EchoTrack: Acoustic Device-free Hand Tracking on Smart Phones

Huijie Chen, Fan Li
School of Computer Science, Beijing Institute of Technology
Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications
Beijing, China chenhuijie, fli@bit.edu.cn

Yu Wang
Department of Computer Science
University of North Carolina at Charlotte
Charlotte, NC, USA
yu.wang@uncc.edu

***Abstract*—This paper explores the limits of acoustic ranging on smart phone in the scenario of device-free hand tracking. Tracking the hand is challenging since it requires continuously locating the moving hand in the air with fine resolution. Existing work on hand tracking relies on special hardware or requires users hold the mobile device. This paper presents EchoTrack, which continuously locates the hand by leveraging mobile audio hardware advances without special infrastructure supported. EchoTrack measures the distance from the hand to the speaker array embedded in smart phone via the chirp's Time of Flight (TOF). The speaker array and hand yield a unique triangle. The hand can be located with this triangular geometry. The trajectory accuracy can be improved with the method of Doppler shift compensation and trajectory correction (*i.e.*, roughness penalty smoothing method). We implement a prototype on smart phone and the evaluation shows that EchoTrack can achieve tracking accuracy within about three centimeters of 76% and two centimeters of 48%.**

I. Introduction

Gesture recognition technology is becoming increasingly popular as a fundamental HCI manner. Users can even interact with various electronic devices in a hand-free manner. This is particularly helpful in the situation that a user's hands are dirty, wet or inconvenient to touch the device. While the existing system, such as Soundwave [1] and WiGest [2], require users to remember the corresponding relations between the hand gesture and the function. Additionally, they only can identify some predefined gestures and being tailored for specific applications. Motion tracking system, such as WiDraw [3] and AAmouse [4] have recently been proposed to help overcome these limitations. However, WiDraw tracks the user's hand with high accuracy with at least a dozen of transmitters and is easy to be disturbed in multi-user situation. AAmouse requires the user to hold the phone and the controlled device requires additional external sensors (two speakers). Some commercial hardware systems such as Kinect [5] and Leap Motion [6] can also track hand motion, however, they require to use dedicated hardware and are not portable.

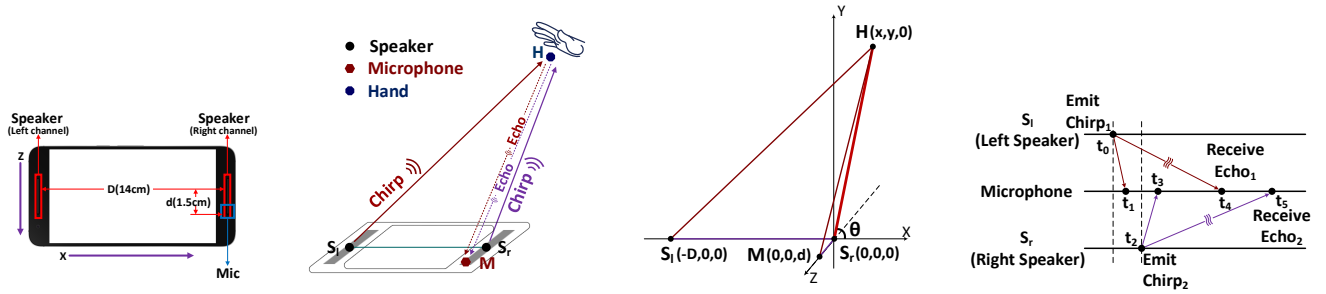
Audio localization technology has been explored with smart phone to achieve centimeter level accuracy in various applications, such as phone-to-phone localization [7], [8], neighbor discovery [9], mobile motion games [10], [11], whereas these techniques need to hold the smart phone in hand. In recent years, the sensors embedded in phone become increasingly diverse and powerful, for example, mobile phone hardware increasingly supports high definition audio capabilities. In particularly, Nexus 6P has two speakers for stereo playback, which are placed on the top and at the bottom of the phone. Besides, its audio chips are capable of recording acoustic signal up to $192kHz$. Such advances could have a significant impact on the accuracy and device-free manner of audio localization.

We introduce EchoTrack, a device-free hand tracking system which can be implemented on commercial phones. The main idea behind EchoTrack is to actively measure the distance from hand to the speaker array on phone by letting phone speakers emit sound and phone microphone sense its reflection. In specific, the hand can be located by a triangle defined by three edge, *i.e.*, the two edges from hand to speaker array on smart phone and one edge defined by the speaker array. Since the length of edge defined by speaker array is fixed, the key challenge is accurate ranging the distance from the hand to each unit in the speaker array. Moreover, the distance from hand to the speaker can be calculated using Time of Flight (TOF), *i.e.*, the time difference between the speaker emits the signal and the microphone records its echo reflected by the hand. However, it may be interfered by multipath effects, environmental noises, imperfect acoustic hardware of phone and Doppler shift of hand movement. Besides, the hand is located with device-free manner, thus the echo detection with low SNR will reduce the ranging accuracy.

EchoTrack addresses the above challenges by applying the following methods. The echo reflected from hand is too weak to be detected clearly, it can be overcome by utilizing chirp with the characteristics of low cross-correlation and high auto-correlation and adding a delay time for the chirp emitted from the other channel. Meanwhile, an enough idle time is added between each sensing period for preventing the interference of multipath effect. Besides, we analyze and compensate the Doppler shift from hand movement and employ smoothing method to improve the trajectory accuracy.

The main contributions of the paper are:

- We demonstrate that a single smart phone can track the moving hand with device-free manner by exploiting acoustic ranging while the existing work requires users to hold the phone in hands, or need specialized hardware.



(a) Nexus 6P with 2 speakers (b) Localization scenario (c) Geometric model (d) Event sequences

Fig. 1: The geometric model transformed from smart phone and the ranging procedure. (a) Front view of Nexus 6P equipped with top and bottom speaker. (b) Localization scenario, in which the chirp is emitted from speaker and its echo is recorded by microphone. The path lengthof chirp flying is calculated with TOA and used to locate the hand. (c) Geometric model for estimating the coordinate of hand. (d) Event sequences in the ranging procedure.

- We design a device-free hand tracking system, which leverages audio hardware advances (*i.e.*, stereo speakers with their high response frequency) of smart phone to track the hand. We design two-channel chirp for accurately calculating TOF and removing the interference of multipath. The trajectory accuracy is improved by deigned method of Doppler shift compensation and roughness penalty smoothing.
- We implement our system on commodity phones. It achieves average motion tracking accuracy of 76% within three centimeters and 48% within two centimeters.

The paper is organized as follows. Sec. II presents the design overview of our system. Detailed descriptions of our system to enable static and continues localization are given in Sec. III and Sec. IV, respectively. Sec. V presents the implementation details and Sec. VI provides our evaluation results. Sec. VII describes related work. Finally, Sec. VIII concludes the paper.

II. Design Overview

Recently, with the emergence of the phone supported stereo playback, we explore the realization of device-free hand tracking on this stereo output phone. The phone supported stereo output is equipped with two homogeneous speakers and spans a wide variety of off-the-shelf phones and tablets, such as Samsung Nexus 10 tablets, HTC One M8, Sony Xperia Z2, Amazon Fire Phone, LG nexus 5X and Google Nexus 6P. Fig. 1(a) shows the front view of Nexus 6P, which is equipped with two speakers on the top and at the bottom of the phone. Moreover, it is also equipped with three microphones. Two of them are only used for noise cancelation, thus we use the microphone at the bottom of front view to record the chirp.

Fig. 1(b) shows the scenario of locating the hand. S_l and S_r correspond to the speaker on the top and at the bottom of the smart phone. M corresponds to the microphone at the bottom of the phone. In order to locate the hand H, speaker S_l and S_r send a chirp in turn to avoid conflict between each other. These chirps are reflected once encountering the hand H. The echoes of these two chirps are recorded by the microphone.

Fig. 1(c) shows the geometric model corresponding to the scenario of locating the hand. We construct a virtual three dimensional coordinate where the origin is speaker $S_r(0,0,0)$. The X-axis is aligned with the line between S_l and S_r, the Z-axis is aligned with the line between S_r and M, and the Y-axis is the line vertical to the screen of phone. In this coordinate, the speaker S_l and microphone M are located at $(-D,0,0)$ and $(0,0,d)$, respectively. Since the hand only waves in the XY-plane, so we can denote the position of hand as $H(x,y,0)$. Obviously, the target can be located with d_{S_rH} and angle θ, where d_{S_rH} represents the distance from S_r to the hand H, and θ is the angle between S_rH and X-axis. d_{S_rH} and θ can be computed from the geometry of the flying path of chirp ($d_{S_lHM} = d_{S_lH} + d_{HM}$ and $d_{S_rHM} = d_{S_rH} + d_{HM}$).

Fig. 1(d) shows the ranging procedure. S_l emits a signal at time t_0 and the microphone M records the signal at time t_1. Time t_0 is unknown, but it can be inferred through the known distance d_{S_lM} between microphone M and S_l. The signal emitted from S_l is reflected once if it encounters the hand and microphone M records its echo at time t_4. The right channel S_r performs the ranging procedure like S_l does. Then d_{S_lHM} and d_{S_rHM} can be estimated as

$$\begin{aligned} d_{S_lHM} &= (t_4 - t_0)\, c = (t_4 - t_1)c + d_{S_lM} \\ d_{S_rHM} &= (t_5 - t_2)\, c = (t_5 - t_3)c + d_{S_rM}, \end{aligned} \tag{1}$$

where d_{S_lM} and d_{S_rM} represent the distance between speakers (S_l and S_r) and microphone M, which are known from Fig. 1(a). c represents the sound speed flying in the air. Time t_1, t_3, t_4 and t_5 can be inferred from the clock of the microphone as the frame counter for smart phone.

Our system, called EchoTrack, includes the following four stages as shown in Fig. 2:

Initialization Stage: The system generates the predesigned two-channel chirp, then initializes the microphone and speakers so that the microphone has been turned on before the speaker works.

Sensing Stage: The microphone continues to record the sound while the speakers send the two-channel chirp peri-

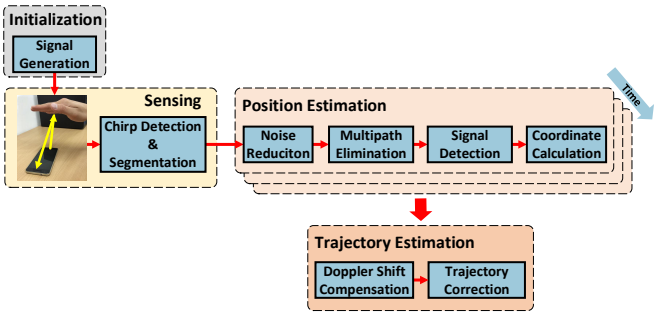
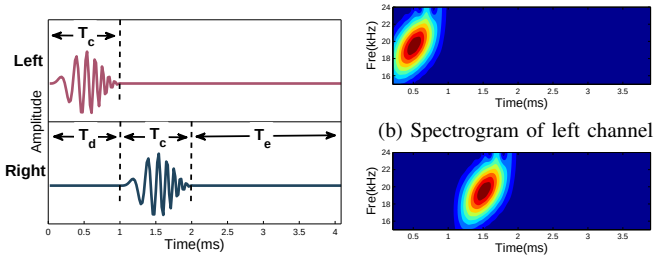


Fig. 2: The system overview of EchoTrack



(a) Time domain of two-channel chirp (b) Spectrogram of left channel (c) Spectrogram of right channel

Fig. 3: Structure of the two-channel chirp: (a) Two chirps will be emitted from the left and right channel. T_d is a delay added into chirp of the right channel for avoiding conflicts. (b) and (c) are the spectrogram of left and right channel.

odically. Then, the audio stream is partitioned into several fragments including original chirp and its echo.

Position Estimation Stage: This stage contains the following steps: noise reduction, multipath elimination, signal detection and coordinate calculation. Firstly, the sound fragment is processed with band-pass filter for eliminating the environment noises, and tailored for removing the multipath interference. Then, the envelope and peak detection locate the original chirp and echo. The round-trip time of chirp can be calculated as dividing the number of samples between the original chirp and echo by the sampling frequency. Finally, the coordinate is calculated using geometric model shown in Fig. 1(c).

Trajectory Estimation Stage: The Doppler shift resulted from hand movement brings out some position estimation error. Therefore, we utilize a compensation method for offsetting the error due to Doppler shift. Besides, we perform a trajectory correction method for improving trajectory accuracy.

III. Position Estimation

This section focuses on the discussion of our basic localization approach assuming that the smart phone and target are static, then the trajectory estimation of moving target is depicted in Sec. IV.

A. Signal Generation

This part discusses how to design the chirp for accurately ranging the distance between hand to each unit in speaker array. Firstly, we need to determine the frequency range of the chirp signal. In [12], frequency responses of various audio interfaces are measured in an anechoic chamber. It shows that the speakers are more frequency selective than microphones and microphone in commercial phone only can record the signal with frequency below $24kHz$. Most of the acoustic based data communication systems use high audio frequency zone which is beyond normal human perception ($16kHz - 22kHz$). Moreover, the majority of the background noises, such as human conversation, music and FM radio have frequencies up to $14kHz$, so we assign the frequency of chirp ranging from $16kHz$ to $23kHz$, which means that the signal bandwidth is $7kHz$. In this frequency band, the ranging accuracy is $2.4cm$, because

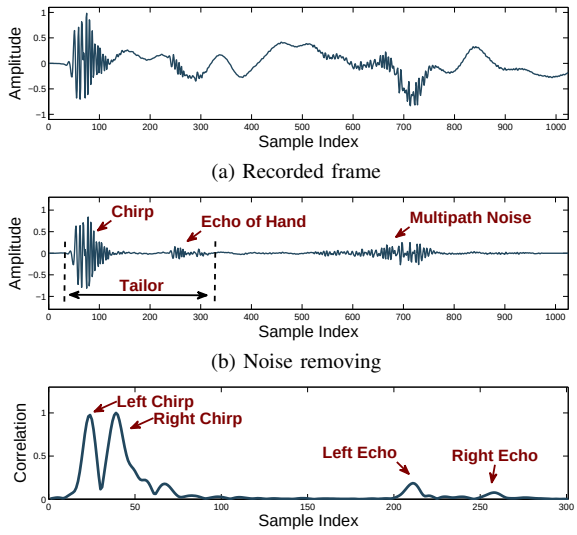
$$\delta = \frac{c}{2B} = \frac{340\text{m/s}}{2 \times 7\text{kHz}} \approx 2.4\text{cm}, \tag{2}$$

where δ represents the ranging accuracy, c and B represent the sound speed in the air and the frequency bandwidth of chirp.

Secondly, we discuss the duration time of the chirp. The longer the duration time has, the more overlapped parts between the original chirp and echo have, which makes it is more difficult to locate the chirp and echo. But when the duration time is too short, the sound energy becomes so weak that the echo may not be detected because of the low SNR. Therefore, we set the duration time with $1ms$ empirically.

Finally, we discuss the conflict avoidance method between the left and right channel of the speakers. There are multiple ways to facilitate multiple-access transmissions using chirp including TDMA and FDMA. Using frequency diversity is not ideal since the frequency bandwidth that it operates on is directly related to the chirp's timing resolution. Ideally, each channel would like to cover the maximum bandwidth to achieve the highest ranging resolution. TDMA suffers from configuration issues since all chirps would need to be scheduled in a collision-free manner. Fortunately, the synchronization between the left and right channel is controlled by the audio module embedded in phone processor. Therefore, we need to design the two-channel chirp with adding a delay in the emitted signal from right channel at each repetition.

Fig. 3(a) shows two-channel chirp in the time domain. The above chirp will be emitted from left channel and the below will be emitted from right channel. T_c and T_d are the chirp duration time and the time delay is added in the right channel signal to improve the detection for the echoes of left and right channel. We set T_c and T_d with $1ms$ empirically. T_e is used for avoiding the interference from multipath effect. Usually, the multipath effect can be ignored after the chirp spreads about $6m$ in indoor environment, since $\frac{6\text{m}}{340\text{m/s}} \approx 17.6\text{ms}$, so we set $T_e = 18ms$. Given these parameters, speaker will emit this chirp once every $20ms(TC_t)$ and achieve a frame rate of $50Hz$. Fig. 3(b-c) show the spectrogram of two-channel chirp. The left and right chirp are generated with up-chirp to make them robust to ambient noises and multipath interference.



(a) Recorded frame

(b) Noise removing

(c) Envelope of match filter

Fig. 4: Detection of chirp and its echo: (a) a originally recorded frame containing the two-channel chirp and echo, (b) the signal after noise removing, (c) the envelope of match filter's output.

B. Chirp Path Measurement

Trajectory estimation needs smart phone to periodically send the two-channel signal and continuously record the sound. The sending process of the two-channel signal can be realized via audio file which includes multiple two-channel signals. The recording thread puts the recorded audio buffer into a cache periodically. The audio processing thread continues to get the audio raw data from the cache and measures the path length of left and right chirp by the following steps.

Framing the recorded sound: We set the length of sliding window according to sensing period time. The overlapped part between sliding window is set to 10% of window's length. Then the recorded sound is cut to frames, which contains the whole two-channel chirp and its echo as shown in Fig.4(a).

Noise reduction and multipath removing: We adopt a band-pass filter to remove background noises which may be caused by air conditioner, human talking and music playing. The low and high cutoff frequency of the band-pass filter are set to $16kHz$ and $23kHz$ respectively. Fig. 4(b) shows the recorded audio processed by band-pass filter. Besides, the indoor multipath noises disturb the echo detection, so we cut out this frame to remove the multipath noises from surrounding obstacles $1m$ away as shown in Fig. 4(b).

Envelope detection: Match filter is adopted widely in radar because of better detection effect to the echo, so it has been applied on ranging technology [13]. The output of match filter contains several small peaks that might interfere with the echo detection. In order to solve this problem, we perform the envelope detection for the match filter's output as follows

$$\mathcal{E}(p,s) = \mathcal{F}^{-1}\left\{\mathcal{Z}\left\{\overline{\mathcal{F}\{p\}}\cdot\mathcal{F}\{s\}\right\}\right\}, \tag{3}$$

where p and s represent the windowed chirp and the recorded signal. $\mathcal{F}$ and $\mathcal{F}^{-1}$ represent Fourier transform and the inverse Fourier transform. $\overline{\mathcal{F}\{p\}}$ represents the conjugate transform of the windowed chirp after Fourier transform. $\mathcal{Z}$ is a function of a one-dimensional vector $X = [x_1, \cdots, x_n]$. This function is written as $\mathcal{Z}(X) = [\mathcal{Z}(x_1), \cdots, \mathcal{Z}(x_n)]$, where $\mathcal{Z}(x_i) = 2x_i$ when $x_i > 0$, and $\mathcal{Z}(x_i) = 0$, otherwise. Fig. 4(c) shows the envelope of the match filter's output, which is easy to identify the frame indexes of the chirps and their echoes.

Peak detection: There are four peaks shown in Fig. 4(c), which correspond to the chirp in left and right channel and their echoes. Our system requires to automatically detect these peaks and calculate the distance between them. The process of peak detection is performed as follows. Firstly, the output of envelope detection needs to be smoothed in sliding window for removing the outlier. The size of the window is set with 15 empirically. Secondly, our system detects all peaks and sort these peaks by their amplitudes in ascending order. Finally, the highest four peaks are identified as the original two-channel chirp and their echoes.

Distance measurement: From previous steps, we can get the frame indexes of original chirp and echo. It assumes that the frame indexes of the chirp from left channel and its echo is I_0 and I_1 and the frame indexes of the chirp from right channel and its echo is I_2 and I_3. Then, the flying path of chirp (d_{S_lHM} and d_{S_rHM}) that send from left and right channel to microphone can be obtained as follows.

$$\begin{aligned} d_{S_lHM} &= (t4-t1)c + d_{S_lM} = \frac{I_1 - I_0}{f_s}c + d_{S_lM} \\ d_{S_rHM} &= (t5-t3)c + d_{S_rM} = \frac{I_3 - I_2}{f_s}c + d_{S_rM}. \end{aligned} \tag{4}$$

C. Coordinates Calculation

The coordinate can be calculated with the geometry of $\triangle S_lHS_r$, so the key is estimating the length of edge d_{S_lH} and d_{S_rH} as shown in Fig. 1(c). We know that $d_{S_lS_r} = D$, $d_{S_rM} = d$ and the path length d_{S_rHM} and d_{S_lHM} can be calculated using Equ. (4).

We assume that the hand only waves in the XY plane. We use Pythagoras theorem to estimate d_{S_rH} and the estimated value is denoted as $\hat{d}_{S_rH}$.

$$\hat{d}_{S_rH} = \frac{d^2_{S_rHM} - d^2_{S_rM}}{2d_{S_rHM}}. \tag{5}$$

Then, $\hat{d}_{S_lH}$ is obtained as

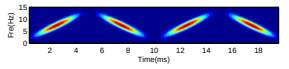
$$\hat{d}_{S_lH} = d_{S_lHM} - (d_{S_rHM} - \hat{d}_{S_rH}). \tag{6}$$

The angle θ can be achieved by the Law of Cosines

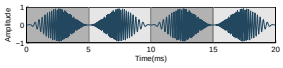
$$\theta = \pi - \arccos\frac{\hat{d}^2_{S_rH} + d^2_{S_lS_r} - \hat{d}^2_{S_lH}}{2\hat{d}_{S_rH}d_{S_lS_r}}. \tag{7}$$

Thus, the coordinate of hand is located as

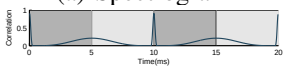
$$H(x,y,0) = (\hat{d}_{S_rH}\cos\theta, \hat{d}_{S_rH}\sin\theta, 0). \tag{8}$$



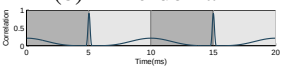
(a) Spectrogram



(b) Time domain



(c) Envelope of correlation with up-chirp

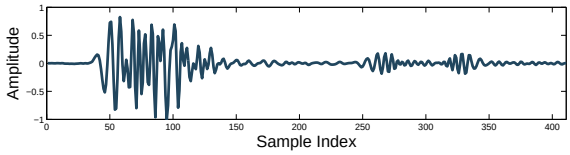


(d) Envelope of correlation with down-chirp

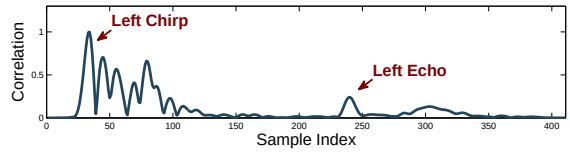
Fig. 5: High autocorrelation about chirp: (a) and (b) the spectrogram and time domain information of signal containing up-chirp and down-chirp. (c) and (d) the envelope of this signal the correlates with up-chirp and down-chirp.

D. Ranging Accuracy Improvement

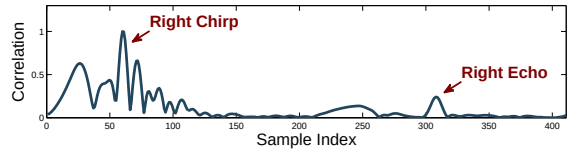
The key of our tracking system is to enable accurate distance ranging with fine resolution. However, we find that the left and right echo can be only distinguished clearly while the distance between them is above $5cm$, which means the frame index number between them is at least 14 if sound speed is set as $340m/s$). The echo peak detection error has a significantly influence on the ranging accuracy. One way to reduce echo peak detection error is adding delay of emitted sound in the right channel which expands the interval frame counter to eliminate the interference. In addition, we generate the left and right chirp using up-chirp and down-chirp. We assume that the left channel emits the sound first.



(a) Recorded signal



(b) Correlation with left chirp



(c) Correlation with right chirp

Fig. 6: Improvement for echo detection. (a) Recorded frame contains left and right chirp and their echoes. (b) Correlation with the signal emitted from left channel. (c) Correlation with the signal emitted from right channel. The first two peaks in (b) and (c) correspond to the original signal and its echo.

Fig. 5 exhibits that this sweep signal has the advantages of low cross-correlation and high autocorrelation, which facilitates precise detection and is robust to device motion, noises, and ambient sounds. We generate a signal which combines with four windowed chirps, the first and third fragment is up-chirp and the second and fourth fragment is down-chirp. Fig. 5(a) and (b) show its frequency and time domain information. Fig. 5(c) and (d) show the envelope of cross-correlation between this signal and up-chirp, down-chirp. Obviously, up-chirp fragment is detected in time $0ms$ and $10ms$ and down-chirp fragment is detected in time $5ms$ and $15ms$.

This method is used for two purposes. First, there are four peaks in the output of match filter when the intentional delay is added in the emitted sound from right channel as Fig. 4(c) shows. The peak index is obtained after peak detection. However, we cannot identify whether the echo is emitted from left or right channel. This method can overcome this problem with low cross-correlation between up-chirp and down-chirp. Second, this method helps to distinguish the overlapped echo. Fig. 6(a) shows a signal combing with up-chirp, down-chirp and their echoes. Fig. 6(b) and Fig. 6(c) show the cross-correlation between this signal and up-chirp, down-chirp. we can see clearly that the chirp emitted from left channel and its echo is higher than the chirp from right channel in the Fig. 6(b). The same phenomenon is shown in Fig. 6(c).

TABLE I: Ranging error results from the Doppler shift

v (m/s)	f_D (Hz)	Ranging Error (cm)	
		48kHz	192kHz
3.9	263	1.4	1.1
3	203	0.7	0.7
2	135	0.7	0.53
1	68	0	0.18
0.5	33	0	0

IV. Trajectory Estimation

The trajectory can be obtained by taking sequential position estimation as fast as possible. We can improve it by overcoming the following challenges.

A. Doppler Shift Compensation

The radial velocity of the hand movement causes the Doppler shift, which increases the ranging error. The Doppler shift can be calculated by $f_D = \frac{v \cdot f_H}{c}$, where f_D corresponds to the Doppler frequency shift, v and c correspond to the speed of hand movement and sound. f_H represents the upper bound of the chirp frequency.

In the evaluation, the sound speed v is $340m/s$ and the encoded chirp's frequency ranges from $16kHz$ to $23kHz$, so $f_H = 23kHz$. The experiment in SoundWave [1] indicates that usually the fastest speed of the hand movement of the most users is $3.9m/s$. Therefore, we only discuss the effort that the radial speed of the hand movement is below $3.9m/s$. The parameters used for the chirp is the same as in Sec III-A. We test the effects of Doppler shift with sampling frequency of $48kHz$ and $192kHz$. The evaluation results are shown in Table I. Obviously, faster hand movement speed causes larger ranging errors.

The experiment in AAmouse [4] shows that the speed does not exceed $1m/s$ when a user moves a mobile device in hand. Additionally, we have conducted experiment on a group of

20 users, about 14 users' hand movement speeds are under $0.8m/s$. If users only make simple gesture such as pull and push, this result in faster speed of hand movement. However, if users want to draw complex (long and curve) line in the air, the speed will be slower. Notice that the effect of hand movement speed on ranging cannot be ignored and compensating the Doppler shift [14] can reduce the ranging error.

B. Trajectory Correction

This system collects lots of position coordinates through continuous localization. The trajectory which covers these points has large errors compared with the ground truth. Trajectory correction method is introduced to solve this problem. Roughness Penalty Smoothing (RPS) [15] is proposed to deal with the situation that the sampling data contains lots of singular points. The objective function S of the RPS is

$$S=\arg\min_{y^*}\left\{\sum_{i=1}^{n}\left(y\left(i\right)-y^*\left(i\right)\right)^2+\lambda\int\left(\partial^2 f\left(x\right)\right)^2 d\left(x\right)\right\}, \tag{9}$$

where $y(i)$ represents the sample point, $y^*(i)$ represents the smoothed point. The latter half of this equation is for penalty. λ is the penalty factor. $\int\left(\partial^2 f\left(x\right)\right)^2 d\left(x\right)$ represents the curvature of the smoothed trajectory.

First, we discuss how to determine $\partial^2 f\left(x\right)$. In fact, if the function $f\left(x\right)$ is a cubic spline, the later part of Equ. (9) can be transformed as

$$\int\left(\partial^2 f\left(x\right)\right)^2 d\left(x\right)=y^{*\mathrm{T}}Ky^*. \tag{10}$$

K can be obtained as $K=QR^{-1}Q^t$. Q and R are expressed as

$$Q=\begin{bmatrix} h_1^{-1} & 0 & \cdots & 0 \\ -h_1^{-1}-h_2^{-1} & h_2^{-1} & \cdots & 0 \\ h_2^{-1} & -h_2^{-1}-h_3^{-1} & \cdots & 0 \\ 0 & h_3^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & h_{n-1}^{-1} \end{bmatrix}_{n\times(n-2)} \tag{11}$$

$$R=\begin{bmatrix} \frac{1}{3}(h_1+h_2) & \frac{1}{6}h_2 & \cdots & 0 \\ \frac{1}{6}h_2 & \frac{1}{3}(h_2+h_3) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{3}(h_{n-2}+h_{n-1}) \end{bmatrix}_{(n-2)\times(n-2)}. \tag{12}$$

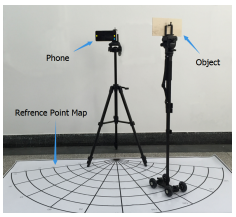
where n is the dimensionality of the smoothed data and $h_i=t_{i+1}-t_i$ $(t_1<t_2<\ldots<t_n)$. t_i represents the time slot of current sampled data. Since the system collects the sampled data periodically, h_i can be set with 1.

Based on the above equation, the objective function in Equ. (9) can be converted to

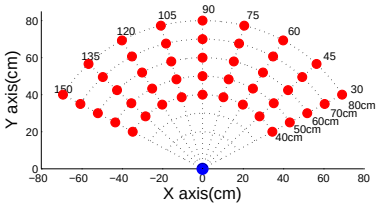
$$S=\arg\min_{y^*}\left\{y^t y-2y^{*t}y+y^{*t}\left(I+\lambda K\right)y^*\right\}. \tag{13}$$

In fact, $y^*=\left(I+\lambda K\right)^{-1}y$ is the optimal solution of S. The following part discusses how to determine λ. The method involves the cross-checking method

$$\arg\min_{\lambda}\left\{n^{-1}\sum_{i=1}^{n}\left(\frac{y\left(i\right)-y^*\left(i\right)}{I-A_i\left(\lambda\right)}\right)^2\right\}, \tag{14}$$



(a) Experimental Setup



(b) Reference Points

Fig. 7: Experimental scenario. (a) A Nexus 6P Phone is fixed on tripod located on point $(0,0)$, pulley is used to move the board along the predesigned path. (b) Red cycles are the positions of reference points. The distance to phone ranges from $30cm$ to $70cm$ and the angle ranges from 45 to 135 degree. The phone is placed on $(0,0)$.

where $A_i\left(\lambda\right)$ is the ith diagonal elements of the matrix A, $A=I+\lambda K$. It is worth noting that other trajectory correction methods (*i.e.*, kalman filter and particle filter) also have the similar effect.

V. Implementation

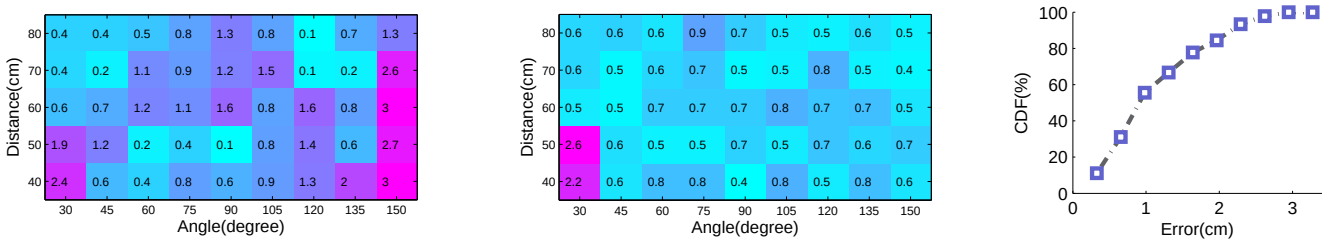
We implement our system for validating the feasibility of the hand localization. Our system is composed with a client and a server. The client sends the two-channel chirp periodically and continuously records the sound, thus the recorded sound will be transmitted to the server. The server performs the position calculation and returns the coordinates back to the client.

The smart phone acts as the client. In specific, the client was implemented on Nexus 6P with Android version 6.1. It supports sampling the sound signal with the rate of $192kHz$. We set the sampling rate of microphone as $48kHz$ for the consideration of processing overhead and requirement in chirp's frequency. In addition, the generation of chirp, recording of acoustic signal in client and transmitting of raw data are implemented with Android API. We use a PC with Intel i7 3.3GHz processor and 4GB memory as our server. The receiving and processing of raw data are implemented with JAVA 1.6. Client communicates with server via Wireless LAN, which is implemented with Sockets interface in JAVA library. This calculation can also be implemented using smart phone like other schemes.

VI. Evaluation

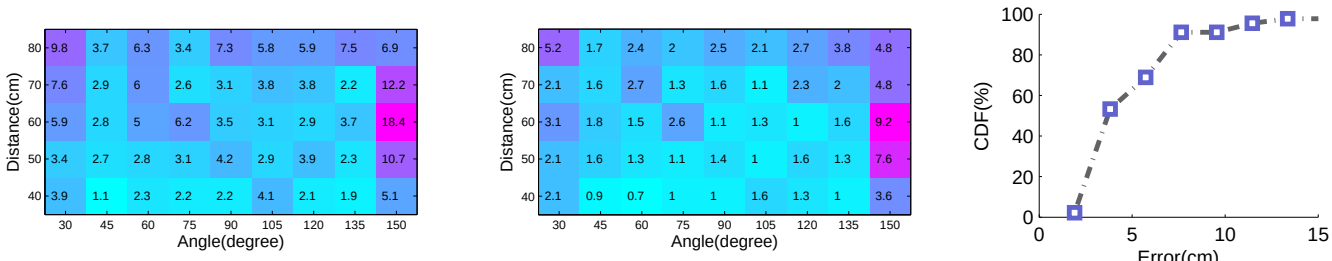
Experimental Setup: The experiment is conducted in a laboratory with the area of $5m\times6m$. The laboratory is not anechoic, therefore it exists multi-path effects and environment noises.

Fig. 7(a) shows the experiment setup. The phone is fixed on tripod. The wood board are fixed on tripod with a pulley at the bottom of this tripod for easy movement. The board and phone are place on the same hight with $70cm$. The size of wood board is $5cm\times18cm$, which is similar to the size of adult hand and it will be tracked by the phone. Some reference points with known coordinates are drawn on the floor to be the ground truth position precisely. Fig. 7(b) shows the locations



(a) Mean error (b) Standard deviation (c) CDF

Fig. 9: Ranging estimation error.



(a) Mean error (b) Standard deviation (c) CDF

Fig. 10: Static localization estimation error.

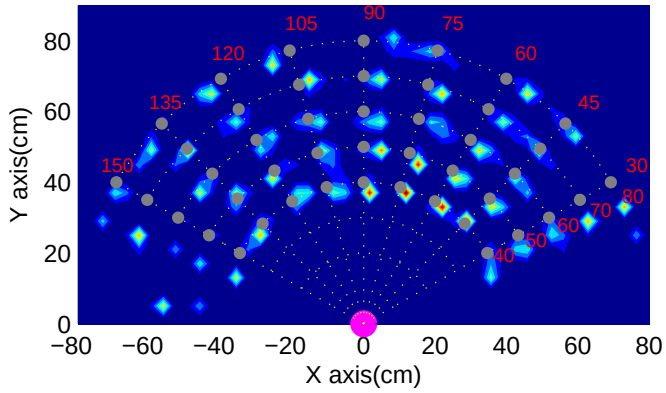


Fig. 8: Static localization distribution map. Each reference point is denoted as gray dot.

of the reference points. The origin of the coordinate system is the phone.

For the static localization, we place the tripod with pulley on each reference point. For continuous localization, the sensing frequency is set as $40Hz$. It is difficult to determine the ground truth while the target is moving. We apply a method in which manually moves the target along the motion path passing through some reference points. The average hand moving speed is $80cm/s$.

We consider the following evaluation metrics:

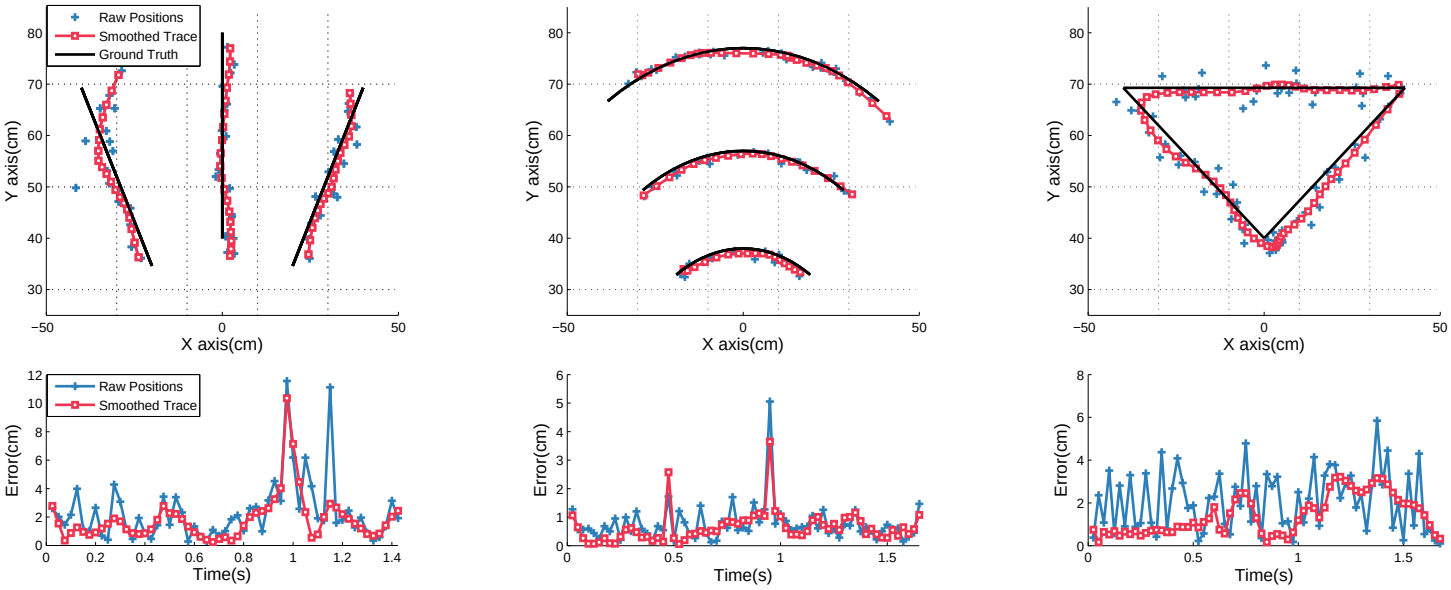
1) **Ranging Error:** the difference between the measured distance and the ground truth distance. Our primary target is small ranging mean error and standard deviation. The results for mean and standard deviation are conducted over 20 trials.
2) **Localization Error:** the Euclidean distance between the calculated coordinate and its corresponding reference point.
3) **Trajectory Error:** the Euclidean distance between the the coordinate of each point in trajectory and the ground truth.

A. *Performance of Distance Measurement and Localization*

Fig. 8 shows the visualized distribution results. In this figure, the reference point is plotted in gray dot to indicate the distance with estimated position. It is seen that when the location is close to the phone and near to the front of phone, the estimations are closer to their ground truth reference points. Distance Measurement from target to phone is a crucial part of EchoTrack. It directly affects the accuracy of the position estimation, so we first look at the ranging performance.

Fig. 9(a-b) show the mean ranging error and standard deviation. It is seen that the locations near the front of the phone have the position error about $1.2cm$. The locations on the sides of phone have larger error about $2.5cm$. Besides, the locations on the sides of the phone are more unstable than the locations near the front of phone. Fig. 9(c) shows that the mean ranging error within $2cm$ of 80%.

Fig. 10(a-b) show the static localization's mean error and standard deviation of ranging error. It confirms our previous measurement observations that the mean error of location estimation is larger toward the periphery. Besides, it also shows that the static localization has larger mean error compared



(a) straight line path and error (b) arc line path and error (c) triangle path and error

Fig. 11: Raw and smoothed position estimates for the paths of straight line, arc line and triangle.

to ranging module which is resulted from the accumulated ranging error. Fig. 10(b) shows that the ranging has higher stability, the reason is that each position is accurate placed, but the static localization is relatively unstable. Fig. 10(c) shows the mean error of localization is within $7cm$ of 80%. We found that a higher accuracy resulted from the fact that the target is facing the mobile phone. In this case, there is a stronger echo reflected to the mobile phone, which leads to the detection of peak position in the matched filter results more accurate.

B. Performance of Continuous Localization

To evaluate the performance of continuous localization, three motion patterns are tested by fixing the smart phone and moving the target along the predefined motion patterns. We choose three basic patterns: three straight lines, three arc line paths and a triangle path with sharp turns which are shown in Fig.11. For each motion path, we compare the accuracy of position estimation between the sampled raw data and the smoothed data with the manner of roughness penalty. We also compare the measurement error versus ground truth for each motion path.

Fig. 11 shows visual inspection effect and error estimates of individual motion paths. Firstly, it shows that the estimated trajectories of curve and triangle paths are acceptable with 0 to $4cm$ error, however, estimation of straight line path has larger error. The reason is that the radial velocity of the hand movement (*i.e.*, moving forward and away from the phone) causes the Doppler shift, which increases the ranging error. Secondly, it indicates that the roughness penalty smoothing method substantially improves the accuracy of estimated trajectories, however, the sharp turns are difficult to compensate, especially for the vertices of the triangle.

Fig. 12 shows the CDF for the position estimation errors across the three motion paths. Position estimation with roughness penalty smoothing achieves less than $3cm$ position error of 76% and $2cm$ position error of 48%. Using raw data alone results in less than $3.6cm$ position error 80% of the time and less than $2.8cm$ position error over 50% of the time.

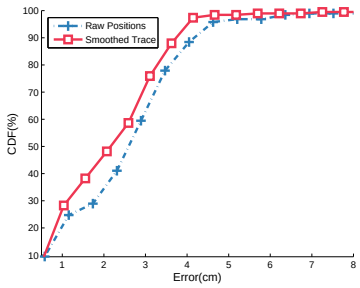


Fig. 12: CDF of position error across all paths.

VII. Related Work

Over the past decade, the acoustics-based approach has provided insights for localization and tracking. In this section, the systems and approaches will be reviewed where acoustic signals are used for locating or tracking.

Acoustics Localization: High-precision position can be achieved with accurate ranging. The ranging accuracy depends on the signal speed and the precision of TOA measurement. BeepBeep [7] ranges the distance between two phone without time synchronization. To identify the mobile phone user is whether a driver or passenger, detect driving phone system [16] assumes that the phone position near the driver's seat

means the phone user is a driver. The proposed acoustic relative-ranging system locates the phone through the processing which records some sound clips from the speakers of four corners inside the car and determines the different TOA. EchoLoc [17] is a hand localization solution which can be enabled on COTS devices. It leverages the microphone embedded on smart phone and stereo speaker. A TDoA based localization scheme [18] is designed for grouping and locating mobile phone users in proximity, which uses internal motion sensors and speakers/microphones to extract the mobility-differentiated TOA. Moreover, CondioSense [19] recognizes the phone context(*i.e.*, pocket, bag, car indoor and outdoor) using active audio sensing. It captures the attenuation and interference effects influenced by propagation medium. Unlike the above approaches, our system is a device-free scheme that locates the hand through acoustics ranging.

Motion Tracking: MoveLoc [20] can track the moving objects without carrying any devices (Device-free). The basic idea is mapping the Doppler frequency shifts of the echo reflected off moving object to the Angle of Arrival (AoA) from beacons. This scheme requires special hardware and cannot be implemented on smart phone. Accurate Air Mouse (AA-Mouse) [4] can track hand movement with high accuracy in real time, but this scheme needs the controlled device equipped with special hardware. Tracko [21] tracks surrounding devices in 3D space using the time-of-flight of the inaudible signals. It exchanges a series of inaudible stereo sounds and derives a set of accurate distances between devices from the differences in their arrival times. Our scheme can be implemented on smart phone and without any special hardware, it is implemented with device-free manner which means it does not require users to hold the phone in hand. FingerIO [22] tracks the finger movement by transforming the device into an active sonar system. In contrast, our system can apply in more complex scenario, which contains surrounding noises and indoor multipath effect. Besides, we use a compensation method to offset Doppler shift due to hand movement. Meanwhile, we utilize a smoothing method for improving the accuracy of tracking trajectory.

VIII. Conclusion

Hand tracking is becoming increasingly popular as a fundamental HCI approach. Users can even interact in-air with various applications running on surrounding electronic devices in the hands-free manner. Our work demonstrates that it is possible to leverage sound signals from smart phone to track the hand motion in the air. We introduce EchoTrack, a tracking motion solution that can be enabled on smart phone as an application. EchoTrack leverages the speaker array and microphone on smart phone to track the detailed trajectory of the user's hand near the smart phone, without requiring the user to touch the device or hold any hardware. We design the two-channel chirp to remove the indoor multipath noise and improve the trajectory accuracy using Doppler shift compensation and roughness penalty smoothing method. We have implemented EchoTrack system in the smart phone and evaluated our tracking motion scheme. The performance shows EchoTrack is capable of continuous localization resolution within mean three centimeters of 76% and two centimeters of 48%. In the future work, we will design a dynamic scheduling mechanism for energy-efficiency and employ the conflict avoidance method (TDMA and CDMA) to be used in multi-user scenarios.

ACKNOWLEDGEMENT

The work is partially supported by the National Natural Science Foundation of China under Grant No. 61370192, 61432015, 61428203 and 61572347, and the US National Science Foundation under Grant No. CNS-1319915 and CNS-1343355. F. Li is the corresponding author.

References

[1] S. Gupta, D. Morris, S. Patel, and D. Tan, "SoundWave: using the doppler effect to sense gestures," in *ACM SIGCHI*, 2012, pp. 1911–1914.

[2] H. Abdelnasser, M. Youssef, and K. A. Harras, "WiGest: A ubiquitous WiFi-based gesture recognition system," in *IEEE INFOCOM*, 2015.

[3] L. Sun, S. Sen, D. Koutsonikolas, and K.-H. Kim, "WiDraw: Enabling hands-free drawing in the air on commodity WiFi devices," in *ACM Mobicom*, 2015, pp. 77–89.

[4] S. Yun, Y.-C. Chen, and L. Qiu, "Turning a mobile device into a mouse in the air," in *ACM Mobisys*, 2015, pp. 15–29.

[5] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE MultiMedia*, vol. 19, no. 2, pp. 4–10, 2012.

[6] F. Weichert, D. Bachmann, B. Rudak, and D. Fisseler, "Analysis of the accuracy and robustness of the leap motion controller," *Sensors*, vol. 13, no. 5, pp. 6380–6393, 2013.

[7] C. Peng, G. Shen, et al., "Beepbeep: a high accuracy acoustic ranging system using COTS mobile devices," in *ACM Sensys*, 2007, pp. 1–14.

[8] J. Qiu, D. Chu, X. Meng, and T. Moscibroda, "On the feasibility of real-time phone-to-phone 3D localization," in *ACM Sensys*, 2011.

[9] K. Wang, Z. Yang, Z. Zhou, Y. Liu, L. Ni, "Ambient rendezvous: Energy-efficient neighbor discovery via acoustic sensing," in *IEEE INFOCOM*, 2015.

[10] Z. Sun, A. Purohit, R. Bose, and P. Zhang, "Spartacus: spatially-aware interaction for mobile devices through energy-efficient audio sensing," in *ACM MobiSys*, 2013, pp. 263–276.

[11] Z. Zhang, D. Chu, X. Chen, and T. Moscibroda, "Swordfight: Enabling a new class of phone-to-phone action games on commodity phones," in *ACM MobiSys*, 2012, pp. 1–14.

[12] H. Lee, T. H. Kim, J. W. Choi, and S. Choi, "Chirp signal-based aerial acoustic communication for smart devices," in *IEEE INFOCOM*, 2015.

[13] D. Graham, et al., "A software-based sonar ranging sensor for smart phones," *IEEE Internet of Things J.*, vol.2, no.6, pp.479–489, 2015.

[14] C. Wang and J. D. Ellis, "Dynamic doppler frequency shift errors: measurement, characterization, and compensation," *IEEE Trans. on Instrumentation and Measurement*, vol. 64, no. 7, pp. 1994–2004, 2015.

[15] P. J. Green and B. W. Silverman, *Nonparametric regression and generalized linear models: a roughness penalty approach*, CRC Press, 1993.

[16] J. Yang, S. Sidhom, et al., "Detecting driver phone use leveraging car speakers," in *ACM Mobicom*, 2011, pp. 97–108.

[17] H. Chen, F. Li, and Y. Wang, "Echoloc: Accurate device-free hand localization using COTS devices," in *ICPP* 2016, pp. 334–339.

[18] H. Han, S. Yi, et al., "AMIL: Localizing neighboring mobile devices through a simple gesture," in *IEEE INFOCOM*, 2016.

[19] F. Li, H. Chen, X. Song, Q. Zhang, Y. Li, and Y. Wang, "CondioSense: High-quality context-aware service for audio sensing system via active sonar," in *Personal and Ubiquitous Computing*, pp, 1–13, 2016.

[20] K. Zhao, D. Fang, et al., "Poster: doppler effect based device-free moving object localization," in *ACM Mobicom*, 2014, pp. 441–444.

[21] H. Jin, C. Holz, and K. Hornbæk, "Tracko: Ad-hoc mobile 3D tracking using bluetooth low energy and inaudible signals for cross-device interaction," in *ACM UIST*, 2015, pp. 147–156.

[22] R. Nandakumar, V. Iyer, D. Tan, and S. Gollakota, "FingerIO: Using active sonar for fine-grained finger tracking," in *ACM CHI*, 2016.