are the four path parameters, $\boldsymbol{p} = [\hat{\theta}, \hat{v}, \hat{r}, \hat{\alpha}]$, associated with a single target.

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 within the mixed signal. Then, we reconstruct the strongest signal based on Equation (7) and subtract this estimated signal from the mixed signal. 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 and can thus be considered as noise. Figure 8 illustratively demonstrates the



(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.

process for estimating parameters for signals reflected from three targets, namely $S1$, $S2$, and $S3$.

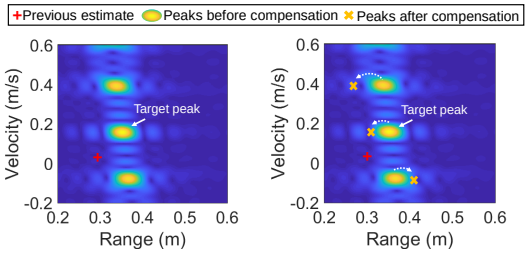
4.2.4 More iterations. Although the parameters for all paths can be estimated from just one iteration, there still exist errors in the estimation. The main reason is that when we estimate the strongest signal, we consider all the weaker signals as noise. After one iteration, we obtain the estimates of signals and the remaining part is the estimated noise. This estimated noise is much closer to the true noise than the "weaker signals + noise". Thus we can iterate the process by applying this more accurate estimated noise for signal estimation in the second round. After this iterative process, we can obtain more accurate parameter estimates. The above iterative process is performed until the difference of each path parameter between two consecutive iterations is smaller than a pre-defined threshold. From our experiments, the average number of iterations is very small, *i.e.*, 2.3.

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 $O(N_{iter} \cdot L \cdot \eta)$.

Reducing search windows. 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) as well as 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 computational cost and maintaining high accuracy. We empirically set $D = 40$ based on the trade-off analysis in Section 7.4.



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

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

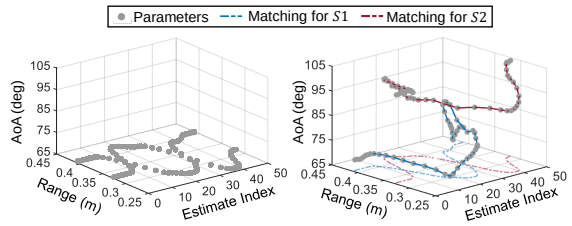
Accurate starting position estimate. An accurate starting position estimate can significantly reduce the computational cost. For acoustic signals, the range estimate of the starting position is particularly accurate. This is another important factor why the optimal parameters can be found very quickly.

5 PERFORMING ACCURATE TRACKING

This section first introduces how to deal with the velocity ambiguity issue, followed by matching signals from two consecutive estimates for each target. At last, we propose a novel method to compute the instantaneous velocity.

5.1 Dealing with Velocity Ambiguity

There is an intrinsic ambiguity issue in estimating the velocity of a moving target using chirp-based signals [22]. Specifically, there would exist multiple peaks, including one target peak and several ambiguous peaks, as shown in Figure 9a. Due to the noise and multipath, the peak with the largest amplitude may not correspond to the target peak, making it infeasible to identify the target peak by amplitude. We propose to apply the continuity property among adjacent estimates to identify



(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, (b) much easier using both range and angle information.

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 very 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 which has the smallest distance with the previous estimate in the velocity-range 2D space. However, this approach still does not work due to the range-Doppler deviation induced by target movement. We thus compensate the raw peak estimates before we apply the continuity property to remove the velocity ambiguity. We illustrate the concept in Figure 9. The estimate from the previous round is marked as a red plus. In Figure 9a, without compensation, two peaks have similar distances to the previous round estimate. After compensating the range-Doppler effect (detail in Section 5.3), the compensated target peak (marked as a yellow cross) is much closer to the previous estimate compared to the other two compensated ambiguous peaks as shown in Figure 9b.

5.2 Matching Signals with Targets

To track multiple 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, human movements are smooth and continuous), and thus, the parameters between two consecutive estimates from the same target should be close. However, when two targets have overlapping trajectories, the parameters of two targets become similar, causing confusion in tracking as shown in Figure 10a.

We propose to employ parameters from more dimensions to address this issue, because it is unlikely that multiple targets share similar parameters in all dimensions at a given timestamp. For example, for simplicity to visualize, Figure 10b shows that the ambiguity in Figure 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 used a weighted L1-norm

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

where $\boldsymbol{p}^t = [\theta^t, r^t, v^t, \alpha^t]$ represents the parameter estimate at time t, and d 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 in order to exhaustively search the optimal pairs. We thus employ the Hungarian algorithm [11, 14] to reduce the computational complexity from $O(n!)$ to polynomial time.

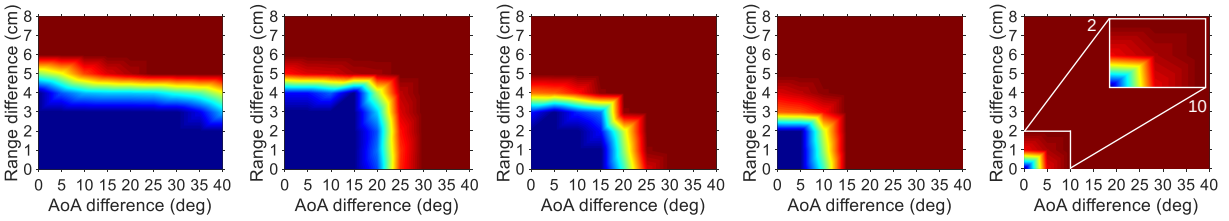
5.3 Refinement of Parameter Estimates

The room for improvement arises from the fact that the velocity estimates are computed as the average velocity across multiple chirps. It is well known that the chirp-based methods have difficulties in obtaining accurate instantaneous velocity [7]. In this section, we propose a novel method to obtain instantaneous velocity by exploiting the fact that the information estimated from multiple dimensions is not equally accurate. Specifically, we rely on more accurate range estimates owing to the low propagation speed to compute the instantaneous velocity. Without the instantaneous velocity, the range deviation is compensated with the average velocity which is not accurate. We can now refine the range deviation with more accurate instantaneous velocity.

We now describe how to obtain the instantaneous velocity. At the beginning when the target is static, the estimated range r_1 is undeviated. The target starts moving and at the next timestamp, the target has moved to another location and the estimated range r_2 is deviated due to the range-Doppler effect. Compared with the range estimate in previous timestamp r_1, 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, i.e., $r_2 = r_1 + vT + \frac{vf_0T}{B}$, where v is the velocity of the moving target. Since both r_1 and r_2 can be accurately obtained, we can thus derive the velocity v from the above equation. This velocity v is obtained from two range estimates from adjacent chirps, which can thus be considered as the instantaneous velocity due to the small duration of one chirp (40 ms). With the instantaneous velocity, we can calculate the movement-caused range deviation and remove the deviation from the range estimate.

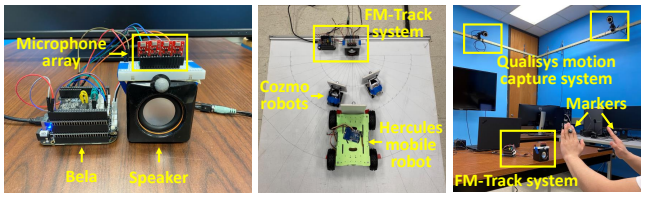
6 IMPLEMENTATION

We implement FM-Track on the Bela platform [27] as shown in Figure 12, which has the flexibility to change the number 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 11: Resolvability comparison. 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.



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

Figure 12: Experiment setup.

microphones and vary the microphone/speaker locations for comprehensive evaluations. The received acoustic samples are fed into a laptop equipped with an Intel i7 processor and a 32 GB memory via a USB cable. The algorithms are implemented in MATLAB.

Acoustic components: The Bela platform is connected with one general-purpose speaker [1] to transmit acoustic signals and an array of up-to four MEMS microphones [2] to receive acoustic signals. Many commercial smart devices are also equipped with multiple microphones and speakers, such as Amazon Echo (five speakers and seven microphones) [4] and Huawei Sound X (two speakers and six microphones) [9].

Acoustic signals: The default chirp signals adopted in our implementation sweep from 16 KHz to 20 KHz with a chirp duration of 40 ms at the sampling rate of 44.1 KHz.

Robots: To enable controlled experiments in terms of velocity and moving trajectory, we mount cardboards on two different kinds of robots. The first one is called Cozmo [28], which can be controlled at a speed granularity of 1 mm/s with a maximum speed of 20 cm/s. The second one is 4WD Hercules Robot [29] whose speed granularity is 2 cm/s with a maximum speed of 2 m/s. Both robots can be controlled to follow a pre-defined trajectory.

Ground truth: As shown in Figure 12c, we employed an optoelectronic motion capture system (i.e., Qualisys [10]), which supports sub-mm-level multi-target motion tracking with a frame rate of 250 Hz, to capture the ground truths of the target movements. The targets, including robots, human hands and fingers, are equipped with passive (reflective) markers to be tracked by an array of six cameras mounted on the ceiling.

7 EVALUATION

For experiments involving moving targets, we mount the targets on mobile robots for precise controlling of target movements. The robots begin from static and accelerate to the maximum speed. The maximum speed is adjusted to meet different requirements of experiments. We further conduct field studies by tracking hands and fingers to demonstrate the feasibility and reliability of FM-Track in real-life scenarios. Unless otherwise specified, we adopt four microphones and four chirps with a bandwidth of 4 KHz to perform three-dimensional parameter estimation. Each experiment is repeated 20 times.

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 [35] and MUSIC [13, 16]. 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) [17]. Because we do not have sufficient data to train an RNN, we emulate their performance by performing estimation using 2D MUSIC and compensating the range-Doppler effect using the actual instantaneous velocity captured by the motion capture system for a fair comparison.

Capability of resolving multiple targets. The key to enable multi-target tracking is the capability of separating signals 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 difference, AoA difference, 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

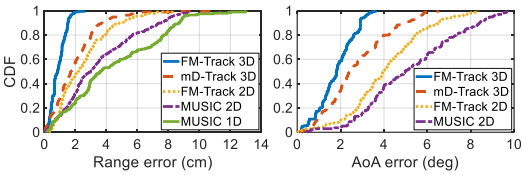


Figure 13: Method comparison.

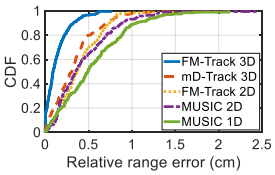


Figure 14: Relative range error.

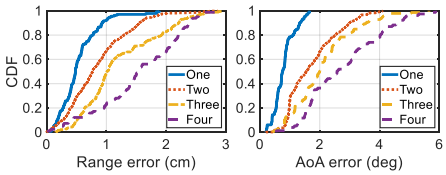


Figure 15: Impact of target numbers.

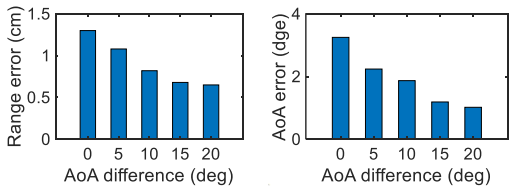


Figure 16: Impact of angle differences.

other target with the help of the robots. Figure 11 depicts the resolvability comparison among different approaches, where cooler (blue) colors indicate "non-resolvable" and warmer (red) colors denote "resolvable". Thus, the smaller the blue region, the better resolvability performance. We observe that, with information from three dimensions, FM-Track 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 close-by fingers.

Tracking accuracy. In this section, we compare the tracking accuracy of FM-Track with two state-of-the-art systems. We quantify the tracking accuracy using three metrics: absolute range error, absolute AoA error, and relative range (displacement) error. The default number of targets is two. We mount two pieces of finger-sized cardboard on two Cozmo robots, one of which is kept static at position $[0.30\ m, 70°]$. We make the second robot move towards the sensing device with 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 true instantaneous velocity captured by the motion capture system.

Absolute range and AoA accuracy. Figure 13 shows the tracking errors for range and AoA, respectively. The median range error for FM-Track 3D is 0.86 cm, outperforming mD-Track 3D by approximately 100%. The median AoA errors achieved with FM-Track 3D and mD-Track 3D are 1.82° and 2.45°, respectively. The proposed system outperforms mD-Track 3D due to its capability of resolving signals with comparable strengths.

Relative range (displacement) accuracy. We compare the relative range errors of different systems in Figure 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 corresponding absolute range error depicted in Figure 13, which supports our discussions of relative vs. absolute ranges in Section 2.3. Our proposed system is able to achieve millimeter tracking accuracy for both the relative and absolute ranges.

7.2 Factors Affecting the Performance

In this section, we evaluate the 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 higher overall resolvability. Owning to the high dimensionality, even with just two microphones, the achieved range resolvability is 3 cm, which means two close-by targets could still be separated and accurately tracked. With four microphones, the resolvability is 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. Furthermore, we study the impact by varying the bandwidth size from 2 KHz to 6 KHz. With larger bandwidth, we can improve the resolution of range and thus, better resolvability. Interestingly, the improvement is marginal when we increase the bandwidth beyond 4 KHz. These results collectively show that, to improve resolvability, it is more efficient to increase the dimensionality of the signals rather than improving the resolution of the signal within a single dimension.

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

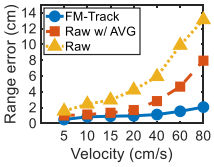
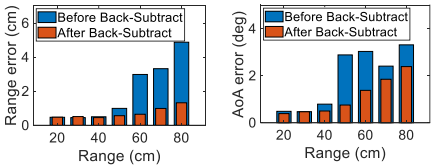
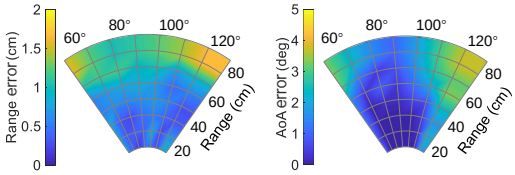


Figure 17: Effectiveness of parameter refinement.



(a) Range error. (b) AoA error.

Figure 18: Effectiveness of Back-Subtract.

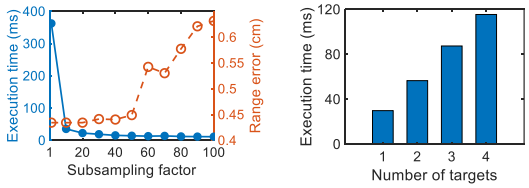


(a) Range error. (b) AoA error.

Figure 19: Starting position error for different positions.

Table 1: Processing time for FM-Track (unit: *ms*).

Pre-processing	Parameter estimation	Signal-target matching	Parameter refinement	Total
13.4	115.5	1.8	0.2	130.9



(a) Subsampling. (b) Number of targets.

Figure 20: Execution time for parameter estimation.

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°).

Effectiveness of parameter refinement. We now show the benefit of the parameter refinement process that we presented in Section 5.3. We compare the results obtained with vs. without the parameter refinement. More specifically, we have two schemes for refinement: 1) refinement with the instantaneous velocity proposed in this work and 2) refinement with the traditional average velocity. As shown in Figure 17, we can clearly observe that FM-Track achieves much better performance compared to the other two methods, which demonstrates the effectiveness of the parameter refinement.

7.3 Starting Position Estimation

An accurate estimation of the starting position of the target is critical for the tracking afterwards. Many existing studies focus on tracking the 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 20 *cm* to 80 *cm* at a step size of 10 *cm*, and the AoA is varied from 60° to 120° at a step size of 10°.

Effectiveness of background subtraction. Background subtraction is a key to accurately estimate the starting position. We compare the estimation accuracy before and after background subtraction. As shown in Figure 18, the effect of the background subtraction is minimal when the target is close to the device. However, when the target is far away from the device, the accuracy after the background subtraction is much higher. The reason is that when the target is close to the sensing device, the target-reflected signal is strong and the interference signals from other surrounding objects can be neglected. However, when the target is further away, the target-reflected signals become much weaker and the interference signals could affect the tracking accuracy more significantly.

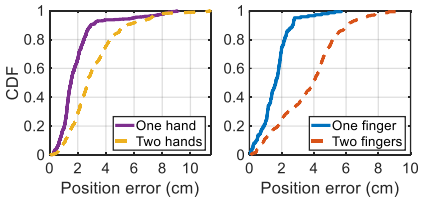
Impact of different positions. Figure 19 shows the starting position estimation accuracy at different positions with the background subtraction. We can see that good performance can be achieved in a sector region. The median accuracy for the range and AoA are 0.68 *cm* and 0.91°, respectively. Specifically, the region spans from $60-120°$ with a maximum distance of approximately 80 *cm*. The performance degrades when the target moves out of this region. The angle size depends on the speaker's radiation pattern [8]. The range is limited by the Signal-to-Noise Ratio (SNR), which can be improved by using a more powerful speaker or by employing additional signal processing techniques [17].

7.4 System Latency

Table 1 shows the processing time for each component of FM-Track. The total processing time is around 130 *ms*, which supports real-time tracking using samples collected in 160 *ms* (4 chirps with 40 *ms* duration for each chirp). Note that parameter estimation accounts for a significant part of the total processing time, which we will discuss in more detail below.

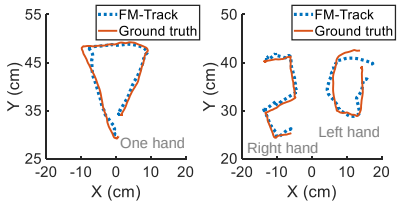
Choosing the subsampling factor D**.** Choosing a large D can reduce the computational time, but also decreases the number of samples in estimating parameters, hence degrading the accuracy. Based on the results shown in Figure 20a, we set $D=40$ to balance the accuracy and computational cost.

Number of targets. The average number of iterations for our algorithm is only 2.3, which is benefited from highly accurate starting position estimation. With more targets, we observe an increase of processing time. However, even for four targets, the median execution time for parameter estimation is as small as 115.5 *ms*.



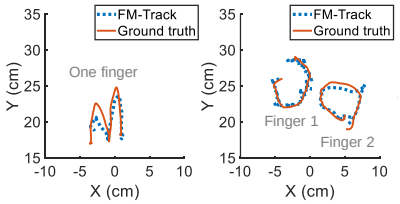
(a) Hand tracking. (b) Finger tracking.

Figure 21: Real-life performance.



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

Figure 22: Hand drawing samples.



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

Figure 23: Finger drawing samples.

7.5 Field Studies

In this section, we conduct experiments to track the movements of human hands and fingers. We convert the range-AoA position into X-Y coordinate and define a new metric to characterize the accuracy, i.e., position error. The position error is the Euclidean distance between the estimated position and ground truth measured by the motion capture system.

Hand Tracking. 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. Figure 21a shows the range and position errors for both one hand and two hands. 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 Figure 22.

Finger Tracking. Similarly to the hand tracking, a volunteer was asked to sit 0.6 *m* away from the device. First, the volunteer was asked to draw shapes with the index finger on the horizontal plane. Then, the volunteer was asked to use the index finger from each hand to simultaneously draw two shapes. As we can observe from Figure 21b, 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 Figure 23. Two fingers can be separated and accurately tracked without an issue with a spacing of 2 *cm*.

8 RELATED WORK

This section discusses three broad categories of prior studies that are relevant to the proposed study herein.

Contact-based acoustic tracking. Owing to its low propagation speed, acoustic signals have been widely employed for fine-grained motion tracking, where users are required to hold a device. AAMouse [40] tracks a mobile device based on Doppler shift at multiple frequencies. CAT [16] and Rabbit [18] employ distributed FMCW devices to estimate the distance between the transmitter and receiver. SoundTrak [42] extracts the phase information from the sinusoidal waves to track a customized finger ring with respect to a smartwatch in 3D space. MilliSonic [30] leverages the phase of the FMCW signal to enable concurrent tracking of multiple mobile devices. All these systems infer human motions by tracking movement of devices.

Contactless acoustic tracking. Many systems have been developed to perform contactless tracking without requiring users to carry any device. Efforts have pushed the granularity of tracking to a millimeter level. FingerIO [19] exploits the auto-correlation properties of OFDM symbols to achieve *mm*-level tracking, while LLAP [33], Strata [41] and the system proposed by Sun *et al.* [25] track fine-grained movements by capturing the phase change of signals. VSkin [26] 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 [17] enables room-scale hand motion tracking by combining a microphone array with a series of signal processing techniques and an RNN. However, these studies have focused mainly on single-target tracking, whereas FM-Track aims to enable fine-grained multi-target tracking.

Contactless multi-target tracking. There have been several attempts to contactlessly track multiple targets using other types of wireless signals, i.e., OFDM-based WiFi signals and chirp-based radar signals. For WiFi signals, WiDeo [11] and mD-Track [35] propose to leverage information from multiple dimensions including angle, ToF and Doppler shift, to isolate the superposed signals reflected from multiple targets. For radar signals, WiTrack2.0 [3] leverages its inherent large bandwidth to support multi-target tracking. All the previous multi-target tracking systems mainly 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 how to accurately track multiple targets using acoustic signals and showcase its feasibility and reliability in real-life applications, i.e, multi-hand and multi-finger tracking. 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 signal-based tracking such as velocity ambiguity and unavailability of instantaneous velocity. We believe the proposed methods can benefit other chirp-based signal tracking such as LoRa and Radar.

REFERENCES

[1] E. 2020. Al-202 high-fidelity usb acoustics system, 2020.

[2] S. E. 2020. Admp401 mems microphones, 2020.

[3] F. Adib, Z. Kabelac, and D. Katabi. Multi-person localization via {RF} body reflections. In *12th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 15)*, pages 279–292, 2015.

[4] Amazon. Amazon echo, 2020.

[5] J. Chauhan, Y. Hu, S. Seneviratne, A. Misra, A. Seneviratne, and Y. Lee. Breathprint: Breathing acoustics-based user authentication. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, pages 278–291. ACM, 2017.

[6] H. Chen, F. Li, and Y. Wang. Echotrack: Acoustic device-free hand tracking on smart phones. In *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*, pages 1–9. IEEE, 2017.

[7] S. Gupta, D. Morris, S. Patel, and D. Tan. Soundwave: using the doppler effect to sense gestures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1911–1914. ACM, 2012.

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

[9] Huawei. Huawei sound x, 2020.

[10] Q. Inc. Qualisys motion capture systems, 2020.

[11] K. Joshi, D. Bharadia, M. Kotaru, and S. Katti. Wideo: Fine-grained device-free motion tracing using {RF} backscatter. In *12th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 15)*, pages 189–204, 2015.

[12] V. Katkovnik, M.-S. Lee, and Y.-H. Kim. 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*, pages 543–547. IEEE, 2002.

[13] M. Kotaru, K. Joshi, D. Bharadia, and S. Katti. Spotfi: Decimeter level localization using wifi. In *ACM SIGCOMM computer communication review*, volume 45, pages 269–282. ACM, 2015.

[14] H. W. Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[15] X. Li, D. Zhang, Q. Lv, J. Xiong, S. Li, Y. Zhang, and H. Mei. Indotrack: Device-free indoor human tracking with commodity wi-fi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–22, 2017.

[16] W. Mao, J. He, and L. Qiu. Cat: high-precision acoustic motion tracking. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, pages 69–81. ACM, 2016.

[17] W. Mao, M. Wang, W. Sun, L. Qiu, S. Pradhan, and Y.-C. Chen. Rnn-based room scale hand motion tracking. In *The 25th Annual International Conference on Mobile Computing and Networking*, page 38. ACM, 2019.

[18] W. Mao, Z. Zhang, L. Qiu, J. He, Y. Cui, and S. Yun. Indoor follow me drone. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, pages 345–358. ACM, 2017.

[19] R. Nandakumar, V. Iyer, D. Tan, and S. Gollakota. Fingerio: Using active sonar for fine-grained finger tracking. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 1515–1525. ACM, 2016.

[20] R. Nandakumar, A. Takakuwa, T. Kohno, and S. Gollakota. Covertband: Activity information leakage using music. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–24, 2017.

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

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

[23] W. Ruan, Q. Z. Sheng, L. Yang, T. Gu, P. Xu, and L. Shangguan. 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*, pages 474–485. ACM, 2016.

[24] S. Shen, D. Chen, Y.-L. Wei, Z. Yang, and R. R. Choudhury. Voice localization using nearby wall reflections. In *The 25th Annual International Conference on Mobile Computing and Networking*. ACM, 2020. Preprint for MobiCom 2020.

[25] K. Sun, W. Wang, A. X. Liu, and H. Dai. Depth aware finger tapping on virtual displays. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*, pages 283–295. ACM, 2018.

[26] K. Sun, T. Zhao, W. Wang, and L. Xie. 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*, pages 591–605. ACM, 2018.

[27] B. Team. Bela platform, 2020.

[28] C. Team. Cozmo smart robot, 2020.

[29] H. Team. 4wd hercules mobile robotic platform, 2020.

[30] A. Wang and S. Gollakota. Millisonic: Pushing the limits of acoustic motion tracking. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 18. ACM, 2019.

[31] J. Wang, J. Xiong, H. Jiang, X. Chen, and D. Fang. D-watch: Embracing "bad" multipaths for device-free localization with cots rfid devices. *IEEE/ACM Transactions on Networking*, 25(6):3559–3572, 2017.

[32] T. Wang, D. Zhang, Y. Zheng, T. Gu, X. Zhou, and B. Dorizzi. C-fmcw based contactless respiration detection using acoustic signal. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):170, 2018.

[33] W. Wang, A. X. Liu, and K. Sun. Device-free gesture tracking using acoustic signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, pages 82–94. ACM, 2016.

[34] Y. Xie, F. Li, Y. Wu, S. Yang, and Y. Wang. D 3-guard: Acoustic-based drowsy driving detection using smartphones. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pages 1225–1233. IEEE, 2019.

[35] Y. Xie, J. Xiong, M. Li, and K. Jamieson. md-track: Leveraging multi-dimensionality for passive indoor wi-fi tracking. In *The 25th Annual International Conference on Mobile Computing and Networking*, pages 1–16. ACM, 2019.

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

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

[38] X. Xu, J. Yu, Y. Chen, Y. Zhu, L. Kong, and M. Li. 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*, pages 54–66. ACM, 2019.

[39] L. Yang, Y. Chen, X.-Y. Li, C. Xiao, M. Li, and Y. Liu. 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*, pages 237–248, 2014.

[40] S. Yun, Y.-C. Chen, and L. Qiu. Turning a mobile device into a mouse in the air. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*, pages 15–29, 2015.

[41] S. Yun, Y.-C. Chen, H. Zheng, L. Qiu, and W. Mao. Strata: Fine-grained acoustic-based device-free tracking. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, pages 15–28. ACM, 2017.

[42] C. Zhang, Q. Xue, A. Waghmare, S. Jain, Y. Pu, S. Hersek, K. Lyons, K. A. Cunefare, O. T. Inan, and G. D. Abowd. Soundtrak: Continuous 3d tracking of a finger using active acoustics. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(2):1–25, 2017.

[43] M. Zhang, Q. Dai, P. Yang, J. Xiong, C. Tian, and C. Xiang. 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):55, 2018.