EchoLoc : Accurate Device-free Hand Localization using COTS Devices

Huijie Chen, Fan Li
School of Computer Science, Beijing Institute of Technology
Beijing Engineering Research Center for High Volume Language
Information Processing and Cloud Computing Applications
Beijing, 100081, China
{2120140397, fli}@bit.edu.cn

Yu Wang
Department of Computer Science
College of Computing and Informatics
University of North Carolina at Charlotte
Charlotte, NC 28223, USA
yu.wang@uncc.edu

***Abstract*—Hand tracking systems are becoming increasingly popular as a fundamental HCI approach. The trajectory of moving hand can be estimated through smoothing the position coordinates collected from continuous localization. Therefore, hand localization is a key component of any hand tracking systems. This paper presents EchoLoc, which locates the human hand by leveraging the speaker array in commercial off-the-shelf (COTS) devices (*i.e.*, a smart phone plugged with a stereo speaker). EchoLoc measures the distance from the hand to the speaker array via the time of flight (TOF) of the chirp. The speaker array and hand yield a unique triangle, therefore, the hand can be localized with triangular geometry. We prototype EchoLoc on iOS as an application, and found it is capable of localization with the average resolution within five centimeters of 73% and three centimeters of 48%.**

***Keywords*-Hand Localization; Acoustic Ranging; Device-Free; COTS;**

I. Introduction

Hand tracking systems, such as WiDraw [1] and AAmouse [2], have recently been proposed to enable the users to interact with various devices more conveniently. However, WiDraw requires at least a dozen of APs to guarantee tracking accuracy and is easy to be disturbed in multi-user situation. AAmouse requires users to hold the phone and the controlled device requires additional sensors (*i.e.*, two speakers). Therefore, these solutions require dedicated hardwares or devices to perform hand tracking. Moving hand trajectory can be estimated through smoothing the position coordinates collected from continuous localization. Thus, the key underlying technical challenge for device-free hand tracking is how to design an accurate hand localization method without any special hardwares.

In recent years, the computation, sensing and storage capabilities of smart phones have been further improved, which enables numerous new sensing applications [3]–[5]. Moreover, mobile phone hardware is increasingly supporting high definition audio capabilities targeted at audiophiles. For example, the audio chips of iPhone 4s are capable of $20kHz$ playback and recording. Such advances could have a significant impact on the accuracy of acoustic ranging and localization.

In this paper, we introduce EchoLoc, a hand localization system performing acoustic ranging without dedicated hardware. EchoLoc leverages the microphone on a smart phone and the left and right channels in a stereo speaker to locate the user's hand near the phone. The two channels in the stereo speaker and the nearby hand yield a unique triangle. The distance between two channels is controlled by the user and can be easily measured. The left and right channel send a chirp and the microphone records its echo reflected by the user's hand. Then the estimation of the distance from the hand to channel is achieved via flying time of the chirp. The user's hand can be located using triangular geometry.

The main contributions of this paper are summarized as follows:

- We demonstrate that commercial off-the-shelf (COTS) devices (*i.e.*, a smart phone and a stereo speaker) can locate the hand by exploiting acoustic ranging. Existing works need users to hold the phone in hand or require specialized hardwares.
- We develop a hand localization system, which leveraging a speaker array to locate the hand. The system designs a two-channel chirp to synchronize the speaker array, sends a chirp and records the echo to measure the distance from the hand to the each speaker and exploits the geometry-based information to locate the hand.
- We implement our hand localization system on a commodity phone and a stereo speaker. Our preliminary investigation shows that it achieves the accuracy with average errors within five centimeters of 73% and three centimeters of 48%.

The paper is organized as follows. Section II explains the geometric model and overview of our system. A detailed description of our system to enable hand position estimation is then given in Section III. Section IV and Section V discuss the implementation details and our evaluation results, respectively. Section VI reviews related works, and Section VII finally concludes the paper.

II. System Overview

Audio localization technology has been explored with smart phones to achieve centimeter level accuracy in various

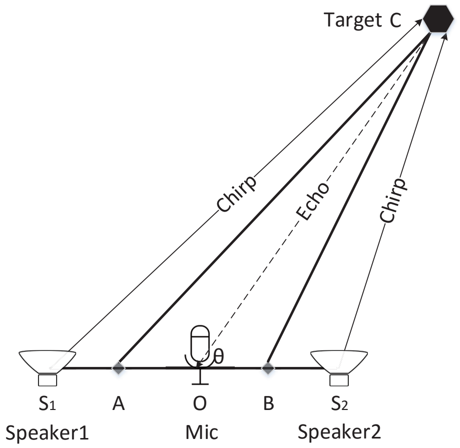


Figure 1: Geometric model. OC and $\theta = \angle COB$ are the distance and angle from the target to the phone, respectively.

applications, such as phone-to-phone ranging [6], $3D$ localization [6]–[8], mobile motion games [9], [10] and indoor localization [11], [12]. Whereas these techniques need to hold the smart phone in hand, we seek to achieve considerable accuracy with device-free fashion. Kinect [13] and Leap Motion [14] achieve high motion tracking accuracy, but they obviously cannot be implemented on COTS devices. Our design goal is to realize accurate hand localization without any special hardware.

Current smart phones are typically equipped with two speakers placed on the top and at the bottom of the phone. These two speakers group into a speaker array and enable acoustic localization. However, the top speaker is mainly used for calling and its frequency response is generally below $12kHz$. The sound with frequency response below $12kHz$ is clearly audible for human, so the speaker on the top of phone does not suitable for acoustic ranging in our system. Moreover, the commercial stereo speakers equipped with left and right channel are widely popular in human's life. It has flat frequency response and reaches $20kHz$, so the smart phone plugged with stereo speakers will improve the accuracy of acoustic localization.

Our hand localization system locates the hand via a smart phone plugged with a stereo speaker. In order to simplify localization, microphone embedded in smart phone is placed between left and right channel of stereo speaker. Fig. 1 shows the localization scenario with the geometric model. $Speaker1$ and $Speaker2$ correspond to the left and right channel of the stereo speaker. Mic corresponds to the microphone on the phone. Then, the remote target C, $Speaker1$ and $Speaker2$ yield a unique triangle. Obviously, the larger of the distance between $Speaker1$ and $Speaker2$, the higher of localization accuracy. However, the ranging model in sonar assume that the distance from target to the phone is much greater than the distance between the microphone and speakers. Considering with the restrictions of the arm length, the hand generally moving near the smart phone in the range about $70cm$. Thus, we place stereo speakers so that the distance between $Speaker1$ and $Speaker2$ is $15cm$. The microphone is placed on the bisects point between $Speaker1$ and $Speaker2$. In this geometric model, each speaker and microphone can be considered as one point. For example, Fig.1 shows that $Speaker1$ and microphone is considered as point A, while $Speaker2$ and microphone is considered as point B.

To locate the remote target C, our system will perform the following audio sensing procedure. $Speaker1$ and $Speaker2$ send a chirp in turn and interval a period of time for avoiding conflicts. The echo of chirp is recorded by the microphone one by one. Then, the path lengths of $\|S_1CO\|$ and $\|S_2CO\|$ can be calculated by multiplying flying time of the chirps with the sound speed. Next $\|AC\|$ is defined as the half of $\|S_1CO\|$, and $\|BC\|$ is defined as the half of $\|S_2CO\|$. With these information, we can estimate the distance $\|OC\|$ from the target to phone and the azimuth $\theta = \angle COB$ that target related to the phone. The target then can be located with the knowledge of $\|OC\|$ and θ.

The overall procedure includes the following sub-process stages:

1) **Initialization Stage:** The system generates the designed two-channel chirp and initializes the audio playing and recording thread so that the microphone has been turned on before the speaker plays.
2) **Sensing Stage:** The speakers send the two-channel chirp periodically. The microphone continues to record the sound. The raw data of recorded sound can be assessed by calling the OS's API.
3) **Ranging Stage:** The recorded sound will be framed as some fragments. Then these fragments will be processed by band-pass filter to eliminate noises. The match filter method recognizes the original chirp and echo for caculating the chirp's round-trip time.
4) **Position Estimation Stage:** The system locates the target through $\|AC\|$ and $\|BC\|$. The position estimation can be further improved by applying amendment procedure.

III. Hand Localization

A. Generating the Two-channel Signal

In this part, we discuss how to design the chirp for ranging and synchronizing the speaker's left and right channel. We first consider how to set the frequency range and duration time of the two-channel signal. For the frequency range of two-channel signal, the chirp needs to be inaudible for human and restricted by the sampling rate. Therefore, we set the frequency range of two-channel signal from $12Khz$ to $20Khz$. For the duration time of two-channel signal, we set

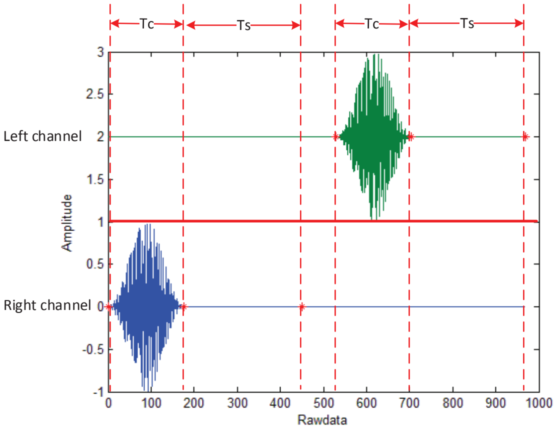


Figure 2: Structure of the two-channel chirp. The blue and green chirp control separately the right and left channel. T_c and T_s correspond to the chirp's duration time and interval time, for avoiding conflicts between right and left channels.

it with $4ms$ empirically considering with the sound energy and echo detection.

Many time synchronization solutions [15]–[19] have been proposed for synchronizing the entire sensor network to a common timing reference. We choose TDMA to synchronize between the left and right channel. The synchronization between the left and right channel can be controlled by the two-channel audio file. Therefore, we need to design the two-channel chirp to schedule the left and right channel.

Fig. 2 shows the time domain structure of the two-channel signal. In this figure, the above green signal controls the left channel, the below blue signal controls the right channel. The left and right channel send windowed up chirp for ranging in consideration of improving the pulse's signal to noise ratio. T_c is the chirp's duration time. T_s is the time interval for avoid the collision between the left and right channels. We set the $T_c = 4ms$ and $T_s = 6ms$. This setting limits the furthest measured distance $d_{max} = 1m$.

B. Processing the Signal

Accurate hand position estimation needs smart phone to send the two-channel signal periodically and record the sound continuously. The periodical sending process of the two-channel signal can be realized via an audio file which includes multiple two-channel signals. The recording thread put the recorded audio buffer into a cache periodically. The audio processing thread continues to get the audio raw data from the cache and handle it by the following steps.

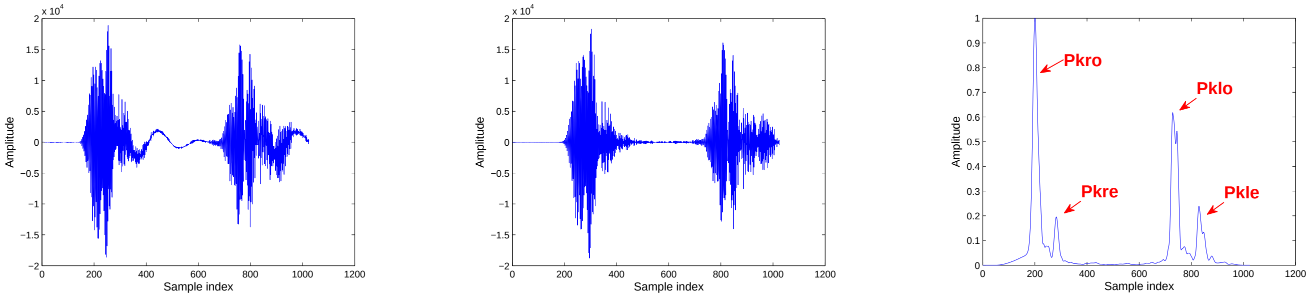
1) **Framing the recorded sound:** We set the length of sliding window according to sensing period time. The overlap part between sliding window is set to 10% of the window's length. Then the recorded sounds are framed to several fragments containing the whole two-channel chirp and its echo. Fig. 3(a) shows such an example frame.
2) **Noise reduction:** Various noises exist in human's life environment, such as air conditioner, talking and music. Noises interfere with the detection of chirp and its echo, thus we use a band-pass filter to remove the surrounding noises. The cut-off frequency of the band-pass filter ranges from $12khz$ to $20khz$. Fig. 3(b) shows the signal after processing by the band-pass filter.
3) **Envelope detection:** Matching filter is widely used in radar system resulted from better identify overlapped wave. The output of match filter corresponds to the correlation between chirp and frame. So the location of chirp and its echo corresponds to the peak in the output of match filer. However, the output of match filter contains several small outliers that interfere with the detection of echo. To mitigate this, we smooth the envelope of the output result. Fig. 3(c) shows the smoothed envelope of the match filter's output.
4) **Peak detection:** To estimate the chirp's flying time, the peak in the smoothed envelope in the previous step needs to be located. Obviously, there are four peaks existing in Fig. 3(c), where $Pkro$ and $Pkre$ respond to the right channel's chirp and its echo. Similarly, $Pklo$ and $Pkle$ respond to the left channel's chirp and its echo. We locate the peak with a sliding window method. The peaks extracted from each sliding window are sorted according to their amplitude, the first four peaks are the original two-channel chirps and their echoes.
5) **Distance Measurement:** From four steps above, we can obtain the time index of original chirp and its echo. We define the frame index of $Pklo$ and $Pkle$ as I_{lo} and I_{le} and the frame index of $Pkro$ and $Pkre$ as I_{ro} and I_{re}. Then $\|AC\|$ and $\|BC\|$ can be obtained from the following equations, where S_f is sampling rate and c is the speed of sound.

$$\|AC\| = \frac{I_{le} - I_{lo}}{2S_f}c. \tag{1}$$

$$\|BC\| = \frac{I_{re} - I_{ro}}{2S_f}c. \tag{2}$$

C. Calculating the Coordinates

Recall that Fig. 1 shows the geometric model of our localization method. In Fig. 1, $\|AC\|$ and $\|BC\|$ are the output of steps above that respond to the distance from the target to the speaker array. The distance from the target C to the phone is $\|OC\|$. Since we assume that $\|OC\|$ is much greater than the $\|AB\|$, then $\|OC\|$ can be estimated as



(a) Recorded frame (b) Denoised frame (c) Output of envelope detection

Figure 3: The processing of the recorded frame. (a) an original recorded frame which contains the two-channel chirp and its echo. (b) the denoised signal. (c) the envelope of the match filter's output for detecting the peak.

follows:

$$\|OC\| \approx \frac{\|AC\| + \|BC\|}{2}. \tag{3}$$

The angle $\theta = \angle COB$ can be obtained by the following equation.

$$\begin{aligned}\theta &= \arccos \frac{(\|AC\| - \|BC\|)(\|AC\| + \|BC\|)}{\|AB\| \times (2 \times \|OC\|}\\ &\approx \arccos \frac{\|AC\| - \|BC\|}{\|AB\|.}\end{aligned} \tag{4}$$

Recall that we set $\|AB\|$ between point A and B to $15cm$ as discussed in Sec.II.

Then, the coordinate (x, y) of the target C can be obtained as follows

$$(x, y) = (\|OC\| \cos \theta, \|OC\| \sin \theta). \tag{5}$$

IV. Implementation

We implement our system to validate the feasibility of the proposed device-free hand localization. Our system is composed with a client and a server. The client (a smart phone plugged with a stereo speaker) sends the two-channel chirp periodically and record the sound continuously, and the recorded sound will be transmitted to the server. The server performs the position calculation and returns the coordinate back to the client.

The smart phone together with the stereo speakers act as client. In specific, since the audio chips of iPhone 4s are capable of recording up to $20kHz$, we implement our hand localization system on iPhone 4s running iOS 9.1. We equip the phone with one stereo speaker, EDIFIER R18T, which has extremely flat frequency response all the way up to $20kHz$. The generation of chirp, the recording of acoustic signal in client and the transmitting of raw data are all implemented with iOS 9.1 API. We use a PC with Intel i5-4590 3.3GHz processor and 4GB memory as our server. The receiving of raw data and the processing of received data are implemented with Java 1.6 at the server side. Client communicates with the server via Wi-Fi based LAN, which is implemented with Socket interface in iOS and Java library. Of course, the calculation in the server can also be implemented at smart phone like in others schemes.

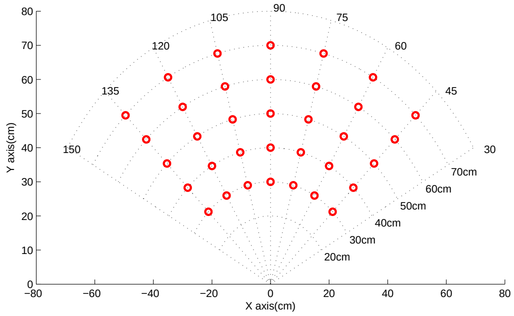


Figure 4: The red circles are the positions of the reference points. The distance to phone ranges from $30cm$ to $70cm$ and the angle ranges from 45 to 135 degree. The phone is placed on $(0, 0)$.

V. Evaluation

A. Experimental Setup

The experiment is conducted in a laboratory with area of $5m \times 6m$. The laboratory is not anechoic, therefore it exists multi-path effects and noises. We placed the phone and stereo speakers on the desk and placed the target outside of the phone. The located target is a rectangle cardboard with the size of $5cm \times 18cm$ which imitate the human hand. In order to control the ground truth position precisely, we draw some reference points on the floor. Fig. 4 shows the locations of the reference points. For the localization test, we place the target on each reference point.

B. Distance and Position Estimation

Distance measurement from the target to the phone is a crucial part of EchoLoc. It directly affects the accuracy of the position estimation, so we first look at the ranging performance. Fig. 5(a) and Fig. 5(b) show the ranging module's mean errors and standard deviations. Fig. 6(a) and Fig. 6(b)

		angle (degree)						
		45	60	75	90	105	120	135
distance (cm)	30	2.3	1.8	1.5	0.20	1.3	1.1	2.9
	40	2.6	1.8	1.1	0.20	0.3	1.6	2.3
	50	3.2	2.1	0.2	0.20	1.3	0.9	2.4
	60	3.4	1.63	1.2	1.20	2.2	0.8	1.2
	70	3.6	1	1.1	0.50	1.3	1.1	0.2

(a) Mean Error

		angle (degree)						
		45	60	75	90	105	120	135
distance (cm)	30	0.2	0.1	0.2	0.15	0.13	0.09	0.15
	40	0.3	0.2	0.1	0.09	0.09	0.09	0.09
	50	0.1	0.2	0.2	0.13	0.15	0.09	0.09
	60	0.1	0.1	0.13	0.15	0.15	0.13	0.15
	70	0.1	0.2	0.09	0.21	0.16	0.15	0.16

(b) Error Standard Deviation

Figure 5: Distance Estimation Error

		angle (degree)						
		45	60	75	90	105	120	135
distance (cm)	30	3.20	2.10	1.60	0.9	1.80	3.1	3.1
	40	8.30	2.2	1.50	1.8	1.30	2.8	5.8
	50	5.10	2.9	1.90	2.1	1.90	4.1	3.1
	60	4.40	2.5	4.90	1.9	2.70	1.9	2.9
	70	5.80	2.70	2.90	2.3	3.3	1.7	4.1

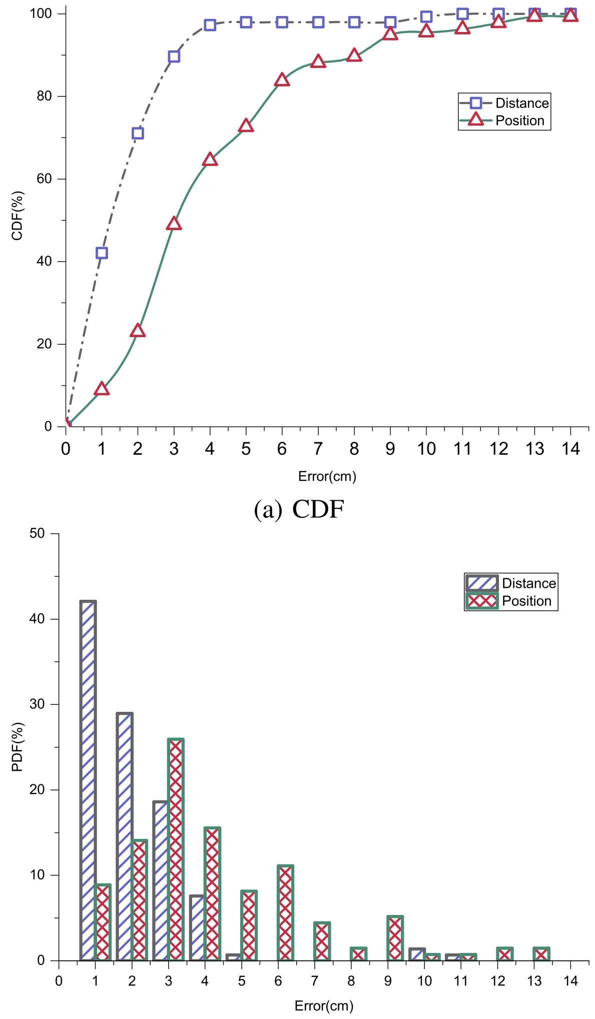
(a) Mean Error of Static Localization

		angle (degree)						
		45	60	75	90	105	120	135
distance (cm)	30	0.83	0.35	0.2	0.53	0.5	1.05	0.3
	40	4.45	0.42	0.32	1.3	0.7	1.06	2.1
	50	0.97	0.54	0.99	1.38	0.4	1.06	1.8
	60	0.67	0.7	1.81	0.80	0.4	1.08	1.3
	70	1.83	1.2	1.58	1.18	2.3	1.07	1.4

(b) Error Standard Deviation

Figure 6: Position Estimation Error

show the localization's mean errors and standard deviations. Note that the localization have larger mean errors compared to the ranging module. It is resulted from the accumulated ranging errors. The similar phenomenon happens toward the periphery where the mean error is larger than that in the center. Moreover, Fig. 5(b) shows that the ranging has higher stability, resulted from each position is accurate placed. But the position estimation is relatively unstable as shown in Fig. 6(b) because of the accumulated errors from distance estimation. We also 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.



(a) CDF

(b) PDF

Figure 7: CDF and PDF of distance and position estimation

Fig. 7 shows the CDF and PDF of the distance and position estimation. They indicate that the average accuracy of distance estimation within five centimeters of about 98% and three centimeters of about 89%. For the position estimation, the average accuracy is within five centimeters of about 73% and three centimeters of about 48%. Notice that these results are only for static localization. If we allow continuous measurements with a simple refinement from the smoothed trajectory, the accuracy can be further improved.

VI. Related Work

High-precision position can be achieved with accurate ranging. The ranging accuracy depends on the signal speed and the precision of TOA measurement. While the traditional ranging technology (TOA) needs synchronously sending time stamp (ST), BeepBeep [6] only needs to exchange the receiving time stamp (RT) communicated using Bluetooth, WI-FI or iBeacons. Furthermore, this approach only uses a microphone, a speaker and a communication component so that it is applicable to most of the smart phones. To identify the mobile phone user is whether the driver or the passenger, Yang *et. al* [20] assume that the phone position near the driver seat means the phone user is a driver, then the proposed acoustic relative-ranging system locates the phone respects to the driver's seat through the processing that recording some sound clips from the speakers of four corners and determining the Different Time Of Arriving (DTOA). Our scheme locates the hand through acoustics localization. Unlike the approaches above, this system does not require users to hold the device in hand. In addition, our scheme can be implemented on smart phone and COTS devices.

VII. Conclusion

Hand tracking systems are becoming increasingly popular as a key HCI approach. While moving hand trajectory can be estimated through smoothing the position coordinates collected from continuous hand localization, hand localization is still very challenging task. This paper demonstrates that it is possible to leverage acoustic signal to estimate the hand position. The proposed hand localization solution, EchoLoc, leverages the microphone embedded on smart phone and stereo speakers to locate the hand and can be enabled on COTS devices. We have evaluated EchoLoc and found it is capable of estimating the hand position resolution with an average error within five centimeters of 73% and three centimeters of 48%. We plan to further improve the accuracy by considering the continuous hand localization and achieve dynamic hand tracking.

Acknowledgment

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

References

[1] L. Sun, S. Sen, D. Koutsonikolas, and K.-H. Kim, "Widraw: Enabling hands-free drawing in the air on commodity wifi devices," in *Proc. of ACM Mobicom*, 2015.

[2] S. Yun, Y.-C. Chen, and L. Qiu, "Turning a mobile device into a mouse in the air," in *Proc. of ACM Mobisys*, 2015.

[3] B. Guo, Z. Wang, Z. Yu, Y. Wang, N. Yen, R. Huang, and X. Zhou, "Mobile crowd sensing and computing: the review of an emerging human-powered sensing paradigm," *ACM Computing Surveys*, 48(1), Article 7, August 2015.

[4] C. Bo, T. Jung, X. Mao, X.-Y. Li, and Y. Wang, "SmartLoc: Sensing landmarks silently for smartphone based metropolitan localization," *EURASIP Journal on Wireless Communications and Networking*, vol. 2016, April 2016.

[5] C. Bo, X. Jian, T. Jung, J. Han, X.-Y. Li, X. Mao, and Y. Wang, "Detecting driver's smartphone usage via non-intrusively sensing driving dynamics," *IEEE Internet of Things Journal*, to appear.

[6] C. Peng, G. Shen, Y. Zhang, Y. Li, and K. Tan, "Beepbeep: a high accuracy acoustic ranging system using COTS mobile devices," in *Proc. of ACM Sensys*, 2007.

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

[8] D. Graham, G. Simmons, D. T. Nguyen, and G. Zhou, "A software-based sonar ranging sensor for smart phones," *IEEE Internet of Things Journal*, vol. 2, no. 6, pp. 479–489, 2015.

[9] Z. Sun, A. Purohit, R. Bose, and P. Zhang, "Spartacus: spatially-aware interaction for mobile devices through energy-efficient audio sensing," in *Proc. of ACM MobiSys)*, 2013.

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

[11] P. Lazik and A. Rowe, "Indoor pseudo-ranging of mobile devices using ultrasonic chirps," in *Proc. of ACM Sensys*, 2012.

[12] K. Liu, X. Liu, and X. Li, "Guoguo: Enabling fine-grained indoor localization via smartphone," in *Proc. of ACM Mobisys*, 2013.

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

[14] 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.

[15] Z. Li, W. Chen, C. Li, M. Li, X.-Y. Li, and Y. Liu, "Flight: Clock calibration and context recognition using fluorescent lighting," *IEEE Transactions on Mobile Computing*, vol. 13, no. 7, pp. 1495–1508, 2014.

[16] Z. Li, W. Chen, M. Li, and J. Lei, "Incorporating energy heterogeneity into sensor network time synchronization," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 1, pp. 163–173, 2015.

[17] P. Lazik, N. Rajagopal, B. Sinopoli, and A. Rowe, "Ultrasonic time synchronization and ranging on smartphones," in *Proc. of IEEE RTAS*, 2015.

[18] M. Uddin and T. Nadeem, "Harmony: Content resolution for smart devices using acoustic channel," in *Proc. of IEEE INFOCOM*, 2015.

[19] S. Ganeriwal, R. Kumar, and M. B. Srivastava, "Timing-sync protocol for sensor networks," in *Proc. of ACM Sensys*, 2003.

[20] J. Yang, S. Sidhom, G. Chandrasekaran, T. Vu, H. Liu, N. Cecan, Y. Chen, M. Gruteser, and R. P. Martin, "Detecting driver phone use leveraging car speakers," in *Proc. of ACM Mobicom*, 2011.