2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)

HyperEar: Indoor Remote Object Finding with a Single Phone

Hongzi Zhu*, Yuxiao Zhang*, Zifan Liu*, Shan Chang† and Yingying Chen‡
*Shanghai Jiao Tong University, China
†Donghua University, China
‡Rutgers University, USA
hongzi@sjtu.edu.cn changshan@dhu.edu.cn yingche@scarletmail.rutgers.edu

***Abstract*—Finding a small object (e.g., keys or a wallet) in an indoor environment (e.g., in a house or an office) can be frustrating. In this paper, we propose an innovative system, called *HyperEar*, to localize such an object using only one single smartphone, based on enhanced time-difference-of-arrival (TDoA) measurements over acoustic signals issued from the object. One major challenge is the hardware limitations of a Commercial-Off-The-Shelf (COTS) phone with a short separation between the two microphones and the low sampling rate of such microphones. HyperEar enhances the accuracy of TDoA measurements by virtually increasing distances between microphones through sliding the phone in the air. HyperEar requires no communication for synchronization between the phone and the object and is a low-cost and easy-to-use system. We evaluate the performance of HyperEar via extensive experiments in various indoor conditions and the results demonstrate that, for an object of 7m away, HyperEar can achieve a mean localization accuracy of about 15cm when the object in normal indoor environments.**

***Index Terms*—object finding; time difference of arrival; indoor environment; smartphone; acoustic source localization**

I. Introduction

It is often the case that people find it is very frustrating to find their personal objects, such as keys and wallets, in an indoor environment. With the proliferation of mobile devices (e.g., smartphones, smartwatches, and tablets), it would be appealing to use one such mobile device to localize small personal objects. To this end, a tag issuing inaudible acoustic signals can be attached to a private object and can be localized by the smartphone of the object owner.

Such an application pose four rigid requirements to a system as follows: *1) far operational distance:* as a tag may be far from the mobile device of a user, especially for a large indoor environment, the system should work well at a long distance. *2) good localization accuracy:* object finding calls for dm- or even cm-level localization accuracy. *3) excellent user experience:* the system should be easy to use and minimize the involvement of users; otherwise, users would get bored and quit using the system. Last but not least, *4) low deployment cost:* the system should rely on devices that are cheap or users already have.

In the literature, existing techniques are insufficient to meet all above requirements. A number of pioneering sound source positioning systems [1]–[6] heavily rely on wireless communication for synchronization. For example, in Cricket [1], ultrasound is used to localize the speaker via Time-of-Flight (ToF) and the speaker and microphones are synchronized through radio frequency (RF) signals. BeepBeep [4] also needs wireless communication for phone-to-phone ranging. Another large portion of existing work [7]–[12] is based on dedicated hardware such as microphone arrays deployed in specific locations in advance. For example, an 8-microphone-array is used on a mobile robot to localize a sound source in 3D [7]. Stefanakis et al [12] propose a system to use a 4-microphone-array to estimate the angle-of-arrival (AoA) of multiple sound source. Several recent studies [13], [14] use smart devices to localize a sound source at mm-level accuracy. For example, keystrokes can be snooped with a smartphone placed by a keyboard [13]. vTrack [14] uses two or three microphones of a smartphone to localize and track a speaker near the phone and can achieve an accuracy of 2.3mm on 0.26m×0.2m regions. These systems can achieve very accurate localization but only work in a short range. WalkieLokie [15] is one recent work that also considers to localize a remote sound source with one single smart device. However, it needs users to continuously walk. As a result, there is no existing solution, to the best of our knowledge, to localizing an small acoustic object with one single smartphone.

In this paper, we propose an indoor acoustic source localization system, called *HyperEar*, which uses a Commercial-Off-The-Shelf (COTS) smartphone to localize a remote speaker in an indoor environment. HyperEar can be deployed as a software (e.g., an app) without any hardware modification. The basic idea of HyperEar is get highly accurate Time Difference of Arrival (TDoA) measures of inaudible acoustic beacons with a smartphone. More specifically, to find the target speaker, a user, holding his/her phone, first selects an appropriate direction and then slides the phone in the air along the direction back and forth for several times. For each onboard microphone, one TDoA can be obtained before and after one slide. With at least two microphones and two corresponding TDoA measures, triangulation can be conducted to estimate the relative location of the phone with respect to the speaker.

Two main challenges lie in the HyperEar scheme. First, with the short distance between both microphones and their limited sampling rate, how to obtain accurate TDoA measurements is not straightforward. For instance, given the small size of a phone, two onboard microphones are separated at most at a distance of about 13-15cm. If the sampling frequency

2575-8411/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDCS.2019.00073

is 44.1KHz as adopted by most smartphones and the sound speed is 343m/s, there would be about only 40 distinguishable TDoA measurements, dividing the space into 40 regions. Consequently, locations in the same TDoA region cannot be discriminated. As such regions expand quickly as the distance from the phone increases, it leads to huge location ambiguity when the speaker is far away from the phone.

To deal with this challenge, HyperEar innovatively incorporates two techniques to increase the density of TDoA regions. First, leveraging the uneven distribution of TDoA regions, the phone is rolled around its z-axis to find the optimal direction, where the speaker stays in the area with densest regions (i.e., with least location ambiguity). Second, the phone is slid along the identified direction in the air and TDoAs are calculated based on the estimated sliding distance of each microphone. In this way, the number of measurable TDoAs is no longer restricted by the size of the phone but the sliding distance of the phone. Therefore, by increasing the sliding distance, the density of TDoA regions can also be increased.

The second challenge is how to accurately estimate the sliding distance of phone with low-end onboard sensors. With unstable hand operations, it is unlikely to achieve perfect movements of the phone. Furthermore, the ever-changing posture of the phone and error-prone acceleration readings result in unexpected rotation and displacement estimation errors.

To tackle this challenge, we leverage an appealing feature of HyperEar, which allows slight phone rotation and displacement from ideal slides. The reason is that the TDoA measurement error, due to unstable sliding movements, occurring on one microphone is about the same amount as that occurring on the other microphone, and they are cancelled with each other during the triangulation calculation. To deal with low-quality acceleration readings, HyperEar leverages the fact that the true velocity of the phone at the starting and at the ending of a sliding movement should be zero, and adopts a linear model to remove the accumulative errors caused by taking the integral of acceleration readings over time.

HyperEar can be used for 2D localization and can be easily extended for 3D localization as well. It has many advantages as follows. First, it reaches the minimum deployment cost with a cheap speaker and the user's smartphone. Second, it requires no synchronization or communication between the speaker and the phone. Last but not least, it is easy to use and has very light workload and elastic requirements for users. We have implemented HyperEar on two types of smartphones, and evaluated the performance of HyperEar through extensive experiments in two indoor environments with different noise types and levels. The results show that, for a speaker of 7m away, HyperEar can achieve a mean localization accuracy of about 15cm in a normal indoor environment and 37cm in a noisy shopping mall.

We highlight our main contributions made in this paper as follows:

- We propose a novel scheme, called HyperEar, to passively estimate highly accurate TDoA of acoustic beacons issued from a distant speaker using a smartphone.
- We implement HyperEar on two Android-based smartphones, i.e., i.e., Samsung Galaxy S4 and Samsung Galaxy Note3, which demonstrates the feasibility of the proposed scheme.
- We conduct a systematic evaluation that shows the high accuracy of HyperEar. The results demonstrate the efficacy of the HyperEar design.

The remainder of this paper is organized as follows. We present the system model, basic concepts of TDoA-based localization and the restrictions of conducting acoustic localization on a smartphone in Section II. Section III introduces the architecture of HyperEar. Finding the direction between a user and the target speaker is elaborated in Section IV. Section V describes how to estimate the phone displacement when the user sliding the phone in the air, given the noisy inertial sensor readings. The procedures of acoustic source localization based on triangulation is introduced in Section VI. We discuss the practical issues that may be encountered in Section **??**. Section VII presents the performance evaluation and experiment results. Section VIII compares HyperEar with related work. Finally, we present concluding remarks of our work and summarize the directions for future work in Section IX.

II. Restrictions of Acoustic Localization on Smartphones

In this section, we introduce the system model and analyze key factors that affect the localization accuracy.

A. System Model

To be practical, we consider a system with only two components:

- **One Single Speaker:** a speaker can be attached to a target personal object and periodically plays an acoustic signal. This signal can be audible or inaudible to human ears, depending on the specific application scenarios.
- **One Single Smartphone:** HyperEar exploits two microphones and the inertial sensors, i.e., the accelerometer and the gyroscope, embedded in a smartphone. Note that it does not need any synchronization or data communication between the phone and the speaker.

B. Basic Concepts of TDoA-based Localization

As we do not know when a particular sound signal is emitted from the speaker, the TDoA of this sound signal can be measured with two microphones. In essence, a TDoA measurement reveals geometry information about the direction of an incoming sound. Consequently, establishing the exact position of this sound in 2D normally requires triangulation, e.g., at least two distinct direction measurements. For example as illustrated in Figure 1, let d_1 and d_2 denote the distance between the speaker and two microphones Mic1 and Mic2, respectively. Suppose Δt_1 is the TDoA measured over Mic1 and Mic2, and then the distance difference Δd_1 from the speaker to this pair of microphones can be represented as

$$\Delta d_1 = d_1 - d_2 = \Delta t_1 \times S \tag{1}$$

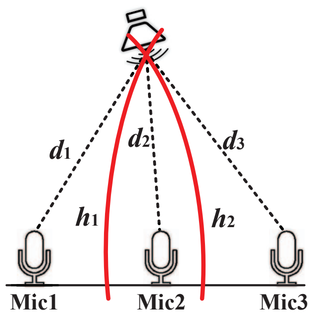


Fig. 1. Localization with triple microphones

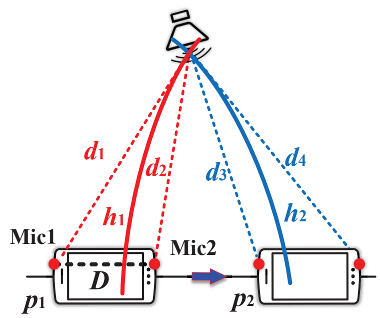


Fig. 2. Naive localization scheme with a two-microphone phone

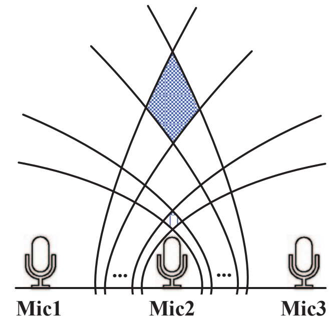


Fig. 3. Location ambiguity increases for far objects

where S is the velocity of sound. All possible positions satisfying Δd_1 lie on the half hyperbola h_1 as illustrated by the left red curve in Figure 1. Similarly, with Mic2 and a third microphone Mic3, another half hyperbola h_2, illustrated by the right red curve, can be estimated and used to derive the relative location of the speaker (i.e., the intersection of h_1 and h_2) with respect to these microphones.

As most COTS smartphones are associated with only two microphones, one naive solution, as illustrated in Figure 2, can be used to localize a phone if the location of the speaker is known. First, a half hyperbola h_1 can be calculated at a position p_1 with the onboard microphones Mic1 and Mic2. Then, the phone can be moved to another position p_2 and a second half hyperbola h_2 can be obtained. As a result, the relative location between the speaker and the phone can also be determined if the moving direction as well as the distance from p_1 to p_2 are known.

C. Challenges for Locating Remote Objects

Several hardware limitations of a COTS smartphone make the above naive solution very challenging when the speaker is far from the phone.

Limited Sampling Rate. When recording, the sound is digitized by the Analog to Digital Converter (ADC) of a microphone at a fixed sampling rate f_s. Therefore, the resolution of TDoA measurements is restricted by the f_s. Though current state-of-the-art audio hardware on smartphones supports sampling rate of up to 192kHz, operating system usually limits this to 44.1kHz, which means the resolution of TDoA measurements is about 0.023ms or the resolution of distance difference Δd (i.e., the distance interval between two adjacent hyperbolas) is about 7.78mm at $S = 343$m/s.

Near Separation between Microphones. The number of distinguishable hyperbolas also depends on the range of possible distance difference Δd, which is bounded by the distance between the two microphones on the phone. As can be inferred from Figure 2, the range of Δd is $[-D, D]$, where D is the distance between the two microphones. Given a sampling rate f_s and D, the number of distinguishable hyperbolas N can be calculated as

$$N = \lfloor 2Df_s/S \rfloor. \tag{2}$$

For example, the distance between the two microphones of a Samsung Galaxy S4 is 13.66cm. With a sampling rate of 44.1kHz, this yields only 35 measurable hyperbolas. As illustrated in Figure 3, with the limited number of distinguishable hyperbolas, the density of hyperbolas drops dramatically as the distance from the microphones increases. While adopting naive TDoA localization schemes can achieve cm- or mm-level accuracy for very near objects, it is very challenging to accurately localize a far sound source. For instance, the localization error of the above naive scheme can reach up to 18.6cm and 266.7cm when the sound source is located at 1m and 5m away from a Samsung Galaxy S4 smartphone, respectively.

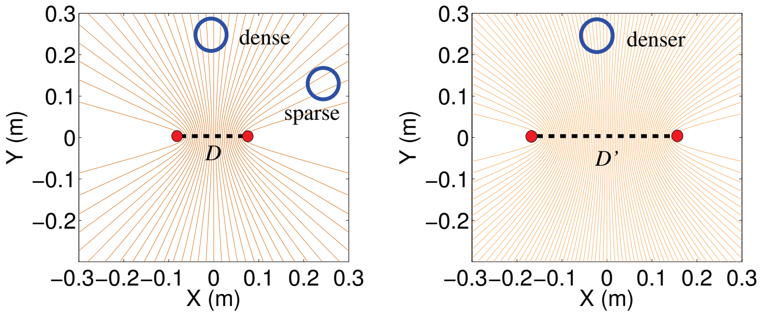
Low-end Inertial Sensors. In the naive scheme, onboard inertial sensors can be used to estimate the information of the moving direction and distance of the phone from p_1 to p_2. Deriving accurate motion information with sampled and error-prone sensor readings, especially under the condition that phone movements are carried out by untrained users, is very difficult.

D. Key Observations on TDoA Measurements

With regard to measuring TDoAs on smartphones, we have two main observations. First, as shown in Figure 4(a), it is clear to find that the distribution of hyperbolas over space is quite uneven, with the central areas having a denser distribution of hyperbolas than other sideward areas. Second, if we increase the separation between two microphones from D to D', as shown in Figure 4(b), the number of hyperbolas will also increase according to (2), leading to a higher density of hyperbolas at remote locations.

III. Overview of HyperEar

We propose an innovative scheme, called *HyperEar*, to solve the problem of indoor smartphone localization with a single remote speaker. The core idea is to expand the TDoA measurement range by moving the phone in the air so that the number of distinguishable hyperbolas is increased, reducing the location ambiguity at a far distance from the phone. Moreover, the motion of the phone is tracked and estimated by processing the noisy inertial sensor readings. As depicted



(a) Uneven distribution of hyperbolas (b) Expanding separation between mics

Fig. 4. Two key observations about TDoA measurements

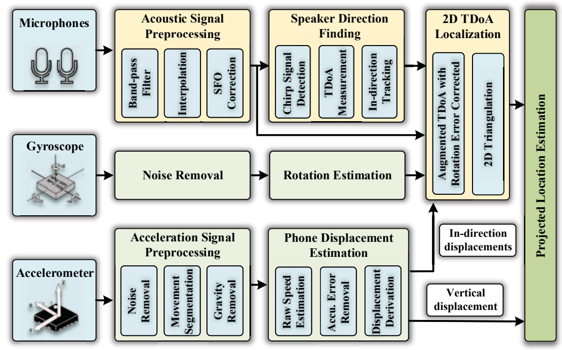


Fig. 5. System architecture of HyperEar

in Figure 5 , the structure of HyperEar incorporates six main components.

Acoustic Signal Preprocessing (ASP). ASP performs three functions to improve the accuracy of TDoA measurements. First, ambient sound being out of the frequency band of the sound signal is filtered out. Second, interpolation is carried out to achieve sub-sample resolution. Third, sampling frequency offset (SFO) errors between the speaker and each of the two microphones are corrected.

Speaker Direction Finding (SDF). The key function of SDF is to find the direction of the speaker. It needs to detect sound signals at both microphones and measure the TDoA for each signal. Based on the relationship of previous TDoA measurements, it gives instructions to help a user find the direction of the speaker.

Motion Signal Preprocessing (MSP). The motion information of the phone is required in HyperEar. This component preprocesses the accelerometer and gyroscope readings. First, it removes high-frequency noise from both signals. It then segments the movement of the phone based on the power level of the acceleration signal.

Phone Displacement Estimation. (PDE) In HyperEar, the phone is required to move in both horizontal and vertical directions. Therefore, the displacements of the phone in each direction are estimated. This component takes the segmented acceleration signals as input to estimate the moving speed and distance of the phone along some direction.

2D TDoA Localization (TTL). The key function of this component is to estimate the distance between the speaker and the phone. Instead of measuring a TDoA based on two microphones at the same position, it measures a TDoA based on two positions at the same microphone. It then integrates the information of the TDoA measurements, the motion estimation, and the direction of the speaker to perform 2D triangulation.

Projected Location Estimation (PLE). PLE tackles indoor smartphone localization in 3D scenarios. Instead of estimate the relative location of the speaker in 3D space, HyperEar calculates the projected location of the speaker on the floor map. In this way, it does not need to know the height information of the speaker and that of the phone.

IV. Speaker Direction Finding

The purpose of finding the direction of the speaker is two-fold. First, if the direction of the speaker is known, we can roll the phone along its z-axis so as to make the speaker sit in the dense area of hyperbolas as depicted in Figure 4(a). Second, this direction information is required in the triangulation calculation. The SDF can find the direction of the speaker based on the fact that, when the speaker aligned with the x-axis of the phone, the TDoA measured on the phone should be zero.

A. Signal Detection and TDoA Measurement

In HyperEar, the speaker periodically plays a chirp signal, in which the frequency first linearly increases and then decreases with time, for its good auto correlation property. We adopt the method introduced in BeepBeep [4] to detect signals at each microphone. To detect the signal, the recorded audio signal at each microphone is correlated with a reference chirp signal. The maximum peak of correlation is concluded as the location of a signal if the value is significantly larger than that with background noise. A TDoA of the ith signal is measured as $t^i_{Mic1} - t^i_{Mic2}$, where t^i_{Mic1} and t^i_{Mic2} represent the timestamps of the ith signal detected at Mic1 and at Mic2, respectively.

B. In-direction Position Tracking

As depicted in Figure 6, when the phone is rolled along its z-axis, the measured TDoA varies in the range of $[-D/S, D/S]$. Let $\alpha \in [0°, 360°)$ denote the angle between the direction of the speaker and the positive direction of y-axis of the phone. The speaker is considered on the right-side of the phone when $\alpha \in [0°, 180°)$ and on the left-side when $\alpha \in [180°, 360°)$. When $t^i_{Mic1} - t^i_{Mic2} = 0$, the direction of the speaker is found, i.e., $\alpha = 90°$ means that the speaker locates in the positive direction of x-axis and $\alpha = 270°$ means that the speaker locates in the negative direction of x-axis. In this case, the phone is at a so-called *in-direction* position and

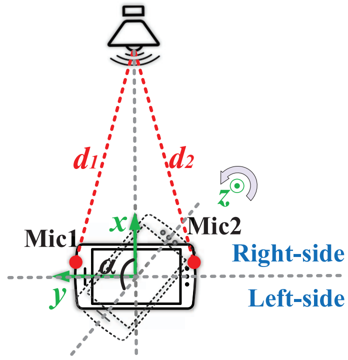


Fig. 6. Illustration of phone rotation and an in-direction position

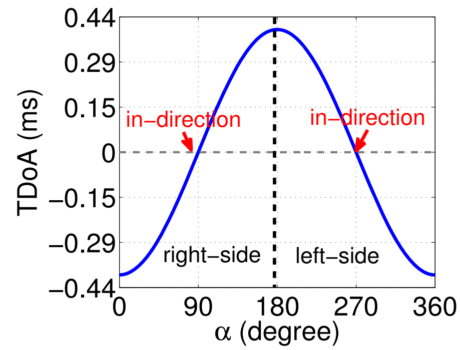


Fig. 7. Relationship between α and the distance difference $d_1 - d_2$

we stop rolling the phone. Figure 7 depicts the relationship between measured TDoAs and α obtained with a Galaxy S4.

V. Phone Displacement Estimation

In HyperEar, the TDoA measurement range is expanded by incorporating phone movements. Therefore, accurate motion information of the phone is key to both TDoA measurement and the final triangulation calculation. The PDE can obtain accurate motion estimation through comprehensive signal processing on raw inertial sensor readings.

A. Motion Signal Preprocessing

1) Noise Removal: We first use gravimeter to cancel the gravity to get linear acceleration data. With low-end inertial sensors, the acceleration and angular speed signals along each axis of the phone contain high-frequency noise. We remove such high frequency noise by passing each signal through a low pass filter. In this work, we use a moving average (SMA) filter, which is the unweighted mean of the previous n samples. We empirically choose the value of n to be 4 to achieve -3dB cut-off frequency at 15Hz with the sampling rate of the accelerometer and gyroscope being 100Hz.

2) Movement Segmentation: In HyperEar, phone movements only consist of simple sliding operations along a given direction. In order to determine the starting and ending points of each slide, we first examine the power levels of the acceleration signals. Particularly, we calculate the power levels of the acceleration signal along y-axis by averaging the accumulative square of the signal amplitude in a sliding time window as shown below

$$P(t) = \frac{1}{W} \sum_{n=t}^{t+W} a(n), \tag{3}$$

where W is the length of the time window and $a(\cdot)$ is the amplitude of the acceleration signal. We empirically take the length of the sliding window as 4 samples (i.e., 40ms with the sampling rate of 100Hz). Figure 8 illustrates the power levels of y-axis acceleration signal when the phone is slid back and forth along its y-axis. We consider that a slide starts when the power levels exceeds a threshold and stops when the power levels goes below the threshold for m samples. An empirical threshold of 0.2 and $m = 8$ are used in this work.

B. Sliding Velocity and Displacement Estimation

An intuitive way to estimate the moving speed of the phone along one axis is to calculate the integral of linear acceleration along that axis over time. For example, Figure 9(a) illustrates the y-axis linear acceleration of one slide as a function of time. The integral speed is depicted as the dashed curve in Figure 9(b). It can be seen that at the end of the slide, the integral speed drifts apart from the ground truth (i.e., zero at the ending point of a slide).

As studied in our prior work [16], the above accumulative error of integral is approximately a linear function of time. Given the fact that the true velocity at both ends of a slide is

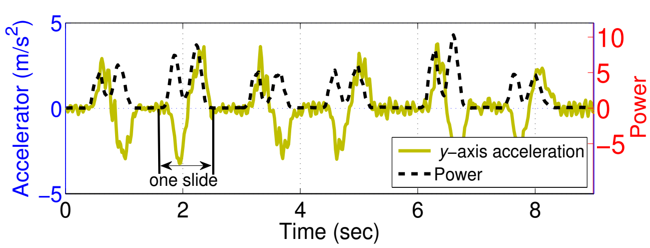
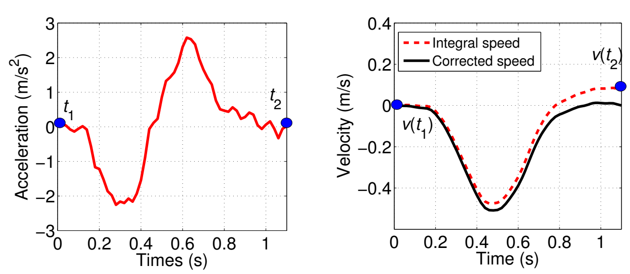


Fig. 8. Segmenting movements based on power of acceleration



(a) Acceleration of a typical slide movement

(b) Linear accumulative errors of integral speeds

Fig. 9. Illustration of speed estimation with accumulative errors removed

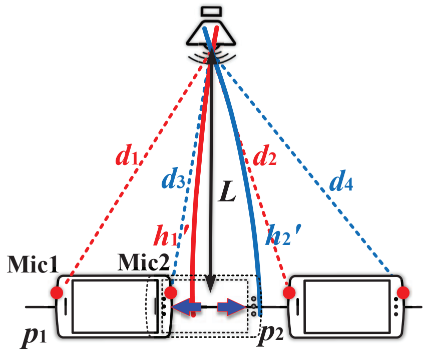


Fig. 10. Augmented TDoA Measurements

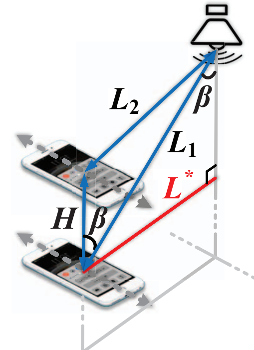


Fig. 11. Projected location on map

zero, the linear model of errors can be derived and utilized to infer accurate moving speed. In specific, let t_1 and t_2 respectively denote the starting and ending points of a slide, and $v(t_1)$ and $v(t_2)$ respectively denote the integral velocity values at t_1 and t_2, as illustrated in Figure 9. For linear accumulative errors, the slope of the linear model can be estimated as

$$err_a = \frac{v(t_2)}{t_2 - t_1}. \quad (4)$$

The instant velocity between t_1 and t_2, therefore, can be corrected as $v^*(t) = v(t) - err_a \times (t - t_1)$, where $v(t)$ is the integral velocity at time t. For example, the solid curve as shown in Figure 9(b) illustrates the corrected speed.

Given the corrected sliding velocity $v^*(t)$, the displacement between any two time instants during a slide can be derived by taking the integral of $v^*(t)$ over time.

VI. Localization through Triangulation

A. 2D Localization based on Augmented TDoAs

Assume that the speaker and the phone co-locate in the same horizontal plane. Without losing generality, we take the case as depicted in Figure 10, where the speaker locates on the right side of the phone. Suppose that the phone hears a signal at position p_1 with the corresponding timestamps at Mic1 and Mic2 being t_1 and t_3, respectively, and hears the next nth signal at position p_2 with the corresponding timestamps at Mic1 and Mic2 being t_2 and t_4, respectively. The TDoA of Mic1 at p_1 and p_2 can be measured as $\Delta t_1' = t_2 - t_1 - n \times T$, where T is the period of signals and n is the number of detected chirp signals according to the algorithm described in Subsection IV-A. Similarly, the TDoA of Mic2 at p_1 and p_2 can be measured as $\Delta t_2' = t_4 - t_3 - n \times T$.

If Cartesian coordinates are introduced such that the origin is the center of the two positions of Mic1 and the x-axis is the inverse y-axis of the phone, then the half hyperbola h_1' can be represented as

$$\sqrt{(x - \frac{D'}{2})^2 + y^2} - \sqrt{(x + \frac{D'}{2})^2 + y^2} = \Delta t_1' \times S, \quad (5)$$

where D' is the estimated sliding distance between p_1 and p_2, and (x, y) is the coordinates of the speaker. Accordingly, the half hyperbola h_2' can be represented as

$$\sqrt{(x - D - \frac{D'}{2})^2 + y^2} - \sqrt{(x - D + \frac{D'}{2})^2 + y^2} = \Delta t_2' \times S, \quad (6)$$

where D is the distance between Mic1 and Mic2 on the phone.

The intersection of h_1' and h_2', i.e., the solution of (5) and (6), is the relative location of the speaker in this coordinate system. In particular, we are interested in the distance L as illustrated in Figure 10, which is the y coordinate of the solution.

B. Projected Location Estimation in 3D

In practice, it is hard to know the stature relationship between a target speaker and the phone. Fortunately, for indoor localization applications, the location of the smartphone on a floor map is concerned. With this condition, the 2D localization based on augmented TDoAs can be extended to more general cases where the speaker and the phone have different statures and do not share a common horizontal plane.

In specific, the phone is required to slide on two horizontal planes with different statures as illustrated in Figure 11. The scheme for phone displacement estimation can also be used to estimate the stature change between the two horizontal planes (e.g., the H as depicted in Figure 11) by conducting the same signal processing procedure on z-axis acceleration readings. Given the estimation of L_1, L_2, and H, the angle β can be calculated as

$$\beta = arccos\frac{H^2 + L_1^2 - L_2^2}{2 \cdot H \cdot L_1}. \quad (7)$$

Therefore, the projected distance L^* can be calculated as $L_1 \times sin(\beta)$.

VII. Performance Evaluation

A. Methodology

We have implemented HyperEar as an application on two models of smartphones, i.e., Samsung Galaxy S4 and Samsung Galaxy Note3. Both phones run Android 5.0 with two

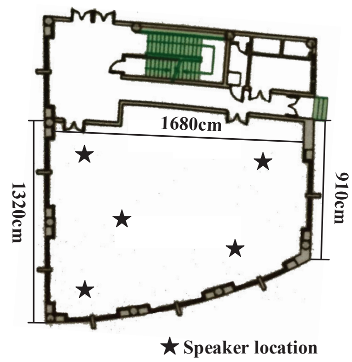


Fig. 12. The meeting room and selected speaker locations for localization experiments

Fig. 13. One of the volunteers sliding an experimental phone in the meeting room

separated microphones that support 16-bit 44.1kHz sampling rate and stereo recording. The distance between the two microphones is 13.66cm and 15.12cm for Samsung Galaxy S4 and Samsung Galaxy Note3, respectively. A cheap desktop speaker with 2W root-mean-square (RMS) power and 150Hz-20kHz frequency response is used. The speaker is mounted on a tripod for both 2D and 3D localization, and connected to a laptop which keeps playing chirp signals on every 200ms.

We evaluate HyperEar in two indoor environments:

- *A meeting room.* The room is approximately 17m×13m with a front stage and ten raws of seats arranged in theatre fashion, as shown in Figure 12. We randomly select five positions in the room to set the speaker. For each position, we let 10 volunteers, four females and six males with stature ranging from 160cm to 187cm, to operate the experimental smartphones according to the specific requirements of each experiment. An example operation of HyperEar is illustrated in Figure 13. We also control the noise level of the room by asking volunteers to keep quite or to chat.
- *A shopping mall.* To set the speaker, we randomly select five positions in the corridor of a shopping mall, which is 95m×16.5m with shops open on both sides. We conduct experiments during off-peak hours when there is background soft music and busy hours when the place is crowded and with advertisement broadcasting.

We evaluate the HyperEar system using the metric of *accuracy* defined as the projected Euclidean distance from the estimated location and the ground truth location of the speaker on the floor map.

B. Effectiveness of Sliding

In this experiment, we first examine the effectiveness of sliding the phone in improving the accuracy of TDoA measurements and 2D localization. For each speaker position in the meeting room, we randomly select five testing positions that are at a distance of 5m away from the speaker. To eliminate the impact of unstable hand operations, in this experiment, we mount each phone on a level slide ruler when sliding. In particular, for each selected position, the ruler is set so that it has the same stature as the speaker and the phone is in-direction when it moves to the center of the ruler. We then slide the phone on the ruler for 50 times with different distances, ranging from 10cm to 60cm with an interval of about 10cm. For each slide, the sliding distance is estimated with PDE (see Section V) and the distance from the speaker to the ruler is estimated with 2D Localization (see Subsection VI-A).

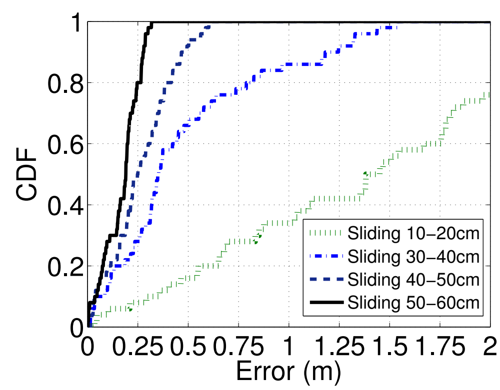


Fig. 14. CDF of 2D localization errors under different sliding distances, Note3 on slide ruler

Figure 14 depicts the cumulative distribution function (CDF) of 2D localization errors over all slides on the Note3. The results on the S4 are similar. It is obvious to see that increasing the sliding range will greatly reduce the 2D localization errors. For example, the average localization error is 18cm when sliding range is 50-60cm comparing to the value of 142cm when the range is 10-20cm. In practice, while it is ideal to slide the phone for longer distances, it is hard for a user to stably control the operation as the sliding distance increases. In HyperEar, slides with an estimated distance over 50cm and z-axis rotation angle less than 20° are automatically selected for use.

C. Impact of Speaker Distance

We then study the effective operation distance of HyperEar. We take a similar setting as in the above experiment except

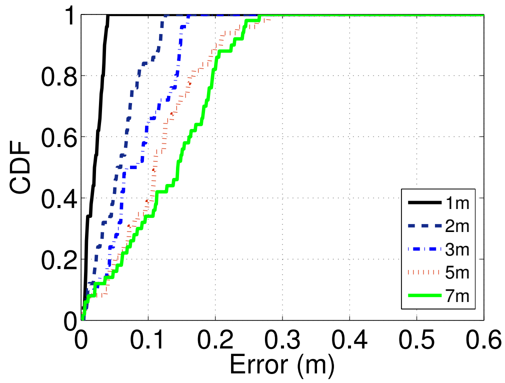


Fig. 15. CDF of 2D localization errors under different operational ranges, S4 on slide ruler

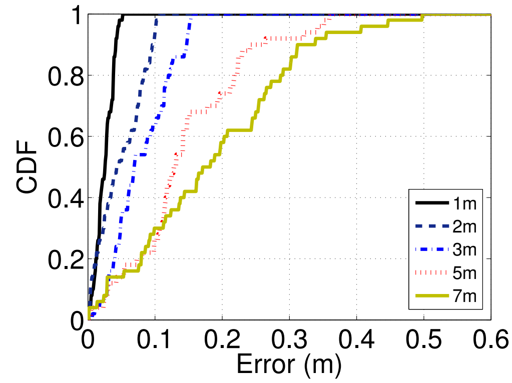


Fig. 16. CDF of 2D localization errors under different operational ranges, Note3 on slide ruler

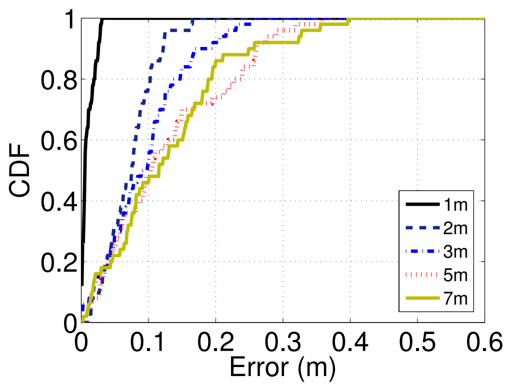


Fig. 17. CDF of 3D localization errors with 5-slide aggregation, S4 in hand

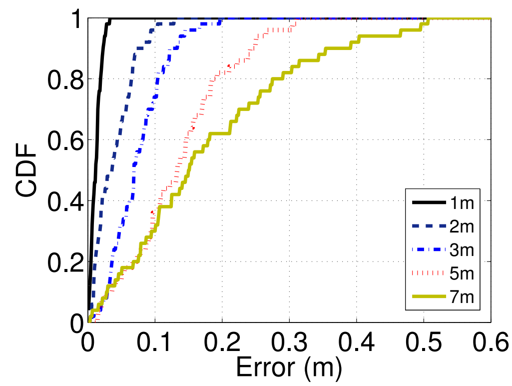


Fig. 18. CDF of 3D localization errors with 5-slide aggregation, Note3 in hand

that each time we slide the phone in a range of 50-60cm on the slide ruler and change the distance between the speaker and the slide ruler from 1m to 7m at an interval of 1m.

Figure 15 and Figure 16 plot the CDFs of 2D localization errors for each distance with the S4 and the Note3, respectively. It can be seen from both figures that: 1) as the speaker distance increases, the 2D localization accuracy gradually decreases; 2) HyperEar achieves better performance on the S4 than on the Note3. For example, the mean and 90%-precision accuracy for 1m distance are 2.0cm and 3.5cm, respectively, and the corresponding values are 14.4cm and 22.3cm when the distance is 7m.

D. 3D Localization

We then examine the performance of the full-version HyperEar smartphone localization system in 3D scenarios. We change the stature of the speaker to 0.5m and randomly select 5 speaker positions in the meeting room. For each speaker position, we change the distance of randomly-selected testing positions from 1m to 7m. For each position of the speaker and each testing position, we ask each volunteer to use each experimental phone to first find the direction of the speaker, slide the phone at one customized stature for five times, change to another customized stature, and slide the phone for five times again.

Figure 17 and Figure 18 plot the CDFs of 3D localization errors for different speaker distances on the S4 and the Note3, respectively. It can be seen that HyperEar can achieve accurate localization in 3D scenario. For example, over a distance of 7m, the mean and 90%-precision localization accuracy on the S4 is 15.8cm and 25.2cm, respectively, and the corresponding values on the Note3 are 19.4cm and 37.5cm, respectively.

E. Different Indoor Environments

We consider the impact of different indoor environments to the performance of HyperEar. In this experiment, we mount the speaker on the tripod for the ease of deployment at five randomly selected positions in both environments as described in the methodology. For each speaker position, five testing positions that are 7m away from the speaker are selected. For each speaker position and each testing position, we ask each volunteer to perform 3D localization using HyperEar with both experimental phones. We conduct the experiment with

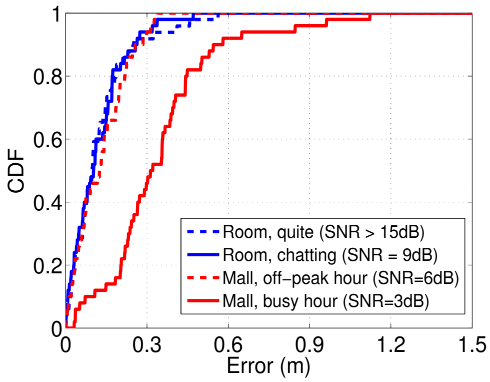


Fig. 19. CDFs of 3D localization errors in two indoor environments with different noise types and levels

different types of noise and control the volume of the speaker so that different signal-to-noise ratio (SNR) values are studied.

Figure 19 plots the CDFs of 3D localization errors on the S4. It can be seen that HyperEar performs stably in the meeting room where the background noise is mainly voice. Recall that we choose a 2-6.4kHz linear chirp signal while human voice is normally lower than 2kHz, which will be filtered out and has little effect to the localization performance. We observe similar results for the shopping mall environment in off-peak hours when there is background music. Though the frequency band of the background noise in the shopping mall overlaps with that of our chirp signal, HyperEar can achieve good performance when SNR is higher than 6dB. When in busy hours, the background noise level dramatically changes over time, making the average SNR low and affecting the localization performance. In the worst case, the mean accuracy is 37.2cm at a distance of 7m.

VIII. Related Work

We classify existing sound source ranging and localization systems into two categories based on system complexity. We further divide each category into two subcategories, i.e., communication-based and communication-free.

A. Dedicated Hardware based

1) Communication-based: In Cricket [1], a listener uses ultrasound to measure the ToF of a beacon with the help of RF signals and calculates its location with three ToF measurements. The system has been reported to perform 100% accurately on 1.2m×1.2m regions. Wang et al [8] implemented a robot navigation system with a distributed array of 24 microphones. As the robot speaks, TDoA of each pair of microphones is measured and used to calculate the robot's localization. The accuracy can reach about 7cm within about 3m range. Diva et al [6] realize a 2D localization system with multiple microphones. The system combines GPS information and wireless sensor network and can achieve an localization accuracy of 10cm within a 11m×11m area.

2) Communication-free: Valin et al [7] use an array of no less than four microphones to find the direction of an acoustic source. The system can reach about 1.7° error when the source is 3-5m away to the array of 8 microphone. Zhang et al [9] present a unified maximum likelihood framework of sound source localization and beamforming. The system uses a microphone array and can achieve AoA accuracy of 6° on a 7m×6m×2.5m region. Stefanakis et al [12] use a 4-microphone-array to estimate the AoA of multiple sound sources. The system uses perpendicular cross-spectra algorithm to derive AoA of signal and count the number of sources. The AoA accuracy is 2° within 1.3m range and the counting accuracy is 94.3% for 3 sources.

B. COTS Mobile Device based

1) Communication-based: BeepBeep [4] is a high accuracy ranging system between two COTS devices. The basic idea of BeepBeep is for each phone to emit a chirp signal, capture two signals (i.e., one from itself and one from another phone), and calculate the relative distance with information exchanged through wireless communication. This method can achieve an accuracy of 5cm within 10m. Qiu et al [5] propose a method based on BeepBeep to realize 3D localization between two smartphones. It can reach an accuracy of tens of centimeters within 5m.

2) Communication-free: Liu et al [13] use a smartphone to snoop keystrokes, reaching mm-level audio ranging. Key strokes are grouped based on TDoA. Acoustic features of key strokes are further used to differentiate keys. vTrack [14] uses two or three microphones of a smartphone and combines TDoA, AoA and power level information to localize and track a speaker near the phone. After that, it uses doppler-effect to track the movement of the speaker. vTrack can achieve an accuracy of 2.3mm on a 0.26m×0.2m region. Although these approaches can achieve extremely high accuracy with single device, they can only work in a very short range. Shake and Walk [20] uses one single microphone to find the direction of a speaker. The basic idea is to detect the frequency change caused by the doppler effect when user moves the phone. It can achieve less than 3° error in 32m distance. Walkielokie [15] is the most related work with HyperEar. It also uses one single smart device to localize a remote sound source. The system requires a user to walk and uses the doppler effect to calculate relative distances of the speaker. Walkielokie can achieve sub-meter accuracy within a range of tens of meters. However, this approach needs the user to continuously walk.

IX. Conclusion and Future Work

In this paper, we have proposed HyperEar, an indoor object finding system based on a single smartphoen. HyperEar overcomes hardware limitations posed by a phone and can achieve 15cm accuracy on average for a desktop speaker of 7m away in normal indoor environments. HyperEar minimizes the system deployment cost by relying on cheap or existing devices, which paves the way for wide application of HyperEar.

The current implementation of Hyper has three main limitations, which direct the way of our future work. First, the system adopts a linear chirp sound signal that is audible to the human ear. While this may be helpful in the application of object finding, constantly broadcasting such sounds in public places will be annoying. In the future, we will examine to use inaudible sound signals and investigate the impact of signal distortion due to frequency selectivity of smartphone microphones. Second, the system assumes a speaker and the phone to be in LoS condition. In the future, we will utilize the mobility of the user (e.g., moving to a nearby position by walk). Last, HyperEar needs a user to perform sliding operations as stable as possible. We will study more reliable schemes, such as stereo vision techniques, to tracking the motion of the phone.

ACKNOWLEDGEMENTS

This work was partially supported by National Natural Science Foundation of China (Grant No. 61772340, 61672151, 61420106010), Shanghai Rising-Star Program (Grant No.17QA1400100), DHU Distinguished Young Professor Program, the National Key R&D Program of China (2018YFC1900703) and 2017 CCF-IFAA Research Fund.

REFERENCES

[1] N. B. Priyantha, A. Chakraborty, and H. Balakrishnan, "The cricket location-support system," in *Proceedings of ACM MobiCom*, 2000.

[2] A. Harter, A. Hopper, P. Steggles, A. Ward, and P. Webster, "The anatomy of a context-aware application," *Wireless Networks*, vol. 8, no. 2-3, pp. 187–197, 2002.

[3] C. V. Lopes, A. Haghighat, A. Mandal, T. Givargis, and P. Baldi, "Localization of off-the-shelf mobile devices using audible sound: architectures, protocols and performance assessment," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 10, no. 2, pp. 38–50, 2006.

[4] C. Peng, G. Shen, Y. Zhang, Y. Li, and K. Tan, "Beepbeep: a high accuracy acoustic ranging system using cots mobile devices," in *Proceedings of the ACM SenSys*, 2007.

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

[6] X. Zheng, S. Yang, N. Jin, L. Wang, M. L. Wymore, and D. Qiao, "Diva: Distributed voronoi-based acoustic source localization with wireless sensor networks," in *Proceedings of IEEE INFOCOM*, 2016.

[7] J.-M. Valin, F. Michaud, J. Rouat, and D. Létourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Proceedings of IEEE/RSJ IROS*, vol. 2, 2003.

[8] Q. H. Wang, T. Ivanov, and P. Aarabi, "Acoustic robot navigation using distributed microphone arrays," *Information Fusion*, vol. 5, no. 2, 2004.

[9] C. Zhang, D. Florêncio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 538–548, 2008.

[10] J. Xiong and K. Jamieson, "Arraytrack: A fine-grained indoor location system." in *Proceedings of USENIX NSDI*, 2013.

[11] X. Alameda-Pineda and R. Horaud, "A geometric approach to sound source localization from time-delay estimates," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 6, pp. 1082–1095, 2014.

[12] N. Stefanakis, D. Pavlidi, and A. Mouchtaris, "Perpendicular Cross-Spectra Fusion for Sound Source Localization with a Planar Microphone Array," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017.

[13] J. Liu, Y. Wang, G. Kar, Y. Chen, J. Yang, and M. Gruteser, "Snooping keystrokes with mm-level audio ranging on a single phone," in *Proceedings of ACM MobiCom*, 2015.

[14] S. Chung and I. Rhee, "vTrack: Virtual Trackpad Interface using mm-level Sound Source Localization for Mobile Interaction," in *Proceedings of ACM UbiComp*, 2016.

[15] W. Huang, X.-Y. Li, Y. Xiong, P. Yang, Y. Hu, X. Mao, F. Miao, B. Zhao, and J. Zhao, "Walkielokie: sensing relative positions of surrounding presenters by acoustic signals," in *Proceedings of ACM UbiComp*, 2016.

[16] H. Han, J. Yu, H. Zhu, Y. Chen, J. Yang, Y. Zhu, G. Xue, and M. Li, "SenSpeed: Sensing driving conditions to estimate vehicle speed in urban environments," in *Proceedings of IEEE InfoCom*, 2014.

[17] R. Nandakumar, K. K. Chintalapudi, V. Padmanabhan, and R. Venkatesan, "Dhwani: secure peer-to-peer acoustic NFC," in *Proceedings of ACM SIGCOMM*, 2013.

[18] Q. Wang, K. Ren, M. Zhou, T. Lei, D. Koutsonikolas, and L. Su, "Messages Behind the Sound: Real-Time Hidden Acoustic Signal Capture with Smartphones," in *Proceedings of ACM MobiCom*, 2016.

[19] W. Wang, A. X. Liu, and K. Sun, "Device-free gesture tracking using acoustic signals," in *Proceedings of ACM MobiCom*, 2016.

[20] W. Huang, Y. Xiong, X.-Y. Li, H. Lin, X. Mao, P. Yang, and Y. Liu, "Shake and walk: Acoustic direction finding and fine-grained indoor localization using smartphones," in *Proceedings of IEEE INFOCOM*, 2014.