Human-centric Computing and Information Sciences

RESEARCH Open Access

Indoor acoustic localization: a survey

Manni Liu[1*], Linsong Cheng[2], Kun Qian[2], Jiliang Wang[2], Jin Wang[3] and Yunhao Liu[1,2]

*Correspondence:
liumanni@msu.edu
[1] Computer Science and Engineering Department, Michigan State University, 428 South Shaw Lane, East Lansing 48824, USA
Full list of author information is available at the end of the article

Abstract

Applications of localization range from body tracking, gesture capturing, indoor plan construction to mobile health sensing. Technologies such as inertial sensors, radio frequency signals and cameras have been deeply excavated to locate targets. Among all the technologies, the acoustic signal gains enormous favor considering its comparatively high accuracy with common infrastructure and low time latency. Range-based localization falls into two categories: absolute range and relative range. Different mechanisms, such as Time of Flight, Doppler effect and phase shift, are widely studied to achieve the two genres of localization. The subcategories show distinguishing features but also face diverse challenges. In this survey, we present a comprehensive overview on various indoor localization systems derived from the various mechanisms. We also discuss the remaining issues and the future work.

Keywords: Indoor localization, Acoustic signals, Time of flight, Frequency-modulated continuous wave, Doppler effect, Phase shift

Introduction

Indoor localization is essential to enable location based services (LBS) such as indoor navigation [27, 35, 37, 84], health rehabilitation [14, 45, 49, 54] and human–computer interaction (HCI) devices [47, 70, 83, 85]. Aiming at higher accuracy, shorter time latency and lower infrastructure requirement, many localization systems have been designed for various scenarios. Generally, localization algorithms can be described as a two-stage procedure [41]. In the first step, geographic information such as distances and angles are measured. In the second step, the target is located using those data. Physical phenomena such as Time of Flight (ToF) [9, 40, 50, 51], Doppler Effect [20, 42, 76, 85] and phase shift [64, 70, 77, 81] assist in the first step. Geometric knowledge [9, 76, 81] and optimization methods [24, 40, 42] are common choices for the second step.

In addition to the acoustic signal on which we will concentrate in this survey, other ways have also been exploited for localization systems in many scenarios. Inertial sensors [49, 62, 87] are frequently utilized due to their highly accessible equipments and straightforward principles. An intuitive idea is using the double integration of the acceleration to estimate the displacement and applying the gyroscope to predict the direction, which however leads to significant location error with even a small measurement error [76]. Although it is challenging to achieve high accuracy with inertial sensors, we can still leverage them as an auxiliary tool in acoustic localization. For example,

Springer Open

Montage [80] attains an initial position through elapsed time between the two time-of-arrivals which is introduced in detail in "Elapsed time between two time-of-arrivals" section, and updates the position through the movement vector from inertial sensors. CAT [42] leverages inertial sensors to improve the accuracy while plugging their readings into an objective function. Tracko [29] applies inertial sensors to correct its 3D estimation. Swadloon [27] applies the accelerometer and the gyroscope to obtain the direction of the acoustic source.

Radio frequency (RF) signals share many characteristics with acoustic signals. To some extent, many ideas applied in acoustic localization are borrowed from RF signals. For instance, FingerIO [46] employs orthogonal frequency-division multiplexing, Starta [77] uses channel impulse response and Guoguo [40] applies decoded symbols, which are originally designed for digital communication. Currently, commercial companies like Google invest a lot in RF localization. Their ongoing products, like Soli [36], can achieve sub-millimeter accuracy. Despite the attractive accuracy, Soli applies millimeter wave, which is very demanding on infrastructure and not applicable for wide deployment. When compared to more common RF signals like Wi-Fi, acoustic localization gains strength as what it requires are mainly microphones and speakers, which are widely equipped on many smart devices. Another advantage of the acoustic signal is that the sound speed is much lower than the speed of the RF signal, which implies potential for higher accuracy [59].

Vision localization is currently widespread on the market, e.g. Kinect [63], Wii [60] and LeapMotion [71], and etc. Different from RF localization, vision localization makes no interference to ubiquitous RF based devices [61]. The main limitation of vision localization is that, it is severely constrained by the lightening condition [78] and suffers from the privacy issue [53]. Currently it is not applicable to smartphones and smart watches because of its high computation overhead and infrastructure requirement [13].

Acoustic localization wins a place as it can achieve a relatively high accuracy and low time latency with equipments already embedded in current smart devices. The acoustic signal is first applied in outdoor localization to detect aircrafts, which is substituted by radars as the RF signal is faster and more effective for long distances. When it comes to indoor localization where GPS does not work well [80], the acoustic signal becomes irreplaceable due to its lower speed, which leads to high accuracy when estimating ToF [59]. Cricket [51] is the first indoor localization system which adopts acoustics and utilizes ToF. It is actually a combination of acoustic signals and RF signals. It has a very impressive accuracy of 12 cm, while being prevented from wide deployment due to its high noise. After Cricket, ToF becomes widespread in acoustic localization. Later in 2012, Doppler effect is introduced in [48] to estimate the motion direction and achieves the mean angular error within 18°. Swadloon [27] further turns Doppler effect to phase shift and achieves a maximum tracking error of 1.73 m in an area of 2000 m^2. AAmouse [76] depends on Doppler effect to track a mobile phone. Doppler effect cannot enable fine-grained localization because of the time-frequency resolution problem [12]. FingerIO [46] uses phase shift to dissolve the unsynchronization issue in ToF based localization. Its high accuracy inspires LLAP [34] and Strata [77], thus phase shift is widely used to acoustic localization.

Localization can be categorized as range-based localization and range-free localization. Range-free localization is more common with the RF signal when high accuracy is not the primary principle [59] and coarse accuracy is enabling to most sensor network applications [23]. Range-based localization is rigid with the accuracy of measured geographic data, on which the acoustic signal performs better due to its lower speed [59]. To the best of our knowledge, existing acoustic localization systems are range-based, which we further divide into two categories: absolute range based localization and relative range based localization. Absolute range based localization lRelative range based localizationeverages absolute distances between the target and different anchor nodes to calculate the coordinate of the target. Each location update is obtained from scratch rather than renewing the previous location. Relative range based localization first obtains an initial location of the target, then updates the location through monitoring the subsequent motion of the target.

This survey classifies acoustic localization systems into absolute range based localization and relative range based localization. We further subdivide each class based on the principles the geographic data is measured. We believe the performance of a localization system is more dependent on the granularity of the geographic data. The rest of the survey is arranged as follows: we summarize the mutual challenges of acoustic localization systems in "Challenges" section and provide a notation table for better reading in "Notations" section. Absolute range based acoustic localization and relative range based localization are introduced in "Absolute range based localization" and "Relative range based localization" sections respectively. In "Future work" section, we analyze the future work. We conclude our survey in "Conclusion" section.

Challenges

In spite of different principles and methods, all acoustic localization systems face the following three mutual challenges, which should be taken into considerable consideration when we design a new system:

- The first one is signal-to-noise ratio (SNR). SNR is the ratio between the power of our desired signal to its background noise [59]. The higher the ratio is, the better the localization system is. If SNR is too low, the receiver may have trouble detecting the desired signal. The received signal is attenuated for long-distance travel and distorted by the communication channel. It is also constrained by the maximum energy the transmitter can provide [50].
- Multipath effect is another common issue [50]. We hope to detect the signal reflected by the target or coming from the direct path. Due to the complex environment, the received signal is the superposition of signals reflected by different objects. Sometimes we assume signals reflected by objects whose distance to the transmitter exceeds a certain distance are too weak to produce an effect. Still, methods are in need to distinguish the target.
- The third mutual challenge is the frequency selection of speakers and microphones. As we have mentioned above, a satisfying localization system shows low requirement on infrastructure, so we would better make use of speakers and microphones embed-

ded in COTS smart devices. For example, EchoLoc [9] employs a smartphone with two speakers to track a hand. It has taken into full consideration that the top speaker and the bottom speaker are designed for various purposes, so their frequency responses are very different.

In addition to the three mutual challenges, different methods are accompanied by new challenges. For example, ToF suffers from the time asynchrony and Doppler effect undergoes the time-frequency resolution problem [12]. Those challenges will be discussed when we introduce specific methods in the following sections.

Notations

(Table 1).

Absolute range based localization

Absolute range based localization monitors the range between the transmitter and the receiver. Usually, the target serves as a transceiver or a reflecting object. In contrary to relative range based localization which studies the displacement of the target, absolute range based localization investigates the flight of the signal between the target and an anchor node. It achieves tracking through continuous localization from scratch. The main mechanism employed in this category is ToF.

Time of Flight

Time of Flight (ToF) is the time it takes for a signal to travel from its transmitter to its receiver [59]. If we denote the absolute range between the transmitter and the receiver by d, we have $d = c \times t$ where c is the speed of sound and t is the ToF. With several

Table 1 Symbols used in this paper

Symbol	Description
A	Amplitude
B	Bandwidth
c	The sound speed
d	Distance
f_{max}	The maximum frequency
f_{min}	The minimum frequency
f_{Rx}	The frequency of the received signal
f_{Tx}	The frequency of the transmitted signal
F_s	Sampling rate
k	The slope of a chip signal
Rx	Receiver
$S(t)$	Received signal
t	Timestamp (usually with a subscript)
t_d	Propagation delay
T	Time interval
Tx	Transmitter
v	The speed of the target
α	Attenuation factor
(x, y, z)	3D location coordinate

anchor nodes (usually one anchor node for 1D localization, two for 3D and three for 3D) and their corresponding absolute ranges to the target, we can locate the target.

There are three primary challenges for ToF [41, 50].

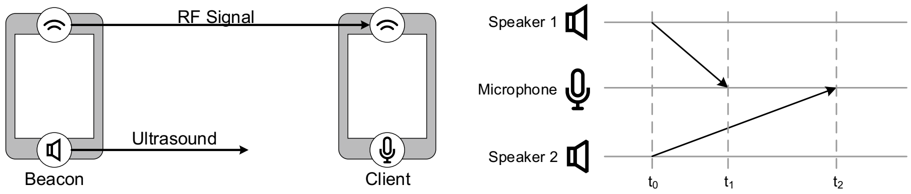
- The first one is the speed uncertainty. We know that the speed of sound varies when the temperature and the humidity change. 1 °C offset in the temperature will cause about 0.606 m/s drift in sound speed when the humidity is 0%. Some smart devices do have in-built thermometers and hygrometers. Systems like [19, 59] provide details on how to improve their accuracy with thermometers equipped in COTS smartphones, but sensors in COTS smartphones often fail to provide us with precise measurements of the surrounding environment.
- The main issue we need to conquer in ToF based localization is about unsynchronized clocks. ToF is measured by the difference of timestamps taken by local clocks from the transmitter and the receiver. It is of high possibility that they are not synchronized. For indoor scenarios where the mignitude of t is usually one hundred-thousandth of c, error caused by time offset is far more serious than that caused by speed offset. Some localization systems seek helps from complicate instruments like getting synced with atomic clocks from GPS. Yet most of them design subtle mechanisms to solve the issue.
- Another challenge for ToF is the sending uncertainty t_s and receiving uncertainty t_r. The sending uncertainty refers to the misalignment between the transmitter timestamp and the actual signal emission time, while the receiving uncertainty is the misalignment between the receiver timestamp and the actual signal reception time. In [50], researchers conducted several experiments to measure the magnitude of $t_s + t_r$ for a COTS mobile phone. They find that the time offset can add up to be several milliseconds. Factors like system load, software delay and interrupt handling delay can cause uncertainty.

ToF can be partitioned into time difference of arrival, one-way time-of-flight and round-trip time-of-flight. In this section, we further divide one-way time-of-flight into elapsed time between two time-of-arrivals and one-way flight with an anchor network. The former one is designed specially for device-to-device localization which asks both devices to be equipped with one microphone and one speaker. The second type requires one side serves as a speaker and the other side as a "listener".

Time difference of arrival

Time Difference of Arrival (TDoA) eliminates the requirement of the emission time [31]. As long as signals are transmitted simultaneously or at a known pattern, we can calculate the distance with TDoA.

Cricket [51] leverages TDoA between the RF signal and the ultrasonic signal to achieve in-building localization. With six beacons deployed on the ceil, Cricket can achieve a median distance error of 12 cm to locate a mobile listener. Each beacon simultaneously emits a specially-designed RF signal and an ultrasonic pulse as illustrated in Fig. 1a. The known speeds of the RF signal and the ultrasonic signal as well as their TDoA are used to calculate the distance between the listener and each beacon.



a Different kinds of signals are emitted simultaneously

b Same signals are emitted from different speakers at a known time schedule

Fig. 1 Two different cases of TDoA

The ID information of each beacon is decoded in their RF signals so the listener can distinguish which beacon corresponds to the calculated distance. The work of Cricket has inspired follow-up research on acoustic localization. Due to the fact that current smart devices are not equipped with ultrasonic sensors, many subsequent works choose inaudible signals among the spectrum of 17–24 kHz considering the sampling rate provided by current smart devices.

In [75], an in-car system is designed to locate the driver's phone by four stereo speakers. When the system is triggered by an incoming phone call, the phone will send an indication to the stereo system through Bluetooth. After the stereo system receives the indication, speakers will transmit high-frequency beeps at a fixed interval as illustrated in Fig. 1b, which are later recorded by the phone. To decide the relative position of the phone between two speakers i and j, the time difference of two speakers emitting signals Δt_{ij} and the time difference of detecting signals $\Delta t'_{ij}$ are compared. If $\Delta t_{ij} > \Delta t'_{ij}$, it means the phone is on the seat next to the speaker j. Otherwise the phone is closer to the speaker i. Since Δt_{ij} is known as a system setting, and $\Delta t'_{ij}$ is calculated by measurements from the phone, the clock unsynchronization issue between the stereo system and the phone is avoided. The main challenge for this system is the heavy multipath environment in the car. The correlation and peak detection method is no longer applicable. In [75], the first sample of a beep is identified by change-point detection. The likelihood that the ith sample X_i comes from a specific beep is defined by $l(X_i) = X_i - \mu$ where μ is the mean value of all the samples from that beep. Next, an evaluation metric $s_i = max\{s_{i-1} + l(X_i), 0\}$ is defined and $s_0 = 0$. The first sample whose s_i exceeds a predefined threshold will be regarded as the starting sample of that beep.

The configuration of infrastructure in [32] is similar to that of Cricket. Multiple speakers are already in place for locating a moving device. In Cricket, the ID information of each anchor is decoded in their RF signals, while rate-adaptive chirps are designed in [32] to distinguish signals from different anchor speakers. Another different point between Cricket and [32] is that the distance to an anchor node is not directly calculated through TDoA in [32]. On the other hand, the coordinate of the target is calculated through a system of equations. Suppose we want to calculate the 3D coordinate of the target, then we need 4 speakers whose coordinates are known, which are denoted by $(x_i, y_i, z_i), i = 1, 2, 3, 4$. Each speaker simultaneously transmits a unique chirp, and they are received by the receiver one after another. If the TDoA of signals from speaker i and j is denoted by T_{ij}, then

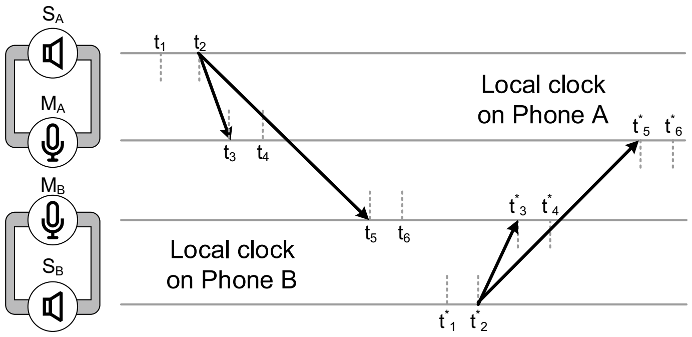


Fig. 2 Measuring the phone-to-phone distance using ETOA

$c \times T_{ij} = \sqrt{(x_i - x)^2 + (y_i - y)^2 + (z_i - z)^2} - \sqrt{(x_j - x)^2 + (y_j - y)^2 + (z_j - z)^2}$ where i and j range from 1 to 4 and they can not be identical. The coordinate of the target (x, y, z) is obtained by solving those equations.

While the TDoA between a RF signal and an ultrasonic signal is applied in Cricket, and the TDoAs among multiple speakers are leveraged in [32, 75], AMIL [21] on the other hand measures the TDoAs of consecutive beeps emitted by a moving transmitter at a predefined pattern. TDoA is actually used for improving accuracy in AMIL. Basically, it applies inertial sensors to track the transmitter's coordinate, which is known to be unprecise because of the double integration. Supposing that the transmitter emits n beeps, its coordinate when emitting the ith beep is (x_i, y_i). (x_0, y_0) is set to be the origin. The displacement between the transmitter and a passive listener from the first beep to the ith beep is $dd_i = \sqrt{(x - x_i)^2 + (y - y_i)^2} - \sqrt{x^2 + y^2}, i = 2, 3, \ldots, n$ where (x, y) is the listener's coordinate. If we denote the time interval between emitting the first beep and the ith beep by Δt and the interval between receiving them by ΔT, we have $dd_i = c(\Delta T - \Delta t)$. $n - 2$ linear equations are thus obtained and the least squares method can be employed to find an approximate solution. To reduce computation overhead, AMIL picks 3 from the n beeps to calculate the coordinate of the listener.

As a conclusion, the clock unsynchronization issue between the broadcasting system and the receiver is avoided in [32, 75], but all the speakers are assumed to share the same local clock. For Cricket, the RF signal and the ultrasonic signal are emitted simultaneously from the same beacon and received by the listener. The internal time systems of both sides are required to be highly consistent when processing the RF signal and the ultrasonic signal. AMIL successfully avoids time subtraction with different local clocks in a novel way, but it does not show advantage over time latency. Since at least 3 beeps are required to locate a 2D position and each beep lasts 50 ms, the total duration adds up to 150 ms. The final time latency can be even higher with the computation time.

One-way Time of Flight

Elapsed time between two time-of-arrivals

BeepBeep [50] defines a new concept as *Elapsed Time between the two time-of-arrivals* (ETOA), which is widely applied to acquire device-to-device (D2D) distances later. The core of BeepBeep can be summarized as two-way sensing, self-recording and sample counting.

BeepBeep demands both devices to be equipped with one speaker and one microphone. Device A and device B emit beeps in turn as shown in Fig. 2. For t_1, t_2, t_3, t_4, t_5, t_6, they are the time A emits its signal, the time A records the emission, the time A receives its own signal, the time A records the reception, the time B receives the signal from A and the time B records its reception of the signal from A. t_1^*, t_2^*, t_3^*, t_4^*, t_5^* and t_6^* are similar notations with respect to the signal transmitted from device B.

If we denote the speaker as S and microphone as M, we have $d\{S_A, M_A\} = c \times (t_3 - t_2)$, $d\{S_A, M_B\} = c \times (t_5 - t_2)$, $d\{S_B, M_B\} = c \times (t_3^* - t_2^*)$, $d\{S_B, M_A\} = c \times (t_5^* - t_2^*)$ where $d\{S_A, M_A\}$ is the distance between S_A and M_A. So are the other three notations. Since the two signals are emitted within one second, the distance R can be approximated by $R = \frac{1}{2}(d\{S_A, M_B\} + d\{S_B, M_A\}) = \frac{1}{2}(c(t_5 - t_2) + c(t_5^* - t_2^*)) = \frac{1}{2}(c(t_5^* - t_3) - c(t_4^* - -t_5)) + \frac{1}{2}d\{S_A, M_A\} + \frac{1}{2}d\{S_B, M_A\}$. As we can see, the last two terms are constants. As for the first two terms, each pair of values is obtained from the same clock and no asynchrony issue will happen. The sum of the first two terms is ETOA. The idea BeepBeep proposed has influenced many subsequent research works, not only self-recording which expedites relevant 3D localization and multiuser localization, but also sample recording which replaces the conventional time measurement and becomes widespread in all the succeeding ToF-based models.

In [55], a 3D localization system is designed on the basis of ETOA. After attaining the D2D distance through ETOA, three additional procedures are done to achieve 3D localization. First, two microphones are required for each device to obtain angle information with the assistance of the law of cosines. Second, a lookup table between the power and the angle is established as cues for angle estimation. Third, a rotation matrix is derived from the data collected by the accelerometer and the digital compass. All the information is fed into Extended Kalman Filter (EKF) to calculate the 3D coordinate of the target.

Tracko [29] is another 3D localization system applying ETOA as well as Kalman filter. In the first step, Bluetooth low energy (BLE) is utilized to detect the presence of other devices. The received signal strength of BLE is also applied to calculate a rough range. Next, the ETOA of acoustic signals generate a more accurate range. Kalman filter combines the rough range from BLE and the relatively accurate range from acoustic signals to determine a 3D position, which is later adjusted by Inertial Measurement Unit (IMU) sensors.

FAR [82] modifies the way BeepBeep detects the arriving time of a signal and achieves lower latency. BeepBeep detects the arriving signal by correlating the recorded profile with the original signal and selects the peak. A autocorrelation is performed on the whole area before cross-correlation on a narrow window which is centered around the autocorrelation peak. The autocorrelation is between a time-domain sample sequence $X = \{p[i](i, = 1, \ldots, n)\}$ and its copied version Y which is delayed by $\frac{n}{2}$. Note that devices emit two identical subsequences in each tone, so X actually consists of two identical parts of which the length is $\frac{n}{2}$. FAR correlates the second part of X with the first part of Y. The key is that we can repeatedly apply the cross-correlation result in ith step to the $(i + 1)$-step, through which the time complexity is reduced from $O(n^2)$ to $O(n)$. In the end, FAR smoothes the proximity area around the autocorrelation peak of $X * Y$ and applies cross-correlation to this narrow area.

Localization can also be applied to D2D file sharing. In [56], a polar coordinate based graph is constructed for each device based on inter-device distances measured by ETOA and angles returned by compasses. The relative position map of all the devices will be depicted on the screen of each involved device. People can share a file to a partner by dragging the file to the icon of the partner's device reflected in the map.

Ping-Pong [24] is a multiuser localization system which combines ETOA and optimization techniques. After calculating pairwise distances through ETOA, it introduces an arbitrary origin to construct a coordinate system and builds an optimization model. If r is a $1 \times N$ matrix where r_{1i} denotes the squared distance between the origin to the device j, R is a $N \times N$ matrix where R_{ij} denotes the pairwise distance between the device i and the device j, Ping-Pong obtains the 3D position matrix X by minimizing $2XX^T = \mathbf{1}r^T + r\mathbf{1}^T - R + \epsilon$ where ϵ is the estimation error. The derivation and details about the objective function can be found in [73].

Sonoloc [16] is a mobile app which can locate hundreds of devices in a large room. During each location updating round, a set of devices T_1 are randomly selected to emit signals, the unselected devices are called passive devices in this round. Sonoloc applies ETOA to calculate the pairwise disance d_{AB} between the transmitter A and the transmitter B, it calculates the coordinates of A and B through minimizing $\sum_{A,B \in T_1} (d_{AB} - \|S_A - M_B\|)^2$ where S_A is the coordinate of A's speaker and M_B is the coordinate of B's microphone. As for the coordinate of a passive device C, Sonoloc calculates the distance difference of C to A and B by TDOA, which is denoted by Δd. The same value Δd can also be inferred through $|\|S_A - M_C\| - \|S_B - M_C\||$. By minimizing the difference between the calculated value and the inferred value, the coordinate of M_C can be obtained.

ETOA is also prevalent when it comes to the initial position acquisition of relative range based localization. For example, Montage [80] and TUM [74] both rely on ETOA to estimate the initial position, while Montage leverages inertial sensors and TUM utilizes particle filter to track continuous motion.

Symbol-based time estimation

One-way flight with an anchor network is common for in-building human tracking [27, 38–40, 46, 81]. ETOA is also one-way flight but its equipment requirement makes it more like a relative ranging method between two mobile devices. The mobile device in this section receives signals from an anchor network and accomplishes self-localization. Anchor speakers achieve synchronization through the wireless communication. Another asynchronization issue is between the mobile and the anchor network.

Guoguo [40] applies a pre-defined symbol sequence. Time of Arrival (ToA) is estimated by detecting the first sample [33] of each symbol. Walsh-Hadamard codes are employed for their othorgonality. Each of the M anchor speakers is assigned a unique sequence consist of L symbols. If the symbol duration is T_s and the guard time between two beacons is T_g, then the round period is $M(LT_s + T_g)$. To increase the update rate, Guoguo devises a symbol-interleaved structure. The frame is divided into symbols. Each beacon transmit one symbol at a time and in turn. No guard time is set. The round duration is reduced to MT_s. To identify the beacon, Guoguo keeps a L length pipeline and iteratively performs the code matching. The synchronizaton round at the beginning is still $LT_s + T_g$ seconds. It requires all the ML symbols to do code matching and explore the time offset. Moreover, the $L - 1$ symbols should be collected for the first update.

Transforming the time interval into the quantum of symbols and their samples is a very straightforward idea. Its implementation in acoustics is not easy due to the limited bandwidth. Guoguo successfully implements the idea but still occupies the frequency band from 15 kHz to 20 kHz. Next, the accuracy of Guoguo is 0.7029 seconds even with the symbol-interleaved structure, which is only eligible to track a low-moving human. Last but not the least, Guoguo applies statistical approaches to estimate ToAs. It implicitly assumes that the noise follows a distribution.

As a short conclusion, BeepBeep and Guoguo both process signals in time-domain. Time-domain signals have a drawback as it is more susceptible to pollution.

Round-trip Time-of-Flight

Round-Trip Time-of-Flight (RTOF) [9, 10, 44, 46, 65, 84, 85] is a favored device-free choice as Fig. 3 shows. The transmitter emits a distinguishable signal. The microphone captures the echo. The distance is estimated through the time between the transmitted signal and its reflection. Due to the round trip, the signal travels longer and attenuates more severely, which makes RTOF more common in around-device interaction.

Cross-correlation is widely employed [18, 22, 50, 55, 59] to determine the arriving time of a signal. The maximum point of the cross-correlation result with respect to the original signal and the received signal will be chosen as the first sample of the received signal. The difficulty is that cross-correlation is expensive in terms of computation complexity. Suppose the original signal is discretely sampled and represented by n time-domain samples $p[1], p[2], \ldots, p[n]$. The received signal, on the other hand, is represented by samples $q[i] (i = 1, 2, \ldots)$. The cross-correlation result is:

$$p * q[i] = \sum_{j=1}^{n} p^*[j]q[i+j] \tag{1}$$

and $\arg\max_i p * q[i]$ is chosen as the first sample of the echo. Cross-correlation is computationally intensive [82]. For smartphones and smart watches, more efficient methods are in need. FAR [82] designs autocorrelation and the time complexity is decreased from $O(n^2)$ to $O(n)$, which however sacrifices the accuracy. Note that Eq. (1) calculates the cross-correlation in time domain. Later the cross-correlation in frequency domain

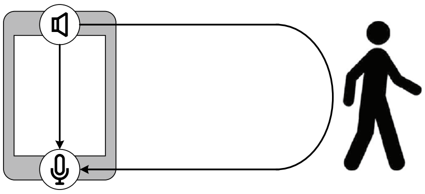


Fig. 3 RTOF

is proved in [19] that it can reduces the time complexity to $O(n \log n)$ while maintaining the accuracy. This cross-correlation substitutes the original cross-correlation and becomes widespread in subsequent RTOF based localization systems.

BatTracker [85] adopts the way introduced in [19] to calculate the distance except it reduces the time duration of the chirp signal from 10 ms to 1 ms for lower time latency. Considering the fact that echoes reflected by objects 5 m away are too weak to be recognized, the time interval between two consecutive pulses is $\frac{5\ m \times 2}{c} \approx 30$ ms. BatTracker further constructs a 3D localization system by picking the top-K strongest peaks of cross-correlation result to establish a reference coordinate.

BatMapper [84] builds digital indoor plans using smartphones with two microphones. The bottom microphone is designed for capturing human voice and the top one is designed for background noise cancellation. Experiments conducted by BatMapper reveal that the bottom microphone is sensitive to lower frequency while the top one has higher noise levels. Considering the heterogeneous characteristics, BatMapper designs a two-pulse signal on the basis of [19]. The first pulse is of higher frequency and longer duration which is suitable for the top microphone. The second pulse is for the bottom microphone with lower frequency and shorter duration. After cross-correlation, the first leak comes from the direct path and is used as the starting point. Due to multipath effect, multiple echoes will be detected apart from the one bouncing off the target. BatMapper picks the top-K strongest peaks and designs a probabilistic evidence accumulation algorithm to map echoes to different reflectors.

EchoTrack [10] is a device-free hand tracking system using smartphones with two speakers and one microphone. The speakers emit chirps in turn and echoes are captured by the microphone. By using cross-correlation introduced in [19], EchoTrack detects echoes from different speakers, calculates the distances to two speakers and obtains the coordinate based on geometry knowledge. The key point of EchoTrack to achieve passive tracking with ToF is the design of its two-channel chirp. At first, the left speaker emits an up-chirp which lasts 1 ms. After an interval of 1 ms, the right speaker emits a down-chirp lasts another 1 ms. The synchronization is achieved through audio module embedded in the phone processor. EchoTack correlates recorded signal with prerecorded up-chirp and down-chirp separately. Left echo and right echo are detected through corresponding cross-correlation. By assigning chirps with different growth trend to the two speakers, those overlapped echoes can be effectively distinguished.

AIM [44] is a smartphone-based acoustic imaging system, but its imaging mechanism also applies echoes of acoustic signals and the way it deals with background noise is enlightening to acoustic localization. The rationale of AIM is based on Synthetic-Aperture Radar (SAR) which is widely used in RF imaging systems [2, 30, 68, 88]. The main idea is moving the transmitter along a distance to simulate a large aperture that helps produce high-resolution images. The user is asked to move his/her phone along a predefined trajectory during which the smartphone periodically emits chirps. After echoes are captured by the smartphone, a 2-stage interference cancellation is applied. Since the speaker and the microphone are omnidirectional, direct path transmission and multipath noises are main interference for AIM. In the first-stage interference cancellation, AIM pre-records direct path transmission in a free space and performs a scaled subtraction to the received superposed signals with the help of Automatic Gain Control (AGC) [5]. In the second stage, AIM applies Least

Square channel estimation [52] to minimize the difference between measured channel and calculated ones $\sum_n (y[n] - \sum_{i\in(U_1\cup U_2)} h_i x[n-i])^2$ where $y[n]$ is the nth received sample, $x[n-i]$ is the $(n-i)$th transmitted sample and $\{h_i\}$ denotes channel taps. U_1 contains all the indices of samples from the target while U_2 contains all the rest. After obtaining $\{h_i\}$ through optimizing the function, the multipath noise and residual direct path interference are removed by $y[n] - \sum_{j\in U_2} h_j x[n-j]$.

ToF is a long-lived method in localization. In this survey, we divide ToF-based localization into TDoA, one-way ToF and RTOF. ToF is still popular nowadays with creative modifications and most of the systems can achieve cm-level accuracy. The main issue challenging ToF-based localization from mm-level accuracy is that, its distance resolution is limited by the sampling rate. The maximum sampling rate a smart device can provide now is 48 kHz. If the sound speed is 343 m/s, then miscounting one sample will lead to the error of 0.71 cm. In practice, it is common to miscount several samples, so the location accuracy is often in the scale of centimeters.

Relative range based localization

Relative range based localization studies the motion of the target. It first obtains the initial location of the target in some way, then updates the location through monitoring the motion of the target. The displacement between two sequential locations is defined as the relative range, in contrary to their respective absolute ranges to pre-defined anchor nodes.

Frequency-modulated continuous wave

Frequency-modulated continuous wave (FMCW) is a common chirp as illustrated in Fig. 4. FMCW is originally introduced in RTOF based localization [9, 84, 85] because it is distinguishable from background noise and has good pulse compressibility [21]. The main feature of FMCW is, its frequency ranges linearly from f_{min} to f_{max} within each period.

Before high-preCision Acoustic Tracker (CAT) [42] proposes *distributed FMCW*, *traditional FMCW* is widely exploited in RF-based localization [1, 3, 6, 17, 28]. *Traditional FMCW* bears a potential peril of the asynchronization between the transmitter and the receiver. Each position update is prescribed within a single sweep of the FMCW. If the starting time stamp of each sweep is normalized as 0 and the current time stamp is t, the transmitted signal $S_{Tx}(t)$ and the received signal $S_{Rx}(t)$ are

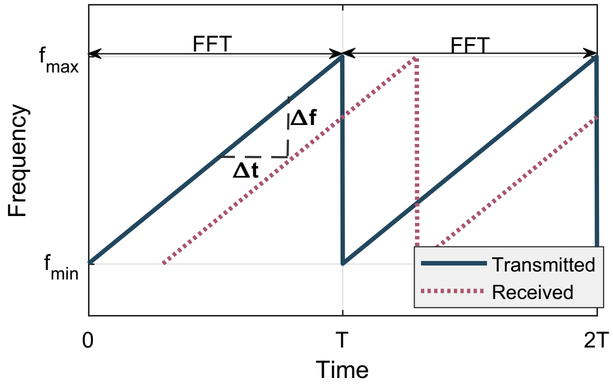


Fig. 4 FMCW

$$
\begin{aligned}
S_{Tx}(t) &= \cos\left(2\pi f_{min} t + \frac{\pi B t^2}{T}\right) \\
S_{Rx}(t) &= \alpha \cos\left(2\pi f_{min}(t - t_d) + \frac{\pi B (t - t_d)^2}{T}\right)
\end{aligned}
$$

After filtering out the high frequency component of their product, the receiver obtains

$$
(S_{Tx} S_{Rx})_{LPF} = \frac{\alpha}{2} \cos\left(2\pi f_{min} t_d + \frac{2\pi B t t_d}{T} - \frac{\pi B t_d^2}{T}\right) \tag{2}
$$

If we plug $t_d = \frac{d+vt}{c}$ into Eq. (2) and ignore terms with $\frac{1}{c^2}$, then $(S_{Tx}S_{Rx})_{LPF}$ can be approximated by

$$
\frac{\alpha}{2} \cos 2\pi \left(\frac{Bv}{cT} t^2 + \frac{f_{min} v}{c} t + \frac{Bd}{cT} t + \frac{f_{min} d}{c}\right)
$$

The frequency of this signal is $f = \frac{Bv}{c} + \frac{Bd}{cT} + \frac{f_{min}v}{c}$ as the mean value of t is $\frac{T}{2}$. When v approaches 0, there will be a peak at $\frac{Bd}{cT}$ in the frequency spectrum. The distance d can thus be estimated by $\frac{cfT}{B}$.

The potential peril is that the construction of Eq. (2) assumes the synchronization between the transmitter and the receiver. To downgrade the peril, CAT [42] introduces a constant time offset as T_0 and then goes through the same deduction of *Traditional FMCW*. The result turns to be $d_n = \frac{cfT}{B} + cT_0$ where n denotes the nth sweep. To eliminate T_0, CAT tracks the target by monitoring its displacement to the initial position.

The rough idea is acquiring the propagation delay. A less sophisticated method could be emitting a highly-compressed pulse and counting the time samples during its travel, which incurs asynchronization and processing delay issue as we introduce in time-based localization. Although CAT dissolves those issues, its ranging resolution is limited by B [81]. The accuracy actually does not show much advantage, that is the reason why CAT collaborates FMCW with IMU sensors and Doppler effect. Rabit [43] further employs the distributed FMCW introduced by CAT to video taping and achieves impressive results.

FMCW is also adopted in ApneaApp [45] and ACG [54]. They are mobile health sensing *apps*. The FMCW is emitted from a mobile device, reflected from the user's body and captured by the same device. Due to the propagation delay, there exist a frequency shift between the emitted signal and the received signal as Fig. 4 shows. The frequency shift can be mapped back to the propagation delay through the constant frequency slope. The frequency shift caused by the chest motion is very minute. If Fast Fourier Transform (FFT) [8] is performed on a single chirp, it is likely that the frequency shift can not be detected in most cases. ApneaApp collects ten chirps and reduces the size of the FFT bin by a factor of 10. This method is applicable as long as the length of ten chirps is smaller than the breathing interval. ApneaApp tracks the chest movement caused by breathing while ACG monitors that of heartbeats. The heartbeat signal is in orders of magnitudes weaker than the breath signal. ACG adopts the FMCW sonar ApneaApp proposes. After the reflected FMCW is captured, ACG first down-converts it to baseband, and then applies ApneaApp's way to select the spatial bin which includes the heartbeat signal. As the heartbeat signal is too feeble, ApneaApp turns to phase measuring in the following steps.

Doppler effect

When there exists a relative motion between the receiver and the transmitter, the received frequency would be different from the transmitted frequency like Fig. 5 demonstrates. Doppler effect [72] quantifies the phenomenon by the following formula:

$$v = \frac{f_{Tx} - f_{Rx}}{f_{Rx}} c \tag{3}$$

The dominant rationale of Doppler effect is the variation of the signal propagation path length [53, 69]. As an alternative way, we can calculate v through Eq. (3) and integrate the result over time to obtain the propagation path length change. With a known initial path length, location information is accessible. This method incurs smaller accumulated error compared to the double integration of accelerations collected by IMU. The key point is converted to how to get a precise and instantaneous f_{Rx}.

According to Nyquist–Shannon sampling theorem, if the sampling rate is at least twice of the maximum frequency, the signal can be recovered through FFT. On the one hand, frequency information is usually resilient to noises. On the other hand, however, the application of FFT incurs the time-frequency resolution problem [12]. The frequency resolution Δf is the ratio of sampling rate F_s over the FFT size. For an analysis window of L time-domain samples, the frequency resolution is:

$$\Delta f = \frac{F_s}{L} = \frac{F_s}{F_s \times t} = \frac{1}{t}$$

where t is the duration of the analysis window. In other words, a smaller frequency resolution contradicts a shorter tracking latency.

AAmouse [76] applies a sampling rate of 44.1 kHz and conducts Short-Term Fourier Transform (STFT) [4] on every 1764 samples. The time latency is 40 ms. The peak frequency from every analysis window will be chosen to calculate the velocity, which will later be considered as the average velocity during this 40 ms and used to calculate the relative range. AAmouse tries to apply zero padding to supply more input data, which actually does not solve the problem. The assumption of the constant velocity during each analysis window will also incur an accumulated error.

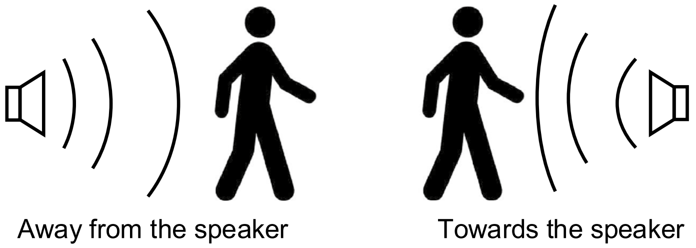


Fig. 5 Doppler effect

Constrained by the time-frequency resolution problem, Doppler effect can only provide a coarse-grained velocity [42, 48, 85] or a direction estimation [7, 11, 20, 66]. Some systems utilize the velocity to improve their tracking accuracy. For example, BatTracker [85] samples signals at a rate of 48 kHz and performs FFT on every 48 samples, which incurs a velocity resolution around 1 m/s. The resulted velocity is used only as a hue to promote the raw tracking prediction made by the IMU. CAT [42] plugs the velocity into an objective function to improve the final accuracy.

Phase shift

Phase-based localization can achieve a balance between high location accuracy and low time latency. Directly measuring the phase of a path is hard [69]. Several methods are introduced in this section to access phase shift. Table 2 shows a comparation between several typical approaches.

To the best of our knowledge, Swadloon [27] is the first acoustic localization system which turns the displacement into phase shift. The transmitter emits the continuous wave. After Band Pass Filter (BPF) and AGC at the receiver side, the signal will be $\cos(2\pi f_{Tx}t + \phi(t))$ where $\phi(t)$ is the phase shift incurred by Doppler effect. The velocity of the mobile device is $v(t) = \frac{c}{2\pi F_{Rx}} \frac{d\phi(t)}{dt}$ and its displacement can be calculated through $\Delta d(t) = \frac{c}{2\pi F_{Rx}}(\phi(t) - \phi(0))$. Swadloon applies Phase Locked Loop (PLL) to calculate $\phi(t)$. PLL is a control system which produces a signal whose phase $\theta(t)$ converges to that of the target signal $\phi(t)$. It consists of phase detector (PD), loop filter (LF) and voltage controlled oscillator (VCO) as illustrated in Fig. 6. PD compares the difference between current generated signal and the target signal. LF is usually a Low Pass Filter (LPF) which filters out high frequency and noise. VCO modulates a new signal according to the result of LPF. The new signal is then fed back to PD. PLL iterates the above steps until the output signal $\theta(t)$ is very similar to $\phi(t)$. The output signal $\theta(t)$ is updated by $\theta_{n+1} = \theta_n + \frac{dJ_{PLL}}{d\theta}$ where $J_{PLL}(\theta) = LPF\{\cos(2\pi F_{Tx}t + \phi(t)) \cos(2\pi F_{Tx}t + \theta(t))\} \approx \frac{1}{2}LPF\{\cos(\phi(t) - \theta(t))\}$. In this way, $\max(J_{PLL}(\theta(t))) = J_{PLL}(\phi(t))$. $\theta(t)$ converges to $\phi(t)$ after enough iterations.

The whole procedure includes BPF, AGC, PLL and linear regression, which incurs large computation overhead. In the experiments, it takes the phone 3.9 seconds to process 1 second of signal samples on average. To achieve realtime tracking, Swadloon lets the phone process 20% of the samples for a trade-off.

FingerIO [46] is a finger tracking system for around-device interaction. Strictly speaking, FingerIO is a ToF-based localization system which applies the normal RTOF procedure. It performs correlation on the received signal with the original signal to identify the timestamp corresponding to the target echo. As we have discussed in "Absolute range based localization" section, this coarse result may contain several samples offset. FingerIO turns to phase shift to fine-tune the result, that is the reason why we put it here for a better comparation. It acheives an impressive result with mm-level accuracy and propels the follow-up research work about phase shift. At the transmitter side, inverse Fast Fourier transform (IFFT) is performed on 64 random bits $\{X_n | n = 1, \ldots, 64\}$ to get 64 time-domain samples $x_k = \sum_{n=0}^{63} X_n e^{\frac{i2\pi kn}{63}}, k = 0, \ldots, 63$. The first 20 samples is appended to the end to form a cyclic suffix. At the receiver side, if the first sample of the echo is correctly decided, FFT can be performed to retrieve the data bits through $X_n = \sum_{n=0}^{63} x_k e^{-\frac{i2\pi kn}{64}}$. If the first sample is mistaken by E samples

Table 2 Comparation between different phase shift based localization systems

	Equipment	Signal	Time latency (ms)	Accuracy
Swadloon [27]	Anchor speakers; smartphones with one microphone and inertial sensors	Continuous wave around 20 kHz	250	The mean error of direction finding is around 2.1° when the range is 32 m, 90-percentile errors are under 0.92 m
FingerIO [46]	Smart devices with one speaker and two microphones	OFDM in the range of 18–20 kHz, the bandwidth of each subcarrier is 375 Hz	5.92	The average accuracy of 2D tracking for smartphones is 8 mm, for smart watches is 1.2 cm
LLAP [34]	Smart devices with one speaker and two microphones	Multiple continuous waves(cosine waves) with frequencies in the range of 17–23 kHz	15	The tracking accuracy for 1D is 3.5 mm, for 2D is 4.6 mm
Strata [77]	Smart devices with one speaker and two microphones	Single-carrier modulated signal with the frequency in the range of 18–22 kHz, 26-bit GSM training sequence	12.5	The distance tracking error is 0.3 cm, the 2D tracking error is 1 cm, the drawing error in 2D is 0.6
Vernier [81]	Two anchor speakers (e.g., speakers on the TVs); smart device with a microphone	Continuous wave (sine wave with the frequency at 20 kHz or 17.5 kHz for 2D)	10	The median error is under 2 mm for 1D when the distance between the speaker and the receiver is 7 m, the median error for 2D is under 4 mm

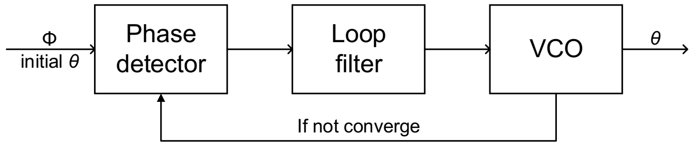


Fig. 6 PLL

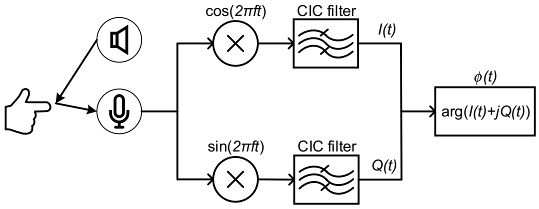


Fig. 7 LLAP

($E < 64$ as the presence of a silence seperation), the corresponding FFT process would be $X_n^E = \sum_{n=0}^{63} x_{k+E} e^{-\frac{i2\pi kn}{64}}$. With the help of the cyclic suffix, $X_n^E = X_n e^{\frac{i2\pi En}{N}}$. The sample offset is converted to phase increment. Since we already know $\{X_n | n = 1, \ldots, N\}$, E can be obtained according to the one-to-one mapping between emitted and received OFDM symbols. After abating the offset, FingerIO reduces the distance error within 1 cm. OFDM actually requires high bandwidth. FingerIO splits 0–24 kHz into 64 subcarriers and sets the audible frequencies to be zero. The bandwidth of each subcarrier is 375 Hz, which is not very robust to frequency offset.

LLAP [70] and PatternListener [86] adopt the coherent detector to obtain a complex signal which can be used to extract the phase information. LLAP is a gesture-tracking system. The transmitted signal is $A\cos(2\pi ft)$, which later gets reflected by the moving finger. If the path length is denoted as $d(t)$, the received signal is $A'\cos(2\pi f(t - \frac{d(t)}{c}) - \theta)$ where θ is caused by hardware delay and phase inversion. LLAP multiplies the received signal by $\cos(2\pi ft)$ and gets $\frac{A'}{2}(\cos(-2\pi f\frac{d(t)}{c} - \theta) + \cos(4\pi ft - 2\pi f\frac{d(t)}{c} - \theta))$. The second term is removed by a low-pass CIC filter. The first term remains and is denoted by $I(t) = \frac{A'}{2}(\cos(-2\pi f\frac{d(t)}{c} - \theta))$. Meanwhile, it multiplies the received signal by $\sin(2\pi ft)$, completes the same procedure above and obtains $Q(t) = \frac{A'}{2}(\sin(-2\pi f\frac{d(t)}{c} - \theta))$. $I(t)$ and $Q(t)$ are the real and imaginary part of the complex signal, and the phase at t can be calculated by $\phi(t) = arctg(\frac{Q(t)}{I(t)})$. The path change between t_1 and t_2 is $d(t_1) - d(t_2) = (\phi(t_1) - \phi(t_2)) \times \frac{\lambda}{2\pi}$. An illustration of the whole process is provided in Fig. 7. PatternListener adopts LLAP's idea and develops an acoustic attack that can crack Android pattern lock. The ideal model is based on the assumption that the received signal only consists of the target echo. Considering multipath effect, the received signal is actually a combination of static-path signals and dynamic-path signals. The dynamic-path signals can be extracted through Local Extreme Value Detection algorithm, which LLAP designs based on Empirical Mode Decomposition algorithm [26]. The dynamic-path signals also contain echoes reflected from other moving surroundings. To solve the issue, multiple frequencies with a fixed interval are emitted simultaneously, different

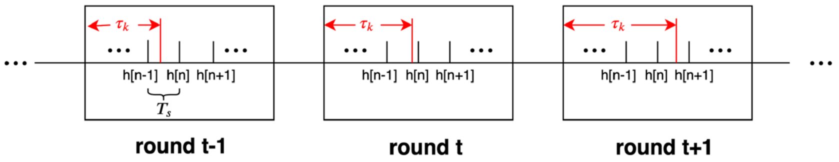


Fig. 8 Strata

outputs are taken in a comprehensive consideration using linear regression and generate the final result.

Channel impulse response (CIR) is also a good choice to get access to phase shift [64, 77]. Suppose the passband signal $x(t)$ is emitted from the speaker, later a superposition of L multiple reflections with different delays $\tau_i (i = 1, \ldots, L)$ and amplitudes a_i is received:

$$y(t) = \sum_{i=1}^{L} a_i x(t - \tau_i) = \sum_{i=1}^{L} a_i e^{-j2\pi f_c \tau_i} s(t - \tau_i) = h(t) * x(t)$$

where f_c and $s(t)$ denote the center frequency of the passband and the baseband signal respectively. $h(t) = \sum_{i=1}^{L} a_i e^{-j2\pi f_c \tau_i} \delta(t - \tau_i)$ is defined as CIR. If CIR is sampled by a time interval of T_s, we can get the nth channel tap $h[n] = \sum_{i=1}^{L} a_i e^{-j2\pi f_c \tau_i} \delta(t - \tau_i) sinc(n - \tau_i W)$. The relationship of consecutive location updating rounds, path delays and channel taps is illustrated in Fig. 8.

Strata [77] combines phase based relative range with channel difference based absolute range. The key finding of Strata is that CIR is influenced by the moving finger. So in each location updating round, Strata adopts a pilot sequence to estimate CIR by Least-Square channel estimation [52]. As soon as CIR is obtained, Strata quickly determines the kth channel tap which is affected by the moving finger in the tth update round. The change of that channel tap from the $(t-1)$th round to the t-round is:

$$\begin{aligned} h_d[k]^t &= h[k]^t - h[k]^{t-1} \\ &= a_{L_k} (e^{-j2\pi f_c (\tau_{L_k}(t-1) + \tau_d(t))} - e^{-j2\pi f_c \tau_{L_k}(t-1)}) \end{aligned} \tag{4}$$

The corresponding phase shift is:

$$\angle(h_d[k]^t) = \angle(e^{-j2\pi f_c \tau_{L_k}(t-1)}) + \frac{\angle(e^{-j2\pi f_c \tau_d(t)})}{2} + \frac{\pi}{2}$$

$\angle(h_d[k]^{t+1})$ is calculated in the same way, and its difference from $\angle(h_d[k]^t)$ is:

$$\angle(e^{-j2\pi f_c \tau_d(t)}) + \frac{\angle(e^{-j2\pi f_c \tau_d(t+1)}) - \angle(e^{-j2\pi f_c \tau_d(t)})}{2}$$

With the assumption that $\tau_d(t+1) = \tau_d(t)$, we can derive $\angle(e^{-j2\pi f_c \tau_d(t)})$ and calculate the relative range through $\Delta d = \frac{\lambda}{2\pi} \angle(e^{-j2\pi f_c \tau_d(t)})$. The absolute range is obtained through minimizing the weighted sum of two differences, one is the difference between measured CIR change and inferred CIR change, the other is the difference between measured delay change and inferred delay change. More details can be found in [77]. The absolute range is introduced only for improving accuracy and calculating the initial location. The final 2D coordinate is acquired with the help of two microphones.

While Strata applies 26-bit GSM pilot sequence due to its good performance on synchronization and channel estimation [52], VSkin [64] employs Zadoff-Chu (ZC) sequence for its low auto-correlation and constant amplitude. VSkin leverages EKF to track channel coefficients and represents the result with I/Q components as $y(t) = (I_{h[n]}, Q_{h[n]})$. The instantaneous curvature of the I/Q trace is $k(t) = \frac{det(y'(t),y''(t))}{\|y'(t)\|^3}$ and the phase shift of the dynamic part during the round $t-1 \sim t$ is:

$$\Delta\theta = 2\arcsin\frac{k(t)|y(t)-y(t-1)|}{2}$$

The relative range is obtained as the way Strata does.

LLAP, Strata and VSkin are all gesture capturing systems. The main difference between LLAP and the two CIR based systems is, LLAP regards the phase shift caused by all the reflectors as an integral after removing possible noises, while the other two operate on individual paths with different delays. As for Strata and VSkin, the former one is designed for in-air gestures happening around 20 cm away from the screen, while the latter one is designed for gestures performed on the surface of a smart device. When it comes to how to decide the target reflection path, Strata first identifies dynamic paths through channel variation and then selects the one with the smallest maximum phase change. VSkin picks the path with largest magnitude change as the target path.

Vernier [81] is another active localization system which achieves very low time latency. The median time latency is around 10 ms while mm-level location accuracy is kept. Vernier first utilizes the law of cosines to obtain the initial position of a device, then applies phase shift to calculate relative ranges and updates the position. The relative range during T is $\Delta d = \frac{c\Delta\phi}{2\pi F_{Tx}} - cT = (N_{max}\lambda - cT) \pm \lambda$ where N_{max} is the maximum number of cycles the signal contains. The approximation error is within a wavelength if we approximate the original formula with $N_{max}\lambda - cT$. If we consider p circles with q samples, where p and q are two integers which make the smallest integer $\frac{p}{q} = \frac{F_{Tx}}{F_s}$, it is proved in the article that the relative phases of q samples are uniformly distributed in $[0, 2\pi]$. For two analysis windows of q samples, the phase shift between their corresponding first samples is an integral multiple of $\frac{2\pi}{q}$. This integer can be obtained through counting the number of local maxima. Vernier achieves an attractive result on time latency, but the experiment setting is kind of fragile since the device moves in a relatively small area.

Future work

Currently the fundamental issue which limits acoustic localization from widespread employment is the severe attenuation of aerial acoustic signals [40]. As we can see in "One-way Time of Flight" section, most RTOF based systems track hand or finger rather than device or human, this is because echoes bouncing off reflectors which are 5–6 m away are too weak to be observed. For remote device or human tracking, more anchor speakers are in need. On the other hand, acoustic localization has irreplaceable qualities as low demand for additional infrastructure to COTS devices, high location accuracy achievement and privacy-preserving property. For further exploitation, we consider the following four aspects of high research potential.

- New patterns of acoustic signals: The emitted signal can significantly affect the performance of acoustic localization. A good signal design is expected to satisfy the three requirements: (1) The signal is supposed to be distinguishable from background noise. (2) The signal works for COTS smart devices. (3) The signal should not disturb human's normal life and incur health damage. Chirps and continuous waves are frequently employed signals. Recently, OFDM and other modulated signals are also borrowed from wireless communication technologies and achieve surprising results.
- Keep up with the development of COTS devices: We must admit that the development of smart devices makes contribution to more advanced localization systems. Back in 2000, Cricket needs specially-designed beacons to locate a mobile device equipped with an API. Each Cricket beacon costs less than 10 dollars but several beacons are required to be deployed on the ceil of a room. EchoTrack [10] in 2017 are already using smartphones with two speakers which are designed to provide users with stereo playback, while BatMapper [84] in the same year makes use of smartphones with one speaker but two microphones for better voice capturing. When it comes to future work, we believe manufacturers will ameliorate hardware capacities, which provides acoustic localization with new possibility.
- Excavation of new observations and phenomena: New observations and phenomena are crucial for breaking current limitations and prompting new research directions. Take Doppler effect for example, it was only introduced to acoustic localization several years ago. Recently an uncared-for observation draws attention as the nonlinearity of the microphone hardware. The nonlinearity of electric components were found producing new frequencies [25]. This property is now employed to make ultrasonics audible to COTS smart devices [57, 58, 79], which may break the dilemma of enabling ultrasonic localization on smart devices.
- Inter-field technologies: Traditionally, triangular geometry is widely applied to compute locations with obtained geographic data. Recent years have witnessed a trend towards the utilization of probability models and optimization techniques. We consider future collaboration with machine learning and deep learning as prospective. For a start, RF-finger [67] tests convolutional neural network (CNN) on RF based gesture recognition and achieves good performance. WordRecorder [15] combines acoustic signals with CNN to recognize handwriting. Long short-term memory (LSTM) neural network is leveraged in [37] for mobile devices tracking.

The demand of localization is getting higher and higher. Body tracking systems are expected to track users without asking them to wear additional devices or completing specific motions. Finger tracking systems are supposed to work when there is an occlusion between the hand and the screen. High accuracy and low time latency are no longer the only expectations.

Conclusion

We present a comprehensive survey for range-based acoustic indoor localization. We divide range-based localization into absolute range based localization and relative range based localization. While absolute range based localization mainly employs ToF, relative range based localization utilizes FMCW, Doppler effect and phase shift. We further analyze the techniques leveraged in the subcategories and show their unique characteristics as well as limitations. Potential research directions are also provided for future study.

Acknowledgements
No acknowledgements.

Authors' contributions
ML writes the survey and other co-authors provide opinions and guidance. All authors read and approved the final manuscript.

Funding
No funding.

Availability of data and materials
No dataset is used for experiment in this survey.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Computer Science and Engineering Department, Michigan State University, 428 South Shaw Lane, East Lansing 48824, USA. [2] School of Software, Tsinghua University, 30 Shuangqing Rd, 100084 Beijing, China. [3] School of Computer and Communication Engineering, Changsha University of Science & Technology, 960 2nd Section, Wanjiali South Road, 410114 Changsha, China.

Received: 3 September 2019 Accepted: 27 December 2019
Published online: 06 January 2020

References

1. Adib F, Kabelac Z, Katabi D, Miller RC (2014) 3d tracking via body radio reflections. In: Proceedings of USENIX NSDI. Berkeley, USENIX Association, pp 317–329
2. Adib F, Hsu CY, Mao H, Katabi D, Durand F (2015) Capturing the human figure through a wall. ACM Trans Graph 34:219:1–219:13
3. Al-Qudsi B, El-Shennawy M, Wu Y, Joram N, Ellinger F (2015) A hybrid tdoa/rssi model for mitigating nlos errors in fmcw based indoor positioning systems. In: Proceedings of IEEE PRIME. New Jersey, IEEE, pp 93–96
4. Allen J (1977) Short term spectral analysis, synthesis, and modification by discrete fourier transform. IEEE Trans Acoustics Speech Signal Process 25:235–238
5. AndroidDev (2019) Automatic gain control. https://developer.android.com/reference/android/media/audiofx/AutomaticGainControl
6. Ash M, Ritchie M, Chetty K, Brennan PV (2015) A new multistatic fmcw radar architecture by over-the-air deramping. IEEE Sens J 15:7045–7053
7. Aumi MTI, Gupta S, Goel M, Larson E, Patel S (2013) Doplink: using the doppler effect for multi-device interaction. In: Proceedings of ACM UbiComp. New York, ACM, pp 583–586
8. Bergland GD (1969) A guided tour of the fast Fourier transform. IEEE Spectrum 6:41–52
9. Chen H, Li F, Wang Y (2016) Echoloc: Accurate device-free hand localization using cots devices. In: Proceedings of IEEE ICPP. New Jersey, IEEE, pp 334–339
10. Chen H, Li F, Wang Y (2017) Echotrack: Acoustic device-free hand tracking on smart phones. In: Proceedings of IEEE INFOCOM. New Jersey, IEEE, pp 1–9
11. Chen KY, Ashbrook D, Goel M, Lee SH, Patel S (2014) Airlink: Sharing files between multiple devices using in-air gestures. In: Proceedings of ACM UbiComp. New York, ACM, pp 565–569
12. Claerbout JF (1992) Earth soundings analysis: processing versus inversion. Blackwell Scientific Publications, London
13. Ding H, Han J, Qian C, Xiao F, Wang G, Yang N, Xi W, Xiao J (2018) Trio: utilizing tag interference for refined localization of passive rfid. In: Proceedings of IEEE INFOCOM. New Jersey, IEEE, pp 828–836
14. Dong Y, Hoover A, Scisco J, Muth E (2012) A new method for measuring meal intake in humans via automated wrist motion tracking. Appl Psychophysiol Biofeedback 37:205–215
15. Du H, Li P, Zhou H, Gong W, Luo G, Yang P (2018) Wordrecorder: Accurate acoustic-based handwriting recognition using deep learning. In: Proceedings of IEEE INFOCOM. New Jersey, IEEE, pp 1448–1456
16. Erdélyi V, Le TK, Bhattacharjee B, Druschel P, Ono N (2018) Sonoloc: Scalable positioning of commodity mobile devices. In: Proceedings of ACM MobiSys. New York, ACM, pp 136–149
17. Gierlich R, Huttner J, Dabek A, Huemer M (2007) Performance analysis of fmcw synchronization techniques for indoor radiolocation. 2007 European conference on wireless technologies. IEEE, New Jersey, pp 24–27

18. Girod L, Lukac M, Trifa V, Estrin D (2006) The design and implementation of a self-calibrating distributed acoustic sensing platform. In: Proceedings of ACM SenSys. New York, ACM, pp 71–84
19. Graham D, Simmons G, Nguyen DT, Zhou G (2015) A software-based sonar ranging sensor for smart phones. IEEE Internet Things J 2:479–489
20. Gupta S, Morris D, Patel S, Tan D (2012) Soundwave: Using the Doppler effect to sense gestures. In: Proceedings of ACM CHI. New York, ACM, pp 1911–1914
21. Han H, Yi S, Li Q, Shen G, Liu Y, Novak E (2016) Amil: Localizing neighboring mobile devices through a simple gesture. In: Proceedings of IEEE INFOCOM. New Jersey, IEEE, pp 1–9
22. Hazas M, Hopper A (2006) Broadband ultrasonic location systems for improved indoor positioning. IEEE Trans Mobile Comput 5:536–547
23. He T, Huang C, Blum BM, Stankovic JA, Abdelzaher T (2003) Range-free localization schemes for large scale sensor networks. In: Proceedings of ACM MobiCom. New York, ACM, pp 81–95
24. Herrera J, Kim HS (2013) Ping-pong: using smartphones to measure distances and relative positions. Proc Meetings Acoustics 20(055):003
25. Horowitz P, Hill W (1989) The art of electronics. Cambridge University Press, New York
26. Huang NE, Shen Z, Long SR, Wu MC, Shih HH, Zheng Q, Yen NC, Tung CC, Liu HH (1998) The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. Proc R Soc Lond A 454:903–995
27. Huang W, Xiong Y, Li XY, Lin H, Mao X, Yang P, Liu Y (2013) Accurate indoor localization using acoustic direction finding via smart phones. CoRR. arXiv:1306.1651
28. Huang Y, Brennan PV, Seeds A (2008) Active rfid location system based on time-difference measurement using a linear fm chirp tag signal. In: Proceedings of IEEE PIMRC. New Jersey, IEEE, pp 1–5
29. Jin H, Holz C, Hornbæk K (2015) Tracko: Ad-hoc mobile 3d tracking using bluetooth low energy and inaudible signals for cross-device interaction. In: Proceedings of ACM UIST. New York, ACM, pp 147–156
30. Karanam CR, Mostofi Y (2017) 3d through-wall imaging with unmanned aerial vehicles using wifi. In: Proceedings of ACM/IEEE IPSN. New Jersey, IEEE, pp 131–142
31. Kaune R (2012) Accuracy studies for tdoa and toa localization. In: 2012 15th International conference on information fusion. IEEE, New Jersey, pp 408–415
32. Lazik P, Rowe A (2012) Indoor pseudo-ranging of mobile devices using ultrasonic chirps. In: Proceedings of ACM SenSys. New York, ACM, pp 99–112
33. Lee JY, Scholtz RA (2002) Ranging in a dense multipath environment using an uwb radio link. IEEE J Select Areas Commun 20:1677–1683
34. Li F, Zhao C, Ding G, Gong J, Liu C, Zhao F (2012) A reliable and accurate indoor localization method using phone inertial sensors. In: Proceedings of ACM UbiComp. New York, ACM, pp 421–430
35. Li T, Chen Y, Zhang R, Zhang Y, Hedgpeth T (2018) Secure crowdsourced indoor positioning systems. In: Proceedings of IEEE INFOCOM. New Jersey, IEEE, pp 1034–1042
36. Lien J, Gillian N, Karagozler ME, Amihood P, Schwesig C, Olson E, Raja H, Poupyrev I (2016) Soli: Ubiquitous gesture sensing with millimeter wave radar. ACM Trans Graphics 35:142:1–142:19
37. Liu H, Li XY, Zhang L, Xie Y, Wu Z, Dai Q, Chen G, Wan CCW (2018) Finding the stars in the fireworks: Deep understanding of motion sensor fingerprint. In: Proceedings of IEEE INFOCOM. New Jersey, IEEE, pp 126–134
38. Liu K, Liu X, Li X (2012) Acoustic ranging and communication via microphone channel. In: Proceedings of IEEE GLOBECOM. New Jersey, IEEE, pp 291–296
39. Liu K, Liu X, Xie L, Li X (2013) Towards accurate acoustic localization on a smartphone. In: Proceedings of IEEE INFOCOM. New Jersey, IEEE, pp 495–499
40. Liu K, Liu X, Li X (2016) Guoguo: Enabling fine-grained smartphone localization via acoustic anchors. IEEE Trans Mobile Comput 15:1144–1156
41. Liu Y, Yang Z, Wang X, Jian L (2010) Location, localization, and localizability. J Comput Sci Technol 25:274–297
42. Mao W, He J, Qiu L (2016) Cat: High-precision acoustic motion tracking. In: Proceedings of ACM MobiCom. New York, ACM, pp 69–81
43. Mao W, Zhang Z, Qiu L, He J, Cui Y, Yun S (2017) Indoor follow me drone. In: Proceedings of ACM MobiSys. New York, ACM, pp 345–358
44. Mao W, Wang M, Qiu L (2018) Aim: Acoustic imaging on a mobile. In: Proceedings of ACM MobiSys. New York, ACM, pp 468–481
45. Nandakumar R, Gollakota S, Watson N (2015) Contactless sleep apnea detection on smartphones. In: Proceedings of ACM MobiSys. New York, ACM, pp 45–57
46. Nandakumar R, Iyer V, Tan D, Gollakota S (2016) Fingerio: Using active sonar for fine-grained finger tracking. In: Proceedings of ACM CHI. New York, ACM, pp 1515–1525
47. Nguyen V, Ibrahim M, Rupavatharam S, Jawahar M, Gruteser M, Howard R (2018) Eyelight: Light-and-shadow-based occupancy estimation and room activity recognition. In: Proceedings of IEEE INFOCOM. New Jersey, IEEE, pp 351–359
48. Nishimura Y, Imai N, Yoshihara K (2012) A proposal on direction estimation between devices using acoustic waves. In: Proceedings of Springer MobiQuitous. Berlin, Springer, pp 25–36
49. Parate A, Chiu MC, Chadowitz C, Ganesan D, Kalogerakis E (2014) Risq: Recognizing smoking gestures with inertial sensors on a wristband. In: Proceedings of ACM MobiSys. New York, ACM, pp 149–161
50. Peng C, Shen G, Zhang Y, Li Y, Tan K (2007) Beepbeep: A high accuracy acoustic ranging system using cots mobile devices. In: Proceedings of ACM SenSys. New York, ACM, pp 1–14
51. Priyantha NB, Chakraborty A, Balakrishnan H (2000) The cricket location-support system. In: Proceedings of ACM MobiCom. New York, ACM, pp 32–43
52. Pukkila M (2000) Channel estimation modeling. Nokia Res Center 17:66
53. Qian K, Wu C, Yang Z, Liu Y, Jamieson K (2017) Widar: Decimeter-level passive tracking via velocity monitoring with commodity wi-fi. In: Proceedings of ACM Mobihoc. New York, ACM

54. Qian K, Wu C, Xiao F, Zheng Y, Zhang Y, Yang Z, Liu Y (2018) Acousticcardiogram: Monitoring heartbeats using acoustic signals on smart devices. In: Proceedings of IEEE INFOCOM. New Jersey, IEEE, pp 1574–1582
55. Qiu J, Chu D, Meng X, Moscibroda T (2011) On the feasibility of real-time phone-to-phone 3d localization. In: Proceedings of ACM SenSys. New York, ACM, pp 190–203
56. Qiu JW, Lo CC, Lin CK, Tseng YC (2014) A d2d relative positioning system on smart devices. In: 2014 IEEE wireless communications and networking conference (WCNC). IEEE, New Jersey, pp 2168–2172
57. Roy N, Hassanieh H, Roy Choudhury R (2017) Backdoor: Making microphones hear inaudible sounds. In: Proceedings of ACM MobiSys. New York, ACM, pp 2–14
58. Roy N, Shen S, Hassanieh H, Choudhury RR (2018) Inaudible voice commands: The long-range attack and defense. In: Proceedings of USENIX NSDI. Berkeley, USENIX Association, pp 547–560
59. Sallai J, Balogh G, Maroti M, Ledeczi A, Kusy B (2004) Acoustic ranging in resource-constrained sensor networks. In: Proceedings of ICWN, p 467
60. Schlömer T, Poppinga B, Henze N, Boll S (2008) Gesture recognition with a wii controller. In: Proceedings of ACM TEI. New York, ACM, pp 11–14
61. Shao S, Khreishah A, Khalil I (2018) Retro: Retroreflector based visible light indoor localization for real-time tracking of iot devices. In: Proceedings of IEEE INFOCOM. New Jersey, IEEE, pp 1025–1033
62. Shen S, Gowda M, Roy Choudhury R (2018) Closing the gaps in inertial motion tracking. Proceedings ACM MobiCom. ACM, New York, pp 429–444
63. Shotton J, Sharp T, Kipman A, Fitzgibbon A, Finocchio M, Blake A, Cook M, Moore R (2011) Real-time human pose recognition in parts from single depth images. In: Proceedings of IEEE CVPR. New Jersey, IEEE, pp 1297–1304
64. Sun K, Zhao T, Wang W, Xie L (2018) Vskin: Sensing touch gestures on surfaces of mobile devices using acoustic signals. In: Proceedings of ACM MobiCom. New York, ACM, pp 591–605
65. Sun Z, Farley R, Kaleas T, Ellis J, Chikkappa K (2011) Cortina: Collaborative context-aware indoor positioning employing rss and rtof techniques. In: IEEE Pervasive computing and communications workshops (PERCOM Workshops). IEEE, New Jersey, pp 340–343
66. Sun Z, Purohit A, Bose R, Zhang P (2013) Spartacus: Spatially-aware interaction for mobile devices through energy-efficient audio sensing. In: Proceeding of ACM MobiSys. ACM, New York, pp 263–276
67. Wang C, Liu J, Chen Y, Liu H, Xie L, Wang W, He B, Lu S (2018) Multi-touch in the air : Device-free finger tracking and gesture recognition via cots rfid. In: Proceedings of IEEE INFOCOM. New Jersey, IEEE, pp 1691–1699
68. Wang J, Xiong J, Chen X, Jiang H, Balan RK, Fang D (2017) Tagscan: Simultaneous target imaging and material identification with commodity rfid devices. In: Proceedings of ACM MobiCom. New York, ACM, pp 288–300
69. Wang W, Liu AX, Shahzad M, Ling K, Lu S (2015) Understanding and modeling of wifi signal based human activity recognition. In: Proceedings of the 21st annual international conference on mobile computing and networking. New York, ACM, pp 65–76
70. Wang W, Liu AX, Sun K (2016) Device-free gesture tracking using acoustic signals. In: Proceedings of ACM MobiCom. New York, ACM, pp 82–94
71. Weichert F, Bachmann D, Rudak B, Fisseler D (2013) Analysis of the accuracy and robustness of the leap motion controller. Sensors 13:6380–6393
72. White D (1982) Johann christian doppler and his effect? A brief history. Ultrasound Med Biol 8:583–591
73. Wilson RS, Walters JH, Abel JS (2004) Speaker array calibration using inter-speaker range measurements. In: Audio Engineering Society Convention 116, Audio Engineering Society, pp 1–8
74. Xu H, Yang Z, Zhou Z, Yi K, Peng C (2017) Tum: Towards ubiquitous multi-device localization for cross-device interaction. In: Proceedings of IEEE INFOCOM. New Jersey, IEEE, pp 1–9
75. Yang J, Sidhom S, Chandrasekaran G, Vu T, Liu H, Cecan N, Chen Y, Gruteser M, Martin RP (2011) Detecting driver phone use leveraging car speakers. In: Proceedings of ACM MobiCom. New York, ACM, pp 97–108
76. Yun S, Chen YC, Qiu L (2015) Turning a mobile device into a mouse in the air. In: Proceedings of ACM MobiSys. New York, ACM, pp 15–29
77. Yun S, Chen YC, Zheng H, Qiu L, Mao W (2017) Strata: Fine-grained acoustic-based device-free tracking. In: Proceedings of ACM MobiSys. New York, ACM, pp 15–28
78. Zafari F, Gkelias A, Leung KK (2019) A survey of indoor localization systems and technologies. IEEE Commun Surv Tutorials 21:2568–2599
79. Zhang G, Yan C, Ji X, Zhang T, Zhang T, Xu W (2017) Dolphinattack: Inaudible voice commands. In: Proceedings of ACM CCS. New York, ACM, pp 103–117
80. Zhang L, Liu K, Jiang Y, Li XY, Liu Y, Yang P, Li Z (2014) Montage: Combine frames with movement continuity for realtime multi-user tracking. In: Proceedings of IEEE INFOCOM. New Jersey, IEEE, pp 799–807
81. Zhang Y, Wang J, Wang W, Wang Z, Liu Y (2018) Vernier: Accurate and fast acoustic motion tracking using mobile devices. In: Proceedings of IEEE INFOCOM. New Jersey, IEEE, pp 1709–1717
82. Zhang Z, Chu D, Chen X, Moscibroda T (2012) Swordfight: Enabling a new class of phone-to-phone action games on commodity phones. In: Proceedings of ACM MobiSys. New York, ACM, pp 1–14
83. Zhao T, Liu J, Wang Y, Liu H, Chen Y (2018) Ppg-based finger-level gesture recognition leveraging wearables. In: Proceedings of IEEE INFOCOM. New Jersey, IEEE, pp 1457–1465
84. Zhou B, Elbadry M, Gao R, Ye F (2017a) Batmapper: Acoustic sensing based indoor floor plan construction using smartphones. In: Proceedings of ACM MobiSys. New York, ACM, pp 42–55
85. Zhou B, Elbadry M, Gao R, Ye F (2017b) Battracker: High precision infrastructure-free mobile device tracking in indoor environments. In: Proceedings of ACM SenSys. New York, ACM, pp 13:1–13:14
86. Zhou M, Wang Q, Yang J, Li Q, Xiao F, Wang Z, Chen X (2018) Patternlistener: Cracking android pattern lock using acoustic signals. In: Proceedings of the 2018 ACM SIGSAC conference on computer and communications security. New York, ACM, pp 1775–1787
87. Zhou P, Li M, Shen G (2014) Use it free: Instantly knowing your phone attitude. In: Proceedings of ACM MobiCom. New York, ACM, pp 605–616

88. Zhu Y, Zhu Y, Zhao BY, Zheng H (2015) Reusing 60ghz radios for mobile radar imaging. In: Proceedings of ACM MobiCom. New York, ACM, pp 103–116

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.