Acoustic Strength-based Motion Tracking

LINFEI GE, Southern University of Science and Technology, China
QIAN ZHANG, The Hong Kong University of Science and Technology, China
JIN ZHANG, Southern University of Science and Technology, China
QIANYI HUANG, Southern University of Science and Technology, China

Accurate device motion tracking enables many applications like Virtual Reality (VR) and Augmented Reality (AR). To make these applications available in people's daily life, low-cost acoustic-based motion tracking methods are proposed. However, existing acoustic-based methods are all based on distance estimation. These methods measure the distance between a speaker and a microphone. With a speaker or microphone array, it can get multiple estimated distances and further achieve multidimensional motion tracking. The weakness of distance-based motion tracking methods is that they need large array size to get accurate results. Some systems even require an array larger than 1 m. This weakness limits the adoption of existing solutions in a single device like a smart speaker. To solve this problem, we propose Acoustic Strength-based Angle Tracking (ASAT) System and further implement a motion tracking system based on ASAT. ASAT achieves angle tracking by creating a periodically changing sound field. A device with a microphone will sense the periodically changing sound strength in the sound field. When the device moves, the period of received sound strength will change. Thus we can derive the angle change and achieve angle tracking. The ASAT-based system can obtain the localization accuracy as 5 cm when the distance between the speaker and the microphone is in the range of 3 m.

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

Additional Key Words and Phrases: strength, sound field, angle tracking, localization

ACM Reference Format:
Linfei GE, Qian ZHANG, Jin ZHANG, and Qianyi HUANG. 2020. Acoustic Strength-based Motion Tracking. *Proc. ACM Meas. Anal. Comput. Syst.* 0, 0, Article 0 (2020), 19 pages. https://doi.org/000000/000000

1 INTRODUCTION

Device motion tracking plays an important role in VR (Virtual Reality) and AR (Augmented Reality) applications. Commercial devices like the HTC Vive [1] use optical methods to localize VR headsets and its controllers. In this optical system, a calibrated laser beacon periodically emits a calibration flash and then sweeps a beam of light across the room. The photosensors on headset will receive the flash and the laser beam. The time difference between the flash and the laser beam will help to derive the relative angle of the photosensor and laser beacon. Several laser beacons and photosensors work together to derive pose and location of the headset. These optical-based tracking methods require expensive and dedicated laser beacons, as well as optical receivers to receive laser. These tracking systems only work on specialized devices, and thus are hard to use in people's daily life.

Authors' addresses: Linfei GE, lgead@connect.ust.hk, Southern University of Science and Technology, Shenzhen, China; Qian ZHANG, qianzh@cse.ust.hk, The Hong Kong University of Science and Technology, Hong Kong, China; Jin ZHANG, zhangj4@sustech.edu.cn, Southern University of Science and Technology, Shenzhen, China; Qianyi HUANG, huangqy@sustech.edu.cn, Southern University of Science and Technology, Shenzhen, China.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2020 Association for Computing Machinery.
2476-1249/2020/0-ART0 $15.00
https://doi.org/000000/000000

Acoustic-based tracking methods are recently becoming popular because of its good avaliability. Acoustic-based systems use speakers and microphones as transmitters and receivers. Some acoustic-based systems use phase information of the sound signal to detect distance variance. Compared with optical systems, acoustic-based systems are cheap and available in people's daily life. General computers and smartphones already have speakers and microphones, therefore, it is easy to implement acoustic-based tracking systems on these devices.

State-of-the-art acoustic-based motion tracking systems [6, 10, 17] have several limitations. Above motion tracking systems are all based on distance estimation. They can achieve multidimensional motion tracking like 2D or 3D tracking by arrays, either it is a speaker array or a microphone array. Assume that there are several speakers and one microphone. These systems derive distance between each speaker and microphone pair. Knowing the locations of these speakers in the speaker array, they can enable multidimensional localization of the microphone. Thus, actually, they are all distance-based systems. In order to achieve accurate localization result, the separations between these speaker anchors need to be large. The array size in CAT [6] even reaches 1 m. Thus, it is hard to implement these systems in a single device like a smart speaker. To better illustrate the limitation, we define TA ratio as the ratio of speaker-target distance over the array size, as shown in Figure 1. These distance-based motion tracking systems usually have a low TA Ratio. CAT [6] has a TA ratio less than 10, while MilliSonic [10] has a TA ratio about 13. With a TA Ratio of 10, a speaker array of size 30 cm is required if we track an object at a distance of 3 m. A normal smart speaker is much smaller than the required array size, so it is almost impossible to implement distance-based motion tracking systems in a single smart speaker device.

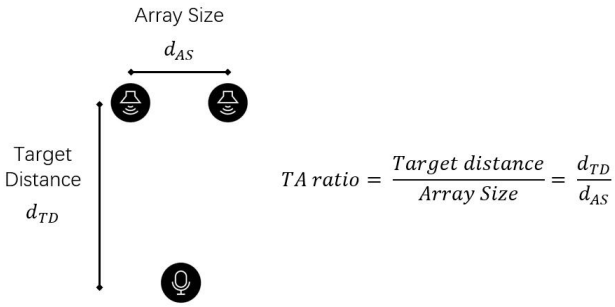


Fig. 1. Speaker array and definition of TA ratio

To address the limitation of low TA ratio, we propose our Acoustic Strength-based Angle Tracking (ASAT) System, which directly tracks angle information and significantly improves TA Ratio. The core idea is based on the strength information of a sound field. When two speakers emit continuous wave (CW) acoustic signals, the superposition of two sound signals will generate a sound field which contains uneven strength distribution. If two signals have different frequencies, the strength distribution will rotate at a frequency of the two signals' frequency difference. That is, at a certain location in the sound field, the microphone will see a regularly changing sound strength. The frequency of the strength change is constant if the microphone keeps static. If the microphone moves, similar to Doppler Effect, the frequency of the strength change will be different from the one when it remains static. The frequency difference is related with the angular speed. Therefore, based on the frequency difference, we will get the angular speed and achieve angle tracking. Besides, we can achieve distance tracking by phase-based methods. Combining the angle tracking and distance tracking, we can further achieve motiong tracking and reach an accuracy of 5 cm. Also, we can even reach a TA ratio more than 100, which is much better than traditional distance-based motiong tracking.

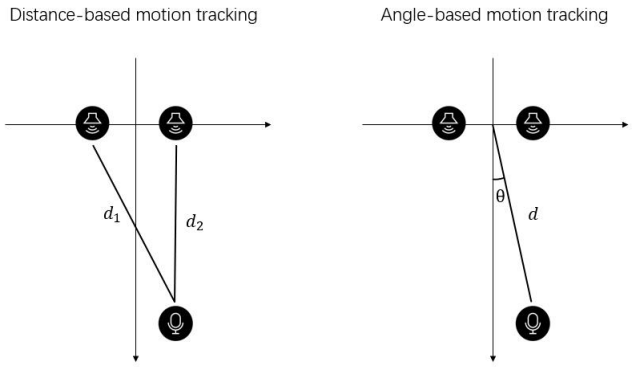


Fig. 2. Comparison of distance-based and angle-based motion tracking.

Here we give the comparison of traditional distance-based motion tracking system and our angle-based motion tracking system. As shown in Figure 2, there are two speakers and one microphone. Traditional distance-based motion tracking system considers two speakers independently and calculates the distance between each speaker and microphone, i.e., d_1 and d_2 separately. Given these two distances, it can achieve 2D localization. Obviously, smaller speaker separation (array size) results in lower accuracy. Our angle-based motion tracking system uses two speakers to generate a sound field and derives relative angle θ from the sound field. Combined with the distance d, we can achieve 2D localization. In theory, the accuracy is not affected by the speaker separation (array size).

The key point of our angle tracking system is to derive angle from the period of the strength variance. We observe that the period will change as the microphone moves. Thus we can calculate the angle from the difference between the observed period and the standard period. To achieve accurate angle tracking, there are still several challenges.

- Different devices have different clock, which causes frequency drift. The frequency drift will increase the tracking error of the system.
- We track angle and distance respectively in our system. How to properly derive the initial location and combine the angle and distance tracking results are important to achieve accurate motion tracking.
- Our system is based on sound field strength. However, the characteristic of the sound field strength is easily affected by multipath and environmental echoes.

We design some mechanisms to solve above mentioned challenges and implement our system on smartphones and achieve accurate angle tracking. We further implement a localization tracking system based on our ASAT System.

This work makes the following contributions.

- We model the sound field generated by two speakers. As far as we know, we are the first to consider two speakers together and model the sound field to derive its strength information.
- We propose a novel strength-based angle tracking method, which breaks the limit of low TA Ratio in traditional distance-based solutions. This makes it possible to achieve high tracking accuracy in a small device like a smart speaker. The angle tracking error is as low as 0.4 degrees.
- We design and implement a motion tracking system based on our proposed angle tracking methods on smartphones. It is able to enhance VR and AR applications without extra cost. The tracking error is as low as 5 cm.

2 SOUND FIELD BACKGROUND

In this section, we will give some basic background of sound and show how we generate a periodically changing sound field.

Our system is composed of two speakers and one microphone. Two speakers generate the sound field and the microphone senses the sound field. We know that if there is a speaker emitting sine wave, a microphone near it will receive it as a sine wave with a certain strength. If we add another speaker and it also emits sine wave with the same strength, there will be a superposition of the sound from these two speakers. Because of the phase difference of the two sounds, the superposition result may have an uneven sound distribution. If their frequencies are the same, only distance differences to the speakers will cause phase difference, and thus the sound field will be static. Some areas have enhanced strength and some have reduced strength. If their frequencies are different, time will also cause phase difference, and thus the sound field will be dynamic. Actually, the dynamic sound field is rotating around the center of the two speakers. If there is one microphone in the rotating sound field, it will sense a periodically changing strength.

We will give mathematic description below. First we introduce the expression of sine sound waves. Then we analyze the sound field generated by two speakers. Finally we will give an intuitional example of the generated sound field.

2.1 Expression of the Sound Wave

A sound wave is a kind of mechanical wave that propagates through a medium by particle-to-particle interaction. Normally, we express a sound wave by,

$$p = p_0 \sin(2\pi f t + \phi), \tag{1}$$

where p is the acoustic pressure, p_0 is the amplitude or strength of the sound wave. f is the frequency and ϕ is the initial phase of the sound wave.

Here, we express the sound wave as a sine wave. In fact, the sound in our daily life is made up by many sine waves. However, in our system, each speaker only plays a single frequency sine wave. The played sine wave can be easily expressed by Equation 1.

2.2 Generation of the Sound Field

Here we use two speakers to generate a sound field. Consider that there are two speakers, each of these speakers plays a sine wave of the same strength. A microphone records the sound in the sound field. The received signals at the microphone's position can be described respectively as follows,

$$p_1 = p_0 \sin(2\pi f_1 t + \phi_1), \tag{2}$$

$$p_2 = p_0 \sin(2\pi f_2 t + \phi_2). \tag{3}$$

The microphone records the superposition of these two speakers. The sound wave is a kind of mechanical wave, therefore, we can directly add them together. The summation is,

$$\begin{aligned} p_{sum} &= p_1 + p_2 \\ &= p_0 \sin(2\pi f_1 t + \phi_1) + p_0 \sin(2\pi f_2 t + \phi_2) \\ &= 2p_0 \sin(2\pi \frac{f_1 + f_2}{2} t + \frac{\phi_1 + \phi_2}{2}) \cos(2\pi \frac{f_1 - f_2}{2} t + \frac{\phi_1 - \phi_2}{2}) \end{aligned} \tag{4}$$

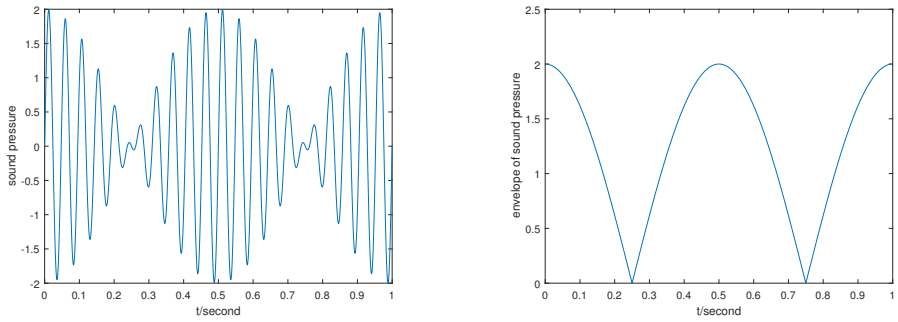
In Equation 4, there is a sine wave with frequency $\frac{f_1+f_2}{2}$ multiplied by a cosine wave with frequency $\frac{f_1-f_2}{2}$. The result is a product of a high-frequency sine signal and a low-frequency cosine signal.

2.3 An Example of the Generated Sound Field

To better illustrate the generated sound field, we give a simulation example in this section.

Figure 3 shows an example of sound wave p_{sum}, which is recorded by the simulated microphone. In this example, we use $f_1 = 20$ Hz and $f_2 = 22$ Hz. These frequencies are low enough to make the result clear in these figures. In Figure 3(a) we can clearly see high-frequency sine wave and low-frequency cosine wave. Figure 3(b) shows its envelope. For sound wave, the envelope represents the sound strength. The envelope, also the sound strength, changes periodically at a frequency of $f_0 = f_1 - f_2$. In our system, we set f_1 and f_2 to be inaudible band frequency near 20 kHz, so people will not hear the transmitted signal. The difference between f_1 and f_2 is small, so it is relatively easy to calculate f_0.

In this background section, we present that if we have two speakers playing sine wave at frequency f_1 and f_2, there will be a periodically changing sound field. A static microphone in this field will sense that the sound strength is changing at frequency $f_0 = f_1 - f_2$. Our Strength-based Angle Tracking System is based on the periodically changing sound field.



(a) An example of sound pressure.

(b) An example of sound pressure envelope.

Fig. 3. An example of sound pressure and its envelope in equation 4

3 STRENGTH-BASED ANGLE TRACKING

The above section gives the background of sound field and analyzes the strength characteristic when there are two speakers and one static microphone. In this section, we will give the detailed description of strength-based angle tracking method.

Intuitively, we leverage Doppler Effect to achieve angle tracking. Considering the sound field at a certain moment, as shown in Figure 4, we will see that there are some areas with high strength and some with low strength. The microphone will sense high sound strength at the bright areas. We use two speakers with different frequencies, so the sound field rotates around the center of the two speakers. In Figure 4, the center of the figure is the center of the two speakers. Considering Figure 3(b) and Figure 4, the bright areas in Figure 4 are corresponding to the peak in Figure 3(b). Dark areas in Figure 4 are corresponding to the valley in Figure 3(b).

A static microphone will sense the strength change just as shown in Figure 3(b). The frequency of the sensed strength change is the frequency difference f_0. If a microphone is moving in this sound field, because of the Doppler Effect, the frequency of strength change that we observe will be different from f_0. The period T_{obs} will also be different. Thus, by calculating the period difference ΔT, we can get angular speed and achieve angle tracking. We give the detailed mathematical analysis below.

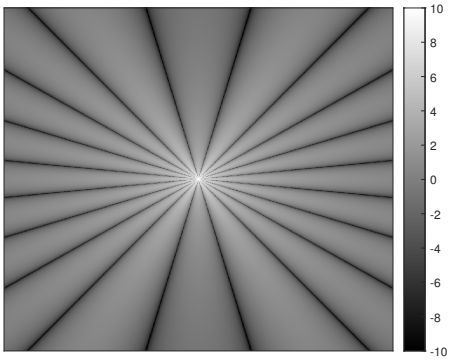


Fig. 4. An example of sound field

Consider the movement from point P_1 (θ_1, r_1) to point P_2 (θ_2, r_2).
At point P_1 (θ_1, r_1), the received signal from two speakers are as follows,

$$p_{11} = p_0 \sin(2\pi f_1 t + \phi_{11}), \tag{5}$$

$$p_{12} = p_0 \sin(2\pi f_2 t + \phi_{12}). \tag{6}$$

The microphone will hear the superposition of these two signals. It can be expressed as,

$$\begin{aligned} p_{1sum} &= p_{11} + p_{12} \\ &= p_0 \sin(2\pi f_1 t + \phi_{11}) + p_0 \sin(2\pi f_2 t + \phi_{12}) \\ &= 2p_0 \sin(2\pi \frac{f_1 + f_2}{2} t + \frac{\phi_{11} + \phi_{12}}{2}) \cos(2\pi \frac{f_1 - f_2}{2} t + \frac{\phi_{11} - \phi_{12}}{2}). \end{aligned} \tag{7}$$

Based on this signal, we calculate the envelope of it to get the strength change. The envelope of the superimposed signal is $\cos(2\pi \frac{f_1 - f_2}{2} t + \frac{\phi_{11} - \phi_{12}}{2})$. Its frequency is $f_1 - f_2$ and its phase is $\frac{\phi_{11} - \phi_{12}}{2}$. The period $T_0 = \frac{1}{f_1 - f_2}$. Similarly, at point P_2, the received signal is,

$$\begin{aligned} p_{2sum} &= p_{21} + p_{22} \\ &= p_0 \sin(2\pi f_1 t + \phi_{21}) + p_0 \sin(2\pi f_2 t + \phi_{22}) \\ &= 2p_0 \sin(2\pi \frac{f_1 + f_2}{2} t + \frac{\phi_{21} + \phi_{22}}{2}) \cos(2\pi \frac{f_1 - f_2}{2} t + \frac{\phi_{21} - \phi_{22}}{2}). \end{aligned} \tag{8}$$

The phase change of the envelope from point P_1 to point P_2 is,

$$\Delta\phi = \frac{\phi_{21} - \phi_{22}}{2} - \frac{\phi_{11} - \phi_{12}}{2}. \tag{9}$$

Because of the phase change, the period that we observed T_{obs} will be different from $T_0 = \frac{1}{f_1 - f_2}$. The difference is,

$$\begin{aligned} \Delta T = T_{obs} - T_0 = T_0 * \frac{\Delta\phi}{2\pi} \\ \Delta\phi = 2\pi \frac{\Delta T}{T_0}. \end{aligned} \tag{10}$$

From Equation 10, we can get phase change from period difference.

For point P_1, if we know its location, it is easy to get its phase difference ϕ_1. From above calculation, we get $\Delta\phi$. Also, we know that $\Delta\phi = \phi_2 - \phi_1$, then we have,

$$\phi_2 = \phi_1 + \Delta\phi = \phi_1 + 2\pi\frac{\Delta T}{T_0} \tag{11}$$

The possible location of point $P_2(\theta_2, r_2)$ is actually a hyperbola. Assume that the hyperbola is,

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1 \tag{12}$$

According to the definition of a hyperbola, we have

$$c^2 = a^2 + b^2, \frac{\phi_2}{2\pi}\lambda = 2a \tag{13}$$

Here, c is defined as the location of the speakers (c,0) and (-c,0). Because c is very small compared to r_2, we think that point P_2 is at the asymptote of the hyperbola. The slope of the asymptote can be expressed as,

$$\frac{b}{a} = \frac{\sqrt{c^2 - (\frac{\phi_2\lambda}{4\pi})^2}}{\frac{\phi_2\lambda}{4\pi}} \tag{14}$$

Then we can derive θ_2, the angle of point P_2,

$$\begin{aligned}\theta_2 &= arctan(\frac{\frac{\phi_2\lambda}{4\pi}}{\sqrt{c^2 - (\frac{\phi_2\lambda}{4\pi})^2}}) \\ &= arctan(\frac{\frac{(\phi_1+2\pi\frac{\Delta T}{T_0})\lambda}{4\pi}}{\sqrt{c^2 - (\frac{(\phi_1+2\pi\frac{\Delta T}{T_0})\lambda}{4\pi})^2}}).\end{aligned} \tag{15}$$

Finally we get the angle θ_2 of point P_2 and achieve angle tracking.

4 SYSTEM DESIGN

In this section, we present how we achieve angle tracking. Then we introduce distance tracking and localization. Also, we discuss about several challenges and their solutions.

4.1 Overview

The overview of our system is shown in Figure 5. There are two parts in our system, the transmitter end and the receiver end.

At the transmitter end, we need to implement a system that controls two speakers playing different content at the same time. We use stereo mode to control two speakers and make them play sine wave at different frequencies. Most audio systems support playing music at stereo mode. Stereo mode output includes a left channel and a right channel. Each channel controls one speaker. Thus, by using stereo mode output, we can control two speakers on one device, whether it is a smartphone or a computer. The played signals are simple sine waves at frequency f_1 and f_2, so they are $\sin(2\pi f_1 t + \phi_1)$ and $\sin(2\pi f_2 t + \phi_2)$. In our system, f_1 and f_2 do not change once the system is setup, so these signals can be either pre-generated or real-time generated. Most devices are compatible with our system.

At the receiver end, the received signal is recorded and processed. The recorded signal is first filtered by a high-pass filter to remove noise. In our daily life, the frequency of most voice is under 10 kHz, including noise. In our system, the transmitter sends detection signals near 20kHz. Thus, we can use a high-pass filter to remove

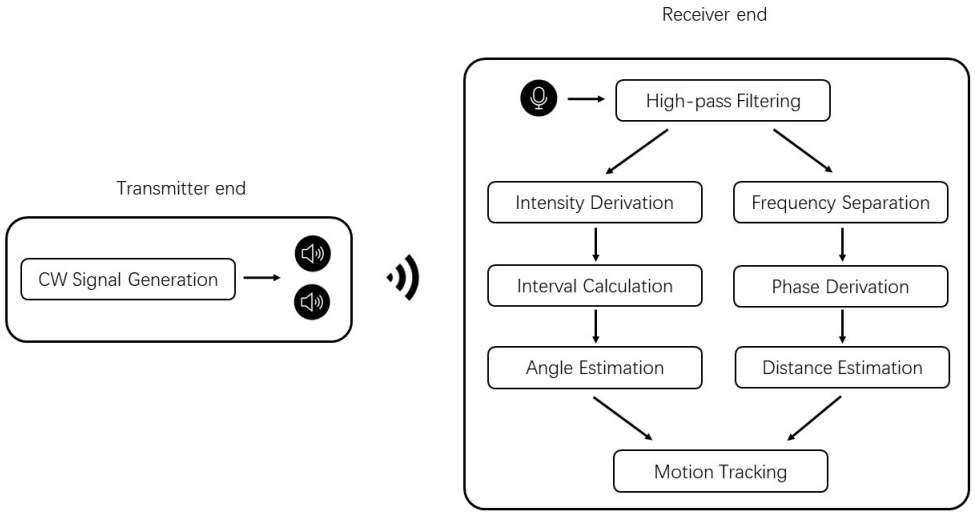


Fig. 5. System overview.

noise in the received signal. Once receiving the sound signal, we apply a high-pass filter with cut-off frequency 18 kHz. The noise filtering removes noise without affecting our detection signal.

One challenge is the frequency drift on different devices. The transmitter end and the receiver end are at different devices, so we need to consider the clock synchronization problem. To solve this problem, the system is required to have a clock calibration. At the beginning, the smartphone is required to keep static near the speaker. The speaker will emits some sine waves at different frequencies, eg. 5, 10, 15, 20 kHz. The smartphone will record these signals and find the frequency difference between the smartphone and the speaker. The frequency difference will be further used in angle and distance estimation to make the result more accurate.

For data processing, there are mainly two parts, angle estimation and distance estimation. For angle estimation, we first derive strength from the received signal. Then, we calculate period of the strength to get period difference. Finally we estimate angle from the period difference. For distance estimation, we first derive the phase of the received signal. Then, from phase change we will calculate distance change and achieve distance tracking. We further combine angle and distance results as well as the initial location to achieve localization. We will give detailed description below.

4.2 Strength-based Angel Tracking

From Section 3 we know that angle information can be extracted from the observed sound field change period. To achieve angle tracking, the key point is to find the observed period T_{obs}.

From Equation 15, we know that the key point of angle tracking is to find the $\Delta T = T_{obs} - T_0$ of the envelope of received signal. Here $T_0 = \frac{1}{f_1 - f_2}$ is the standard period, and T_{obs} is the period that we observed. To achieve angle tracking, we need to get an accurate period T_{obs}. There are several steps that can ensure an accurate period T_{obs}. We first derive the strength to observe the sound field change. Then we calculate the period T_{obs}. Finally we get angle tracking result from period T_{obs}.

Strength Derivation. To get the strength of the received signal, it is important to derive the envelope. The process of strength derivation is actually finding the envelope of the received signal.

The received signal is still a sine wave near 20 kHz. In our system, the sample frequency is 48 kHz. Thus, for the sine wave, each period has $48/20 = 2.4$ sample points. In theory, we can calculated the strength of each period. However, it is not accurate enough, because each period contains less than 3 points and it is not able to indicate the strength accurately. Thus, we choose to calculate the strength for each group of 6 sample points. A group of 6 sample points ensures that there are at least 2 periods and 2 periods are enough to indicate the strength.

In our system, we only care about the strength change instead of the absolute strength value. So for strength calculation, we choose square sum as the indicator. Strength is indicated by peak values of original sound signal. Square sum calculation properly describes the peak values' change, also the strength. Thus, we choose to use the square sum and it is accurate enough to show the strength change.

Period Calculation. The key step of angle estimation is to find the observed period T_{obs}.

From Equation 4, we know that the envelope is the absolute value of a cosine wave. Thus, valleys are easier to find than peaks. In Figure 3(b), we can observe that the valley of the curve is much easier to find. By these valleys we get the interval of each period T_{obs}.

Angle Estimation. In Equation 15, we give the relation of the angle and ΔT. With observed period T_{obs} and standard period T_0, we have ΔT and further derive the angle θ_2.

4.3 Distance Estimation and Localization

To achieve 2D tracking, we need not only the angle tracking but also the distance tracking. In our system, we choose phase-based distance tracking, which is efficient and accurate.

Phase-based distance tracking use the phase information to derive distance change. PAMT [5] already achieves mm-level motion tracking using phase-based method. In our system, we use simple sine wave as the sensing signal. Consider a speaker emits the wave and a microphone receives it. The distance between the speaker and the microphone can be expressed as $d = \lambda * (N + \frac{\phi}{2\pi})$. λ is the wave length and N is an integer. ϕ is the phase which is between 0 and 2π. If the distance d changes, the phase ϕ will also change. Knowing the phase change, we can achieve distance tracking. The detailed phase-based distance estimation is given below.

Phase Derivation. First, we need to derive phase from received signal. Assume that the received signal is S_r. It contains two sine waves at frequency f_1 and f_2. We multiple S_r by sin and cos. The frequency could be either f_1 or f_2. Without loss of generality, here we choose f_1

$$S_{r,sin} = S_r * sin(2\pi f_1 t) \tag{16}$$

$$S_{r,cos} = S_r * cos(2\pi f_1 t) \tag{17}$$

Then a low-pass filter is applied to $S_{r,sin}$ and $S_{r,cos}$. Phase ϕ_d is derived by

$$\phi_d = S_{r,sin} + S_{r,cos} * i, \tag{18}$$

where i is the imaginary unit. Then an "unwarp" operation is applied to solve the phase ambiguity problem. A $2\pi N$ will be added to it.

Distance Estimation. With frequency f_1, the wave length of the sound signal is

$$\lambda = v_s / f_1, \tag{19}$$

where v_s is the sound speed in air. The distance change Δd is

$$\Delta d = \frac{\phi_{d1} - \phi_{d2}}{2\pi} * \lambda. \tag{20}$$

Localization. With angle change $\Delta\theta$ and distance change Δd, we can easily find the location if we know the initial location. In the next section, we will introduce how we get the initial location.

4.4 Initial Location Estimation

To achieve motion tracking, besides angle and distance tracking results, we also need the initial location. In our system, the microphone is required to move along a predefined path at the beginning to get the initial position.

As Figure 6 shows, the microphone is required to go along a calibration path which is perpendicular to y axis. Generally, we define the direction of speakers as the direction of y axis. Assume that it moves from location A to location B. The cross point of calibration path and y axis is the point C. While moving, we continuously track the angle and distance change. If the microphone moves from A to B, We can observe that the distance first decrease then increase. By finding the minimum point of distance, we know when it is at point C. Knowing the start point A, we get the angle change $\Delta\theta_{AC}$ and distance change Δd_{AC} when moving from A to C. Also, we define d_{AO} and d_{CO} as the distance between AO and CO.

$$d_{AO} - d_{CO} = \Delta d_{AC} \tag{21}$$

$$d_{AO} sin(\Delta\theta_{AC}) = d_{CO} \tag{22}$$

By solving these two equations, we get d_{AO}. Also, we know $\Delta\theta_{AC}$. The initial location of point A is $(\Delta\theta_{AC}, d_{AO})$ in polar coordinates.

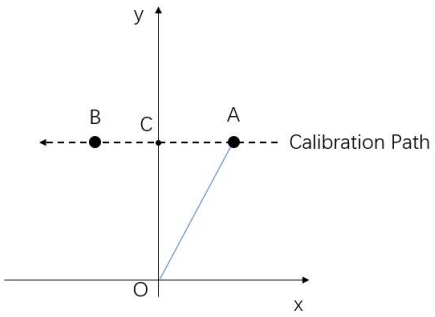


Fig. 6. Calibration.

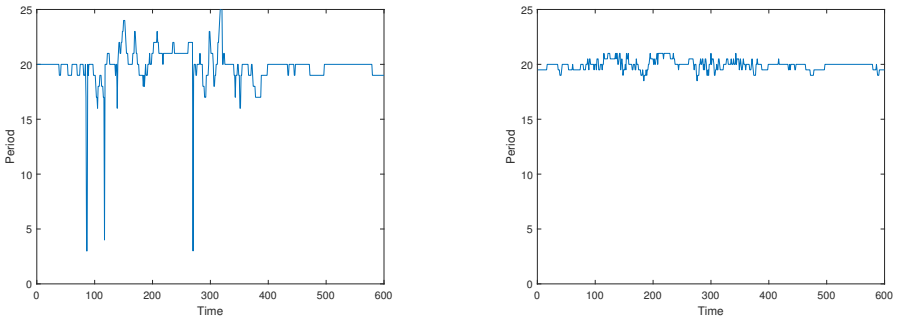
4.5 Multipath and Diversity

An important challenge is the influence of multipath and environmental echoes. At relatively far locations, multipath effect is a problem for our strength-based motion tracking. Our system is based on signal strength, which is easily affected by multipath effect. Sometimes, in "Period Calculation", it is hard to find an avaliable period from the affected signal. Thus, we use Time-Division Multiplexing to improve the availability of period at far locations.

We observed that if we only use one group of frequency, for example 20000/20100 Hz, the sound field strength change is very small sometimes. To solve this, we use two group of frequencies to provide diversity and higher reliability. We use Time-Division Multiplexing, so they will not interfere each other. We get two groups of data and it provides redundancy and increase the reliability.

Figure 7 shows the comparison of period result without/with TDM. In Figure 7(a), some points are very small because the corresponding periods are not available. After using TDM, in Figure 7(b), unavailable points are

much less. Thus, the tracking result is more accurate. Figure 8 shows angle tracking result at 3 m distance and it shows that multiple groups of frequencies do reduce the angle tracking error at far locations.



(a) Period calculation result without TDM. (b) Period calculation result with TDM.

Fig. 7. Period calculation result without/with TDM

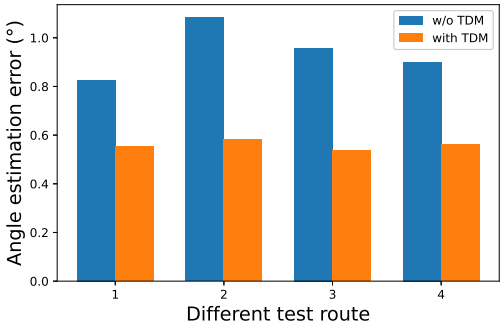


Fig. 8. Comparison of w/o TDM and TDM.

5 EVALUATION

5.1 System implementation

We implement our system on Android smartphones and laptops. At the transmitter end, we use a laptop to control speakers to emit signals at different frequencies. As we mentioned above, we use a stereo output mode to control these two speakers and make them to play sine waves of different frequencies. Two speakers are placed together with a distance of 8 cm. Actually the separation distance does not influence the performance. We choose to use the separation of 8 cm because it is the minimum separation due to the shell of the speaker.

At the receiver end, we use a smartphone to record the acoustic signal. The recording sampling rate is 48 kHz, which is supported by almost all smartphones. We use 16 bit quantification to make the voice accurate enough to identify the acoustic strength. 16 bit quantification means that there are 2^{16} different strength levels. Pulse-code modulation (PCM) encoding is used to record the sound. PCM encoding does not change the waveform, so we

can directly analyze it. The received signal is then passed to a laptop. We use MATLAB to process the signal and analyze the tracking result.

To make our evaluation as accurate as possible, we use a linear actuator with stepper motor to accurately control the movement. The movement resolution is up to 0.03 mm. As Figure 9 shows, we use a laptop as the signal generator and power supply for the speakers. A smartphone moves around and record the signal.

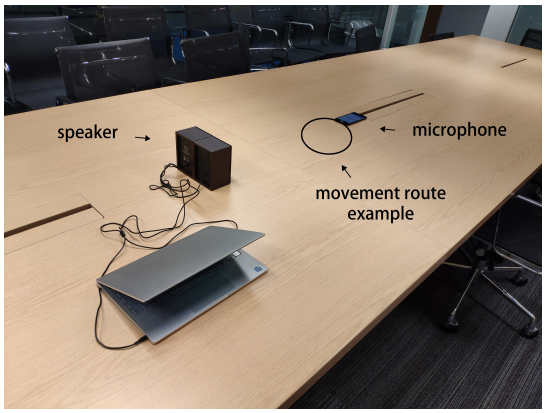


Fig. 9. Evaluation scenario.

We first evaluate angle tracking accuracy. Then we test the distance tracking accuracy. After that, we evaluate the overall motion tracking accuracy. Finally, we evaluate the possible influence of several factors.

5.2 Angle Evaluation

We evaluate the angle tracking performance by making the smartphone move around the speakers by a certain angle. The error is defined as the difference between estimated angle and ground truth. We first mark some points on the ground as the reference points. The smartphone moves from one point to another. Also, we measure the angle change among these points as ground truth. We measure and evaluate multiple times to reduce measurement error.

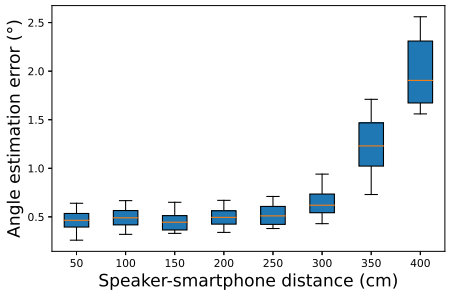


Fig. 10. Angle estimation error.

Figure 10 shows the angle estimation error v.s. distance. We find that in a certain range, our system can achieve an estimation error about 0.4 degree. However, as the distances become larger, affected by multipath effect, the angle estimation error becomes larger, even larger than 2 degrees.

5.3 Distance Evaluation

We use linear actuator with stepper motor to control the movement of the smartphone. While the smartphone is moving, it also keeps recording, so we can get the phase change from the received signal. Thus we can derive the distance change from it. The stepper motor controller is programmable, so we can predefine the movement and use it as the ground truth. The accuracy of the device is up to 0.03 mm, which is accurate enough for the evaluation.

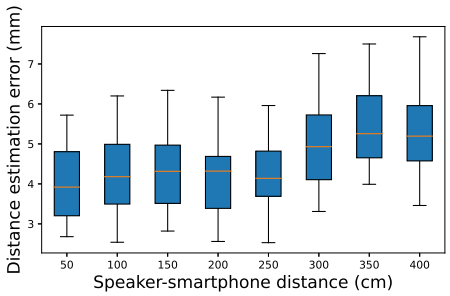


Fig. 11. Distance estimation error.

Figure 11 shows the distance estimation error v.s. distance. We use the phase-based method to derive the distance information. This kind of method has been well investigated. We find that it performs well at all distances. Most of the distance estimation errors are less than 5 mm. It is accurate enough for our motion tracking.

5.4 Motion Tracking Evaluation

We conbine angle tracking and distance tracking results together to evaluate the motion tracking. The motion tracking results could be used in human tracking or gesture recognition. Thus, we test the system's motion tracking ability by making the smartphone to go through a route, and then compare the estimation result and the ground truth route. The average difference between the estimation result and the ground truth route is the motion tracking error. We test several different routes. Circle routes are classical because it has both distance change and angle change. Figure 12 shows the comparison of estimated route (blue line) and ground truth (orange line).

We evaluate the tracking method in two ways. One is to predefine a route and make the smartphone go through the route. The other is to make the smartphone draw some figures and record the route. The results of these evaluations are shown in Figure 13 and Figure 14.

Figure 15 gives the CDF curve of overall motion tracking result within the range of 3 m.

The tracking estimation results shows that our system can achieve a localization error about 5 cm.

5.5 Influence of other Factors

Separation Distance. We evaluate the influence of the separation of two speakers. To reach a separation of 1 cm, we use two identical smartphone speakers as the transmitter. However, the strength of smartphone speaker is much smaller than normal speakers at 20 kHz, so we use signals near 10 kHz to make the sound strength become comparable with normal speakers. This will not change the results about the influence of separation distance.

From Figure 16, we find that tracking accuracy is not affected by different separation distance. Thus, we can implement our system in the case with a small speaker separation distance, even a separation distance of 1 cm.

Noise Influence.

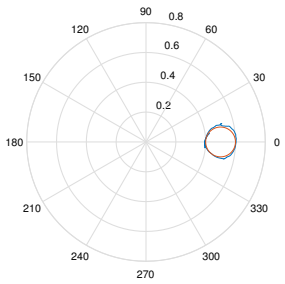


Fig. 12. Motion Tracking Example.

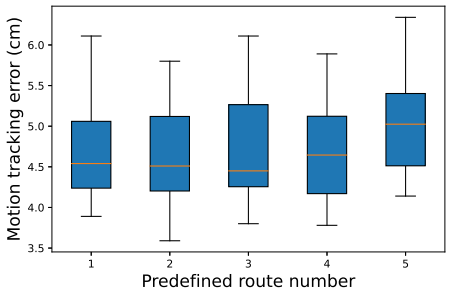


Fig. 13. Motion Tracking error of predefined route.

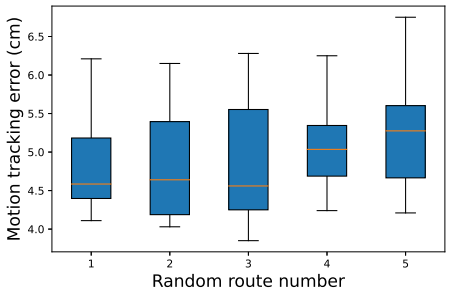


Fig. 14. Motion Tracking error of random drawing.

We also evaluate the influence of the noise. Our system is based on sound field strength. Thus general voice like talking and music playing in daily life may influence our system. We test several scenarios in different noise level. The results are in Figure 17. The results show that background noise significantly affects the system performance because our system is based on acoustic strength. Once the noise strength is relatively high, the sound field will be affected, and it will be hard to detect the angle information.

Different Environment. We evaluate our system in several different rooms to find whether the environment will influence the tracking performance. We choose three places: 1. A meeting room. 2. A living room. 3. A

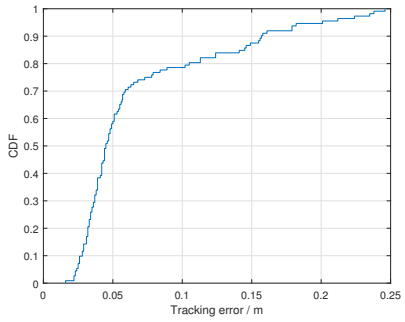


Fig. 15. Tracking error CDF.

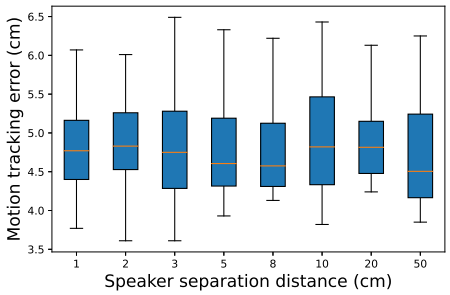


Fig. 16. Motion tracking error v.s. different separation distance.

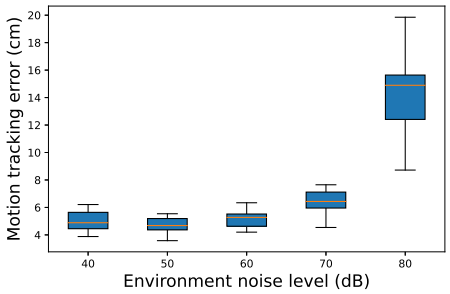


Fig. 17. Motion tracking error v.s. different noise level.

corridor. The evaluation results in Figure 18 show that different environments have similar motion tracking error. Different environments do not influence the tracking performance.

Different Phones. We use several different devices to record the signal to find whether different devices will influence the tracking performance. We use three devices: 1. Xiaomi MI5 2. SAMSUNG S6 3. A laptop. The evaluation results show that the errors are similar. Different devices will have similar performance in our system.

Multiple Users. We use several smartphones to record the signal at the same time to see the influence of multiuser. The result shows that there is not much influence whether there are 1,2 or 3 users. The reason is that

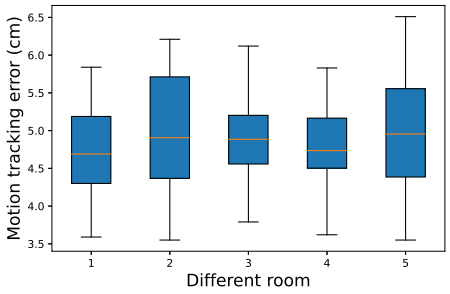


Fig. 18. Motion tracking error v.s. different room.

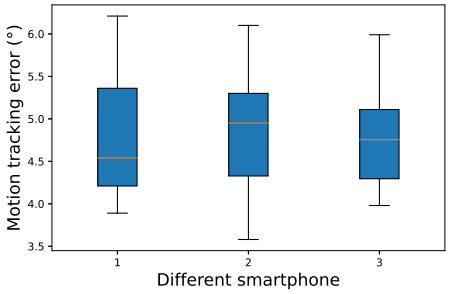


Fig. 19. Motion tracking error v.s. different phone.

the smartphone need only to record the sound without interaction with speakers, so multiple smartphones will not influence the motion tracking accuracy.

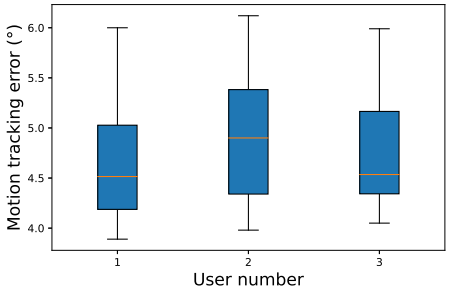


Fig. 20. Motion tracking error v.s. multiple users.

6 RELATED WORK

Here, we review existing works on motion tracking and classify them based on the employed sensors/signals.

6.1 IMU-based Motion Tracking

IMU (Inertial Measurement Unit) sensors are widely assambled in smartphones, and thus they are commonly used for motion tracking. [2] and [3] show that we can fuse multiple IMU information including linear acceleration and rotational rate to achieve motion tracking. However, the tracking error will increase as time goes by, because we need double integration of acceleration to get displacement and the error accumulates during integration.

6.2 RF-based Motion Tracking

Radio Frequency (RF) is used on tracking and localization for many years. RF signals propagate at light speed. It can sense objects far away from the sensor in a short time. Thus, radars are widely used to detect moving objects in outdoor systems. For indoor systems, WiFi signal is widely used in RF sensing, because WiFi devices are ubiquitous in people's daily life.

ArrayTrack [13] achieves indoor localization using multiple-input multiple-output (MIMO) techniques. By several MIMO WiFi devices, ArrayTrack has a median localization error of 23 cm. WiDraw [9] enables hands-free drawing in the air with commodity WiFi cards. It achieves hands-free motion tracking by analyzing Angle-of-Arrival values based on WiFi CSI. WiDraw uses 25 antennas and achieves a tracking median error less than 5 cm. The implemented word recognition system can reach an accuracy of 91%. Tagoram [14] uses commercial RFID to achieve tracking. The median error is 12 cm for unknow path. RF-IDraw [11] uses 8 RFID antennas with different spacing to track the tag with high resolution and low ambiguity. The median error is 3.7 cm.

6.3 Acoustic-based Motion Tracking

In recent years, acoustic-based sensing is more and more popular. Compared with RF signals, acoustic signals propagates much slower, and thus is easier to handle. Also, acoustic devices like speakers and microphones are very cheap and they are widely assembled in devices like smartphones and laptops.

Cricket [8] uses both ultrasound and RF signal and estimates the delay of two signals to derive distance. It achieves 12 cm error with 6 beacon nodes. Swadloon [4] achieves meter-level indoor localization using the phase of acoustic signals. However, meter-lever accuracy is not sufficient for VR or AR applications. Strata [16] achieves device-free tracking using acoustic signals. However, its detection range is less than 1 m. LLAP [12] also implements phase-based tracking and the 1D accuracy is less than 1 cm. FingerIO [7] is a finger tracking solution to achieve interaction with devices. It uses a special designed Orthogonal Frequency Division Multiplexing (OFDM) signal because of its good autocorrelation property. By analyze the echo of the OFDM signal, it derives the motion of a moving finger. The 2-D finger tracking of FingerIO can achieve an average accuracy of 8 mm. AAMouse [15] tracks people's hands to imitate a mouse. It uses the existing speaker and microphone on a smartphone. The speaker emits inaudible sound pulses and uses the frequency shift to estimate the speed and displacement. With a quick calibration and its initial location, AAMouse achieves real-time localization. SoundTrak [17] achieves tracking with a speaker ring and several microphones. The user wears a ring with a speaker. The speaker sends an acoustic signal at a specific frequency. Some microphone nodes at different locations receive the signal. By analyzing the phase information of each microphone, it will be able to localize the finger's position. CAT [6] achieves device tracking by multiple speakers and a single microphone. It uses multiple speakers to transmit inaudible sound at different frequencies. A smartphone with microphone receives these signals and continuously estimates the speed and distance. It uses a distributed Frequency Modulated Continuous Waveform (FMCW) system to accurately estimate the distance and achieve localization. With the help of IMU, CAT can reach a median error less than 8 mm.

CAT and SoundTrak are similar to our system, while we have a better TA ratio than them. CAT has a TA ratio about 10 and SoundTrak's TA ratio is less than 2. Our system can achieve a TA ratio of 100.

7 DISCUSSION

We present a novel motion tracking system leveraging strength-based angle tracking. Traditional distance-based motion tracking systems have low TA Ratio and it limits the application scenario. Our strength-based angle tracking solves this problem by using the sound field characteristics. We use two speakers to play sine waves at different frequencies. These two speakers will generate a sound field whose strength is periodically changing. By analyzing the strength of the received signal, we know the sound field strength change and derive the angle change from it. Based on the angle tracking result, we further design a tracking system and achieve localization. The evaluation results show that our system can achieve motion tracking accuracy about 5 cm.

In our strength-based system, we consider two speakers together, while distance-based systems usually consider each speaker individually. Two speakers generate a sound field and help to derive the motion of the microphone. The idea of considering two speakers together may inspire more applications in motion sensing.

However, there are still some limitations.

- Our system is based on strength, whereas strength may be affected by noise in the background. Although we try to remove these noises, it is hard to completely remove them. The strength of the detection signal in our system is about 75 dB. Once the noise strength is near or higher than the detection signal, the motion tracking performance will be much lower.
- So far, we can only achieve 2D motion tracking, because we only use two speakers. In theory, we can apply more speakers to achieve 3D motion tracking. However, it is hard to model the sound field generated by three speakers. Analyzing the sound field generated by three or more speaker is also one of our future directions.
- Like many tracking-based system, we suffer from interruption problem and accumulated error. In this kind of system, we have to run the system uninterruptedly. Once we stop and restart, we have to do cailbration again to calculate the initial location.

8 CONCLUSION

We introduce a novel motion tracking system which achieves strength-based motion tracking. We consider two speakers together and make them generate a periodically changing sound field. By using the sensed strength of the sound field, we derive angle information and achieve angle tracking. We further implement a motion tracking system leveraging the strength-based angle tracking system. The strength-based motion tracking system breaks the limitation of low TA Ratio of traditional distance-based systems. We evaluate our system on multiple smartphones and scenarios. The evaluation results show that our system can achieve motion tracking accuracy about 5 cm in a range of 3 m.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61701216, Shenzhen Science, Technology and Innovation Commission Basic Research Project under Grant No. JCYJ20180507181527806, Guangdong Provincial Key Laboratory under Grant No. 2020B121201001, Guangdong Innovative and Entrepreneurial Research Team Program (2016ZT06G587), Shenzhen Sci-Tech Fund (KYTDPT20181011104007), RGC under Contract CERG 16203719, 16204418, the Guangdong Natural Science Foundation No. 2017A030312008 and NSFC No. 62002150.

REFERENCES

[1] [n.d.]. HTC VIVE. https://www.vive.com/. Accessed May 15, 2020.
[2] Raúl Feliz Alonso, Eduardo Zalama Casanova, and Jaime Gómez García-Bermejo. 2009. Pedestrian tracking using inertial sensors. (2009).
[3] Hassen Fourati. 2014. Heterogeneous data fusion algorithm for pedestrian navigation via foot-mounted inertial measurement unit and complementary filter. *IEEE Transactions on Instrumentation and Measurement* 64, 1 (2014), 221–229.

[4] Wenchao Huang, Yan Xiong, Xiang-Yang Li, Hao Lin, Xufei Mao, Panlong Yang, and Yunhao Liu. 2014. Shake and walk: Acoustic direction finding and fine-grained indoor localization using smartphones. In *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*. IEEE, 370–378.

[5] Yang Liu, Wuxiong Zhang, Yang Yang, Weidong Fang, Fei Qin, and Xuewu Dai. 2019. PAMT: Phase-based acoustic motion tracking in multipath fading environments. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2386–2394.

[6] Wenguang Mao, Jian He, and Lili Qiu. 2016. CAT: high-precision acoustic motion tracking. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. ACM, 69–81.

[7] Rajalakshmi Nandakumar, Vikram Iyer, Desney Tan, and Shyamnath Gollakota. 2016. Fingerio: Using active sonar for fine-grained finger tracking. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 1515–1525.

[8] Nissanka B Priyantha, Anit Chakraborty, and Hari Balakrishnan. 2000. The cricket location-support system. In *Proceedings of the 6th annual international conference on Mobile computing and networking*. 32–43.

[9] Li Sun, Souvik Sen, Dimitrios Koutsonikolas, and Kyu-Han Kim. 2015. Widraw: Enabling hands-free drawing in the air on commodity wifi devices. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*. 77–89.

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

[11] Jue Wang, Deepak Vasisht, and Dina Katabi. 2014. RF-IDraw: virtual touch screen in the air using RF signals. *ACM SIGCOMM Computer Communication Review* 44, 4 (2014), 235–246.

[12] Wei Wang, Alex X Liu, and Ke Sun. 2016. Device-free gesture tracking using acoustic signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. 82–94.

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

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

[15] Sangki Yun, Yi-Chao Chen, and Lili Qiu. 2015. Turning a mobile device into a mouse in the air. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*. 15–29.

[16] Sangki Yun, Yi-Chao Chen, Huihuang Zheng, Lili Qiu, and Wenguang Mao. 2017. Strata: Fine-grained acoustic-based device-free tracking. In *Proceedings of the 15th annual international conference on mobile systems, applications, and services*. 15–28.

[17] Cheng Zhang, Qiuyue Xue, Anandghan Waghmare, Sumeet Jain, Yiming Pu, Sinan Hersek, Kent Lyons, Kenneth A Cunefare, Omer T Inan, and Gregory D Abowd. 2017. Soundtrak: Continuous 3d tracking of a finger using active acoustics. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 30.