SyncEcho: Echo-Based Single Speaker Time Offset Estimation for Time-of-Flight Localization

Hiroaki Murakami[1,*], Takuya Sasatani[1], Masanori Sugimoto[2],
Issey Sukeda[1], Yukiya Mita[1], Yoshihiro Kawahara[1,*]
[1]The University of Tokyo, Tokyo, Japan
[2]Hokkaido University, Hokkaido, Japan
[*]{murakami, kawahara}@akg.t.u-tokyo.ac.jp

ABSTRACT

Low-cost and accurate indoor location information can add spatiotemporal context to information systems, enabling new location-aware applications. Time-of-Flight (ToF)-based acoustic localization using speakers and microphones allows for localization accuracy within a few tens of centimeters, outperforming RF-based techniques. However, ToF-based localization requires synchronization between the speaker and microphone, *i.e.*, the time offset between them must be known. Previous time offset estimation methods required custom hardware for speakers, limiting their practical use. Estimating the time offset using a single, unmodified speaker is essential for leveraging widely deployed speakers and enhancing coverage. This paper presents the first method for time offset estimation using a single speaker and a microphone, enabled by two key factors: (i) a time offset computation method that utilizes higher-order floor-ceiling reflections as multiple geometrically-constrained virtual speakers, and (ii) a signal processing pipeline that isolates these critical reflections from numerous others by leveraging the speaker's frequency-dependent radiation pattern. Experiments show that the proposed technique can achieve time offset estimation with a 90th percentile error of 259 μs at a 5 m distance. Furthermore, we implemented a ToF localization system based on SyncEcho, demonstrating a 11.0 cm localization accuracy with a 90th percentile error.

CCS CONCEPTS

• **Networks** → **Location based services**; • **Human-centered computing** → *Ubiquitous computing*.

KEYWORDS

time of flight, ranging, indoor localization, acoustic sensing

ACM Reference Format:
Hiroaki Murakami[1,*], Takuya Sasatani[1], Masanori Sugimoto[2], and Issey Sukeda[1], Yukiya Mita[1], Yoshihiro Kawahara[1,*]. 2024. SyncEcho: Echo-Based Single Speaker Time Offset Estimation for Time-of-Flight Localization. In *The 22nd ACM Conference on Embedded Networked Sensor Systems (SenSys '24), November 4–7, 2024, Hangzhou, China.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3666025.3699369

SenSys '24, November 4–7, 2024, Hangzhou, China

ACM ISBN 979-8-4007-0697-4/24/11…$15.00
https://doi.org/10.1145/3666025.3699369

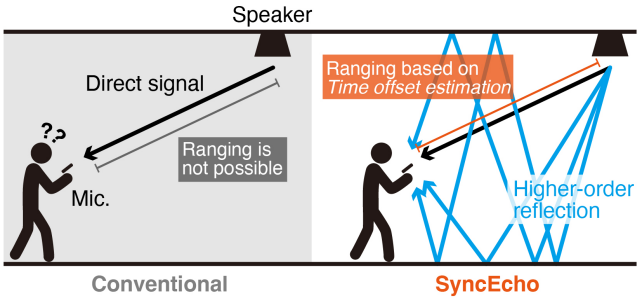


Figure 1: Overview of SyncEcho. SyncEcho uses vertical higher-order reflections of acoustic signals to estimate the time offset between the speaker and a microphone needed for ToF ranging.

1 INTRODUCTION

Indoor location information provides critical context to information systems, and developing low-cost and accurate localization systems using existing infrastructure will accelerate the deployment of ubiquitous computing applications. The widespread use of mobile devices motivates and enables the development of such location-aware applications. One useful example is indoor navigation systems for large and complex buildings (*e.g.*, shopping malls, offices), which cover areas unreachable by outdoor navigation systems. Furthermore, indoor environments produce unique needs, including optimized building energy management [6], home/office automation [34], and empowering new AR/VR applications [40]. Based on these emerging needs, the market value of indoor location information is expected to exceed 29.8 billion dollars by 2028 [22], reinforcing the great demand for indoor positioning technology. However, the complex and enclosed nature of indoor environments prevents one-fits-all solutions; for instance, global navigation satellite systems (GNSS), which cover most cases for outdoor navigation, lack feasibility because GNSS signals do not reach many indoor environments.

These needs and the commodification of mobile and wearable devices led to the investigation of many indoor localization technologies. Existing approaches include, radio frequency (RF)-based approaches such as Wi-Fi [14], Bluetooth low energy (BLE) [13], Ultra wide band (UWB) [10], radio-frequency identification (RFID) [7], visible light-based approaches [16], and acoustic localization [12].

Table 1: Smartphone-based time offset estimation methods for acoustic ranging

Principle	Additional hardware	Multiple speakers	Accuracy [μs]
Dedicated hardware [30]	Modulated LED	None	17.4 (1 m distance, 90th percentile)
RToF-based [33]	Microphone	None	1.9 (3 m distance, average)
TDoA-based [17]	None	Four	720 (4.5 m × 5.5 m area, average)
SyncEcho	None	None	259 (5 m distance, 90th percentile)

Among them, our work is based on acoustic localization using speakers, which has two major advantages:

(1) **Ubiquity and compatibility:** Announcement and music speakers are commonly found in indoor commercial and residential spaces. Additionally, most smartphones have microphones. This infrastructure allows for indoor positioning without the need for dedicated hardware.
(2) **Accurate ranging with generic hardware:** The relatively slow speed of sound waves compared to RF signals enables accurate position estimation based on the time of arrival, using generic hardware.

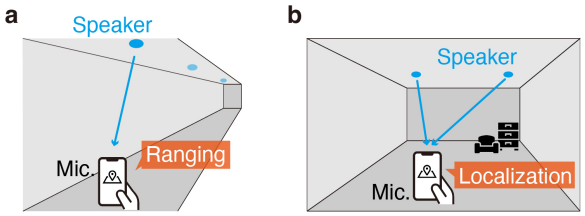


Figure 2: Example usage scenarios of SyncEcho. (a) The system can detect position in narrow environments (*e.g.*, corridors and train platforms) based on ToF ranging. (b) SyncEcho can also detect 2-D or 3-D positions using multiple speakers.

1.1 Challenge: Time Offset Estimation

In acoustic localization, the time of flight (ToF)—the time difference between the speaker emitting sound and the microphone receiving it—is preferred for its high-accuracy ranging capability. Furthermore, unlike the time difference of arrival (TDoA) [12, 18, 35], which requires constant access to multiple speakers to gain any location information, ToF independently performs ranging by each speaker, which is later aggregated to compute location. This feature remarkably improves the robustness of localization because it significantly mitigates the requirements of the speaker-microphone position relationship and, importantly, can allow localization with a minimal number of speakers by leveraging geometrical features inherent to indoor environments (see Fig. 2).

Estimating the time offset between the speakers and microphone accurately, easily, and robustly is one of the most critical challenges when implementing a ToF localization system. Accurate ToF localization requires microsecond-order synchronization between the speakers and the microphone. However, once this time offset is estimated, ToF ranging between all speaker-microphone pairs is enabled, allowing robust and accurate localization. Thus, easily deployable time-offset estimation systems are critical for enhancing the utility of ToF localization.

Currently, time-offset estimation methods, including round-trip time of flight (RToF)-based [33], TDoA-based [17], and dedicated hardware-based approaches [30], have significant limitations. RToF approaches require the installation of speaker-microphone setups that raise privacy concerns, especially in public spaces. TDoA approaches need many speakers, increasing deployment costs. Approaches using dedicated hardware, such as additional LEDs, present cost and complexity challenges. As shown in Table 1, existing methods require additional hardware or multiple speakers, highlighting the significant challenge of developing a microsecond-order time offset estimation using a single, unmodified speaker.

1.2 Our Approach

This paper proposes SyncEcho, the first approach to estimate time offset using a single ceiling speaker. The design of SyncEcho is based on the two following insights. Firstly, reflected signals can be perceived as signals from "mirror" virtual speakers located on the opposite side of the reflector. Higher-order reflections via the floor and ceiling from a ceiling-mounted speaker can be treated as virtual speakers aligned on a line perpendicular to the floor, and the distance between these virtual speakers and the microphone can be formulated by three unknowns, including the time offset. Secondly, considering a ceiling-mounted speaker, the radiation angle of the higher-order reflection signals via the floor and ceiling becomes smaller, mitigating the decay of signal intensity with increased order of reflections. Conversely, the radiation angle of the reflected signals via the wall becomes larger, showing a steep decay of signal intensity with increased order of reflections.

Based on our findings on the radiation angle characteristics of higher-order reflection signals and our observations on the frequency dependency of speaker directivity, SyncEcho identifies first- to third-order reflected signals via the floor and ceiling by transmitting a high-frequency chirp signal from a speaker facing the floor. SyncEcho estimates the time offset using the direct signal and these signals (see Fig. 1). This design makes SyncEcho compatible with widely adopted in-ceiling speakers facing the floor. Such speaker systems are widely used in both commercial and residential spaces for their aesthetic appeal and ability to distribute sound evenly throughout the room [31].

The high-frequency chirp signal that SyncEcho uses is also unobtrusive to most people, ensuring a pleasant acoustic environment. Finally, the estimated time offset remains valid for tens of minutes before the drift error becomes significant; thus, ToF localization can be maintained as long as the user passes through areas where SyncEcho is available at least once during this period.

Our contributions are summarized in the following:

- To the best of our knowledge, **SyncEcho is the first approach that can estimate the time offset using a single, unmodified speaker**. It leverages vertical higher-order reflections via the floor and ceiling to compute the time offset between the speaker and microphone, enabling ToF localization. We assessed its performance under various settings, and the results indicate that SyncEcho achieves a 90th percentile error of 259 μs. Furthermore, **SyncEcho-based ToF localization achieves a 90th percentile error of 11.0 cm.**
- We explore the properties of higher-order reflection signals and the system design that enhances the required signals. Our investigations revealed that utilizing high-frequency signals from speakers facing the floor boosts the SNR of vertical higher-order reflected signals underpinned by the frequency characteristics of the speaker's radiation angle. We verified the consistency and reproducibility of this feature by experimenting with three different speakers.
- SyncEcho identifies vertical higher-order reflected signals using a single microphone. To achieve this, we design a signal processing pipeline utilizing the geometric constraints and successfully identify the needed high-order reflection out of numerous candidate signals.

2 RELATED WORK

2.1 Acoustic Ranging/Localization

Here, we introduce relevant acoustic ranging/localization techniques to highlight the advantages of SyncEcho in terms of usage scenarios. Accurate time offset estimation using smartphones is known to be challenging [17], making ToF-based localization [11, 27] with smartphones difficult. Consequently, most acoustic localization research is oriented toward developing (i) time synchronization-free localization techniques that do not require time offset or (ii) novel techniques for estimating the time offset.

2.1.1 Time Synchronization-Free Localization Techniques. RToF sequentially transmits and receives acoustic signals between nodes and measures the "round trip" time to estimate the distance, requiring each node to have a speaker and a microphone. BeepBeep [25] is a representative example, estimating the distance between two smartphones with an error below 3 cm. This approach has been extensively studied for estimating the distance between devices because it eliminates the influence of time offset [8, 37]. However, its application in position estimation has not progressed significantly because installing speakers equipped with microphones is impractical in public spaces due to privacy concerns. Furthermore, bidirectional communication limits the number of users [4, 10].

Other studies proposed TDoA-based approaches [12, 18, 35], and GPS [24] is one famous example. The TDoA-based approach uses the arrival time difference of acoustic signals transmitted from two speakers to calculate the surface where the target is located. TDoA requires more speakers than ToF for position estimation, but it does not require time synchronization between the speaker and microphone. Furthermore, unlike RToF, it does not necessitate dedicated hardware. As a result, it is currently the most widely applied localization technique for smartphones. However, there are challenges, such as significant positioning errors caused by installing speakers on the ceiling (*i.e.,* at the same height), leading to errors in the height direction. Additionally, acoustic signals are frequently blocked by people and facilities, making TDoA more prone to failure compared to ToF-based localization due to the need for more accessible speakers.

AoA-based approaches use antenna or microphone arrays to estimate the angle of arrival of the transmitted signal [5, 15]. Similar to ToF-based approaches, three anchors are typically required for 3D positioning. As these arrays physically co-exist and are synchronized by hardware, it enables localization without the influence of time offset. However, the high infrastructure requirements pose a barrier to implementation, and the AoA approach is often used for locating unknown sound sources [3, 28] in acoustic sensing.

2.1.2 Techniques for Estimating Time Offset. Next, we introduce approaches for estimating time offset. SCALAR [33] estimates the time offset between two mobile devices by exchanging acoustic signals, similar to BeepBeep. SCALAR achieves a timing accuracy of 1.9 μs by using OFDM signals and considering sampling clock drift. This corresponds to a ranging accuracy of 0.39 mm. Sugimoto *et al.* proposed a camera-based time offset estimation using optimally modulated illumination and achieved a timing accuracy of 17.4 μs at the 90th percentile. SyncSync is a ToF-based localization system that uses Sugimoto *et al.*'s approach. SyncSync uses LED-synchronized speakers as transmitters and a smartphone built-in microphone and camera as a receiver, achieving localization errors within 10 mm. The time offset estimation approach closest to our SyncEcho is Lazik *et al.*'s method [17]. They achieve a mean timing accuracy of 720 μs by using four speakers.

SyncEcho has a clear advantage over relevant approaches in the point that it achieves accurate time offset estimation (259 μs) using a single, unmodified speaker, which are features that no other approach could achieve.

2.2 Multipath-Assisted Positioning

Many multipath-assist sensing techniques have been proposed in prior work. Here, we introduce these technologies to emphasize the technical advancement of SyncEcho in the context of multipath-assisted positioning.

SALMA [9] is a UWB-based positioning system leveraging reflections from four walls. This system uses four directional antennas held by an anchor to identify these reflections. VoLoc [28] and Symphony [34] focus on estimating the position of voice inputs to smart speakers. Given that smart speakers are typically positioned near walls for power supply, they capitalize on the pronounced wall reflections. BatMapper [38] and SAMS [26] transmit acoustic signals from smartphones and receive reflections via walls to compute the distance to these walls and subsequently derive floor plans. BatTracker [39], akin to BatMapper and SAMS, calculates the distance to walls but primarily aims for relative position estimation for tracking. EchoSpot [20] uses speakers and microphones fixed in the environment, achieving device-free positioning by harnessing reflections involving humans and those via both humans and walls.

To the best of our knowledge, SyncEcho is the first approach to harness higher-order reflections, specifically those of the second order and beyond, for positioning. Our methodology ingeniously

capitalizes on the radiation angle characteristics of the speaker, enhancing the SNR of the desired reflections. By exploiting the geometric constraints of virtual speakers and the inherent geometric redundancy, SyncEcho uniquely identifies reflections using a single microphone.

3 PRINCIPLE

3.1 Time Offset Estimation Model Using Vertical Higher-Order Reflection

SyncEcho uses higher-order reflected signals via the floor and the ceiling to estimate δ, the time offset of the speaker relative to the microphone's clock (Fig. 3b), with a single speaker. Here, we describe the principle to calculate the time offset δ. Note that henceforth, n-th-order reflected signals ($n \in \{0, 1, 2, ...\}$) refer to signals propagated via the floor and ceiling n times, and the direct signal is referred to as 0th-order reflected signal.

The reflected signals via the floor and the ceiling can be treated as signals from virtual speakers located at mirror-image positions with respect to the real speaker's location [23] (see Fig. 3a). $P_0 = (x', y', z')$ is the position of the speaker mounted on the ceiling, and $P_n = (x', y', z'_n)$ is the position of the n-th virtual speaker representing the n-th-order reflected signal. Because virtual speakers are at mirror-image positions, the x and y coordinates of all virtual speakers match with the actual speaker. The height of the n-th virtual speaker z'_n is

$$z'_n = (-1)^n z' + \frac{1}{2}\Big((2n-1)(-1)^n + 1\Big)h, \tag{1}$$

where h is the ceiling height, and z' and h are assumed to be known. Note that z' and h are the same value when the speaker is embedded in the ceiling.

Fig. 3b shows the timeline of events. We denote $X = (x, y, z)$ as the position of the microphone, t_n as the time that the signal transmitted from P_n arrives at X, and I as the transmission interval time. The system repeats this timeline as a cycle with duration I. When the signal is transmitted at $t = 0$ on cycle m of the speaker's clock, and their signals are received on cycle m' of the microphone's clock as shown in Fig. 3b, the distance between the n-th virtual speaker and the microphone can be expressed as:

$$||P_n - X|| = \sqrt{d^2 + (z'_n - z)^2} = c(t_n - \delta), \tag{2}$$

where $d = \sqrt{(x'-x)^2 + (y'-y)^2}$ and c is the sound speed. By applying the least-square method to Eq. 3, we can obtain three values, namely δ and d, z.

$$J = \sum_{n=0}^{3} r_n(\delta, d, z)^2, \tag{3}$$

such that

$$r_n(\delta, d, z) = \sqrt{d^2 + (z'_n - z)^2} - c(t_n - \delta). \tag{4}$$

Theoretically, we can calculate δ using up to second-order reflections, but SyncEcho uses up to third-order reflections to locate the pattern of responses attributed to high-order reflections, which is described later in §4.3.

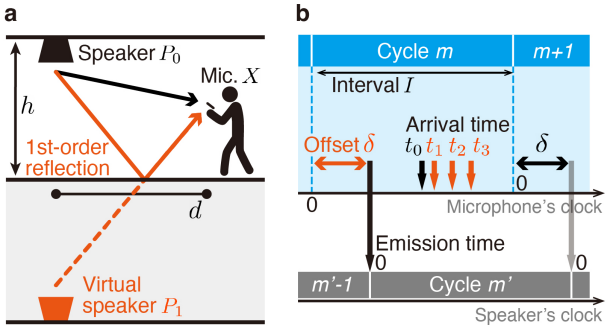


Figure 3: (a) The virtual speaker represents the reflected signal. (b) Overview of the event timeline of SyncEcho and the relationship with the speaker's and the microphone's clock.

3.2 Distinguishing Vertical and Horizontal Reflection based on Speaker Directivity

To estimate time offset based on the described model, it is critical to distinguish the vertical (floor/ceiling) reflections from the horizontal (wall) reflections. To this end, we discovered that the speaker's directivity enhances the vertical higher-order reflections and diminishes horizontal higher-order reflections. SyncEcho leverages this key feature to distinguish vertical and horizontal high-order reflection.

This characteristic can be explained by observing how the signal intensity varies with respect to (i) propagation distance, (ii) number of reflections, and (iii) speakers' radiation angle. Generally, acoustic signals attenuate with (i) increasing distance and (ii) the number of reflections; thus, due to factors (i) and (ii), the signal level decreases as the order of reflection increases. Regarding (iii) speakers' radiation angle, signal intensity decreases as the radiation angle increases for most speakers, as shown in Fig. 4a [19].

Remarkably, this directivity poses different effects on vertical reflections and horizontal reflections. When assuming a ceiling speaker, the radiation angle gets smaller with an increased order of vertical reflections and larger with horizontal reflections, as described in Fig. 4. This factor enhances vertical higher-order reflections and diminishes horizontal higher-order reflections.

Considering factors (i) to (iii), the vertical reflections show a slow decay with increasing order of reflections, while the horizontal reflections show a steep decay. SyncEcho leverages this principle to distinguish vertical and horizontal reflections.

4 SYSTEM DESIGN

Based on the sensing principles described above, we next proceed to the system design of SyncEcho. The design of the transmitter signal, extraction methods of the vertical higher-order reflection signals, the time-offset estimation process, and the sampling clock offset correction are discussed herein.

4.1 Signal Design for Enhancing Vertical Higher-Order Reflection

To accurately detect the propagation time of each signal, we transmit a chirp signal, which is commonly used in acoustic sensing

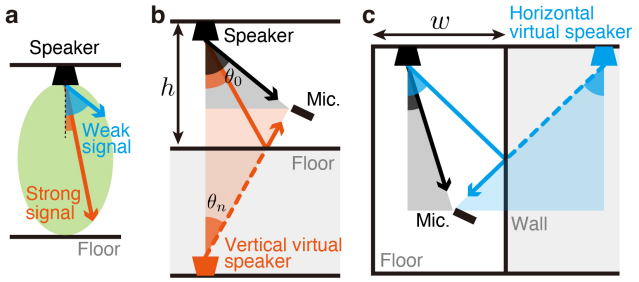


Figure 4: The relationship between the speaker's directivity, reflection order, and radiation angle: (a) Typical speakers emit stronger signals towards smaller angles. (b) Vertical reflections correspond to smaller angles, and (c) horizontal reflections correspond to larger angles.

systems [2, 21, 32]. A linear chirp can be expressed as:

$$s(t) = \sin 2\pi(f_0 t + \frac{k}{2}t^2), \quad (5)$$
$$k = \frac{f_1 - f_0}{T},$$

where f_0 is the start frequency, f_1 is the end frequency, and T is the sweep time from f_0 to f_1 and we used $T = 20$ ms in this work. To suppress clicking noise, known to occur when the speaker output rapidly fluctuates, we also apply a Hann window to the transmitter chirp signal [18].

The time of arrival of chirp signals can be computed by the envelope signal calculated as the cross-correlation between the transmitted signal and the recorded data. Furthermore, we apply a smoothing operation to this envelope to eliminate outliers. Each peak in the smoothed envelope corresponds to the time of arrival of either the direct or reflected signal.

To determine the appropriate frequency range for the chirp signal, we investigated the SNR of the direct and reflected signals in the envelope calculated using three different frequency bands. Because different speakers may have different optimal frequency bands, this experiment was evaluated using three speakers: two tweeters (Fostex FD28D, Fostex PT20K) and one subwoofer (Fostex PW80K). The frequency bands of the transmitted signal were categorized into three groups: low-frequency band (4 kHz - 7 kHz), mid-frequency band (10 kHz - 13 kHz), and high-frequency band (16 kHz - 19 kHz). The experiment setup is the same as the setup later described in §3.2. Fig. 5a to 5c show the envelopes observed using FT28D, PT20K, and PW80K speakers, respectively. Fig. 5a and 5b demonstrate that using a chirp signal with a higher frequency bandwidth as the transmitted signal tends to decrease the SNR of unwanted reflected signals (orange portion) and improve the SNR of vertical n-th-order reflected signals via the floor and ceiling. This trend is also observed in Fig. 5c, where using a chirp signal with a mid-frequency bandwidth decreases the SNR of unwanted reflected signals and improves the SNR of vertical n-th-order reflected signals compared to using a chirp signal with a low-frequency bandwidth. However, it is evident in Fig. 5c that using a chirp signal with a higher frequency bandwidth leads to a decrease in the SNR of vertical second- and third-order reflected signals (green portion).

Fig. 6 compares the envelope when the speaker is facing the floor with the envelope when it is facing the microphone. Remarkably, Fig. 6 shows that when the speaker is facing the floor, the SNR of vertical third-order or higher reflected signal is higher compared to when the speaker is facing the microphone, which also supports our observation that speakers facing the floor enhances vertical higher-order reflections. Taking these results into account, SyncEcho emits a higher frequency chirp signal from a speaker facing the floor to make vertical higher-order reflected signals stand out. Note that the appropriate frequency band also depends on the speaker.

4.2 Extracting Higher-Order Reflections

The design of the transmitted signal allows us to identify large peaks in correlation values as potential candidates for vertical n-th-order reflections. In this paper, we use cell averaging constant false alarm rate (CA-CFAR) [29] to detect peaks in correlation values that are relatively large. Let us denote the time of arrival of the peak detected by CA-CFAR as τ_i ($i \in \{0, 1, 2, ...\}$) (see Fig. 7a). This value represents one of the potential arrival times for the vertical n-th-order reflections.

In SyncEcho, it is necessary to determine the times of arrival of the vertical zeroth- to third-order reflected signals, denoted as t_0 through t_3. Since the direct signal arrives the earliest, we can readily assume τ_0 to be t_0. However, it is challenging to uniquely determine t_1 to t_3 because other τ_i could potentially be the times of arrival of other reflections. In this section, we calculate the potential time windows in which the vertical n-th-order reflected signals can arrive based on geometric considerations, thereby narrowing down the candidate for t_1, t_2, and t_3.

This geometric constraint-based peak filtering reduces the likelihood of incorrect time offset estimates in §4.3 by removing obvious outliers in advance. This approach is particularly effective in scenarios with multiple reflections. Additionally, this pre-filtering improves processing efficiency, reducing the computation time required for time-offset estimation as we will evaluate in §5.2.3.

4.2.1 General Design of Time Window. We calculate the time window for t_1 based on t_0 as the reference time. The difference between t_0 and t_1 increases as the microphone approaches the speaker relative to the xy plane and decreases as it moves away. When the microphone is located directly below the speaker (*i.e.,* when the microphone is aligned in a straight line with the real speaker and the first-order virtual speaker), the time difference of arrival has a maximum value of $(|z_1' - z| - (z_0' - z))/c = 2z/c$ [23]. Therefore, the time window for t_1 is derived as follows:

$$0 < t_1 - t_0 \leq \frac{2z}{c}$$
$$t_0 < t_1 \leq t_0 + \frac{2z}{c}$$

Assuming the height of the microphone, z, to be $z_{min} < z < z_{max}$, the above inequality can be rewritten as:

$$t_0 < t_1 < t_0 + \frac{2z_{max}}{c}. \quad (6)$$

We consider all τ_i that satisfy Eq. 6 as potential candidates for t_1. Let $t_1^{(j)}$ ($j \in \{1, 2, ...\}$) represent one of these candidates. Generally, the height of the microphone is higher than the floor and lower than

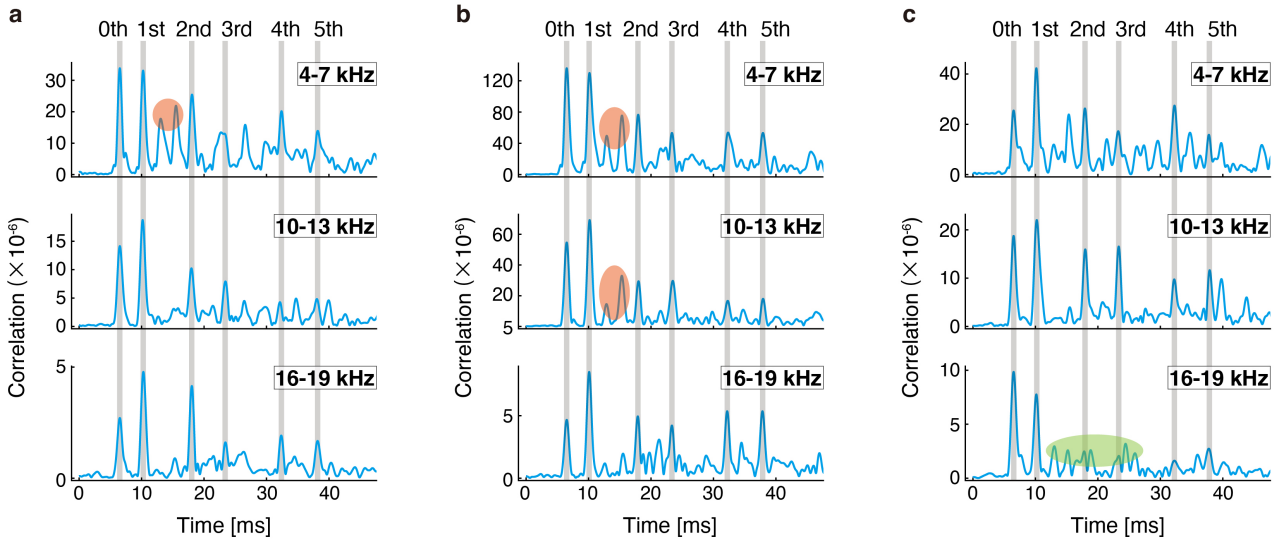


Figure 5: The envelopes of the received signals observed using (a) FT28D, (b) PT20K, and (c) PW80K.

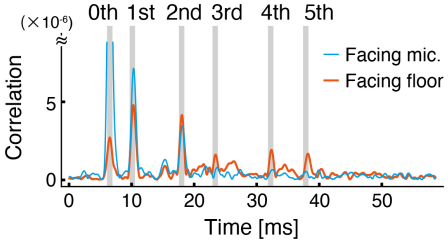


Figure 6: Comparison of the signal envelope received from floor-facing and microphone-facing speakers.

the installed speaker, so we can assume $0 < z < z'$. Furthermore, if the user is holding a smartphone built-in microphone, the range of z can be further restricted (*e.g.*, 0.7 m $< z <$ 1.5 m).

Similar to the time window for t_1, the time windows for t_2 and t_3 are given by:

$$t_0 < t_2 < t_0 + \frac{2h}{c}, \tag{7}$$

$$t_0 < t_3 < t_0 + \frac{2(h + z_{max})}{c}. \tag{8}$$

From the above equations, we can obtain the candidates for t_2 and t_3. However, the time windows for t_2 and t_3 are larger compared to the time window for t_1, resulting in a higher number of candidates for t_2 and t_3. This significantly increases the possibility of selecting incorrect times of arrival when finally determining one each for t_2 and t_3 from the pool of candidates. Therefore, we focus on t_1, which has the smallest number of candidates.

4.2.2 Time Window for t_2. Taking into account the time difference of arrival between the real speaker and the second-order virtual speaker, we can express t_2 as a value determined by d and z as:

$$c(t_2 - t_0) = \sqrt{d^2 + (z_2' - z)^2} - \sqrt{d^2 + (z_0' - z)^2}$$

$$t_2 = \frac{1}{c}\left(\sqrt{d^2 + (z_2' - z)^2} - \sqrt{d^2 + (z_0' - z)^2}\right) + t_0, \tag{9}$$

where z_0', z_2', and t_0 are known values. While there is a range for z, we only have the information that d is greater than or equal to 0, which prevents us from constraining the time window for t_2. Consequently, we consider representing the aforementioned equation as a function of z using t_1.

Assuming any $t_1^{(j)}$ as t_1, the relationship between d and z can be derived from the time difference of arrival between the real speaker and the first-order virtual speaker.

$$c(t_1^{(j)} - t_0) = \sqrt{d^2 + (z_1' - z)^2} - \sqrt{d^2 + (z_0' - z)^2}$$

After simplification, the equation can be rewritten as:

$$d^2 = \alpha z^2 - \beta, \tag{10}$$

where

$$\alpha = \frac{4z'^2 - \gamma^2}{\gamma^2},\ \beta = \frac{4z'^2 - \gamma^2}{4},\ \gamma = c(t_1^{(j)} - t_0).$$

Given that $d^2 = \alpha z^2 - \beta \geq 0$, it follows that $z \leq -\gamma/2, \gamma/2 \leq z$. Therefore, z must satisfy both $z \leq -\gamma/2, \gamma/2 \leq z$ as well as $z_{min} \leq z \leq z_{max}$. This implies $\max(\gamma/2, z_{min}) \leq z \leq z_{max}$. By utilizing Eq. 10, Eq. 9 can be reformulated as a function dependent on z:

$$T_2(z) = \frac{1}{c}\left(\sqrt{\alpha z^2 - \beta + (z_2' - z)^2} - \sqrt{\alpha z^2 - \beta + (z_0' - z)^2}\right) + t_0 \tag{11}$$

Therefore, the time window for t_2 is given by:

$$\min_z T_2(z) < t_2 < \max_z T_2(z) \tag{12}$$

Fig. 7b shows the general shape of the graph of $T_2(z)$, and the specific time windows for $0 \leq z \leq z'$ (under normal conditions) and $0.7 \leq z \leq 1.5$ (assuming the user is standing and holding

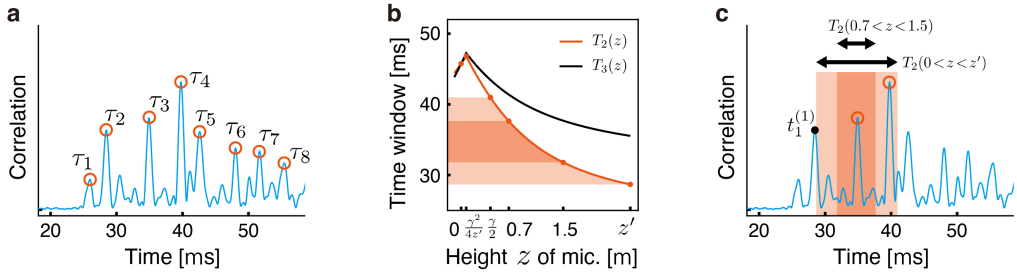


Figure 7: (a) Examples of candidates for vertical reflections detected by CA-CFAR. (b) Overview of how the time windows T_2 and T_3 are narrowed down. (c) Candidates for t_2 computed by the time window based on T_2.

the microphone). Restricting the microphone's height effectively reduces the number of candidates for t_2, as shown in Fig. 7c. We consider all τ_i that satisfy Eq. 12 as potential candidates for t_2. Let $t_2^{(k)}$ ($k \in \{1, 2, ...\}$) represent one of these candidates.

4.2.3 Time Window for t_3. Similar to $T_2(z)$, t_3 can be expressed as:

$$T_3(z) = \frac{1}{c}\left(\sqrt{\alpha z^2 - \beta + (z_3' - z)^2} - \sqrt{\alpha z^2 - \beta + (z_0' - z)^2}\right) + t_0$$

Fig. 7b shows the general shape of the graph of $T_3(z)$. We consider all τ_i that satisfy Eq. 13 as potential candidates for t_3. Let $t_3^{(l)}$ ($l \in \{1, 2, ...\}$) represent one of these candidates.

$$\min_z T_3(z) < t_3 < \max_z T_3(z) \tag{13}$$

4.3 Time-offset Estimation

Finally, we estimate the time offset δ using the four times of arrival computed in §4.2. Let the combinations which are constructed from $t_0, t_1^{(j)}, t_2^{(k)}, t_3^{(l)}$ obtained in §4.2 be $\mathrm{C}^{(i)}$. Note that in $\mathrm{C}^{(i)}$, $t_2^{(k)}$ and $t_3^{(l)}$ are values corresponding to the range calculated based on $t_1^{(j)}$, and $t_0 < t_1^{(j)} < t_2^{(k)} < t_3^{(l)}$ is also satisfied.

We apply the least-squares method to Eq. 3 with respect to all $\mathrm{C}^{(i)}$ to calculate $J^{(i)}$, which is the residual sum of squares (RSS), and estimate $\delta^{(i)}$. The SyncEcho methodology introduces redundancy by formulating four equations to solve for three unknowns: δ, d, and z. Consequently, the root sum of squares yields a minimal value for correct combinations of times of arrival, while producing a larger value for incorrect combinations. Therefore, the $\mathrm{C}^{(i)}$ that minimizes J is the correct combination, and its corresponding $\delta^{(i)}$ is the estimated time offset.

4.4 Correction of Sampling Clock Offset

Mobile device audio clocks are known to exhibit errors due to hardware imperfections and temperature variations [33]. A slight discrepancy between the audio clocks of the speaker and the microphone can result in clock drift, as shown in Fig. 8a. We operate under the assumption that this clock drift is linear and employ the least squares method, using multiple time offset measurements, to estimate both the initial clock offset and the presumed clock skew. Fig. 8b demonstrates the clock skew error relative to the number of measurements. Notably, both the mean error and variance diminish when the number of measurements surpasses 20. SyncEcho

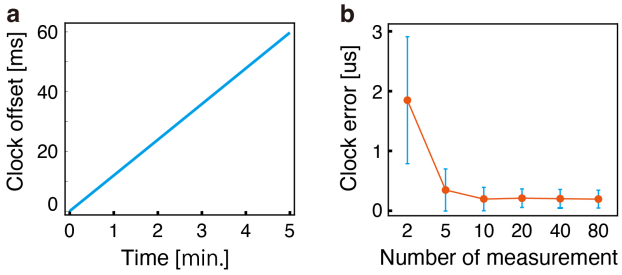


Figure 8: (a) Clock drift observed over a 5-minute measurement period, and (b) average and standard deviation of clock skew errors relative to the number of measurements used in linear regression.

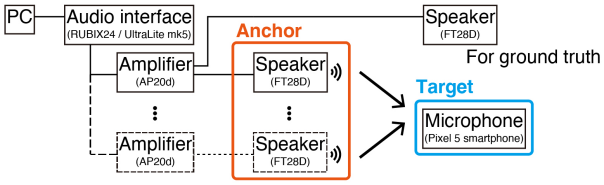


Figure 9: Experiment setup.

requires a span of five seconds for clock drift correction, given its capability to conduct four measurements per second.

5 EVALUATION

5.1 Experiment Setup

Fig. 9 shows the basic measurement setup. For the transmission system, we used a PC (MacBook Pro) to generate audio signals, an audio interface (ROLAND RUBIX24), an amplifier (Fostex AP20d), and speakers (Fostex FT28D). The main speaker outputs a chirp signal in the range of 16 kHz to 19 kHz with a duration of 20 ms, and the transmission interval I was set to 250 ms. For the receiver (*i.e.,* microphone), we used a Google Pixel 5 smartphone and set the sampling rate to 48 kHz. We 16-fold up-sampled the recorded data and then analyzed the data offline with Mathematica.

To evaluate the performance of the time offset estimation, we use two speakers: an anchor speaker to estimate the time offset and a reference speaker for acquiring the ground truth of the time

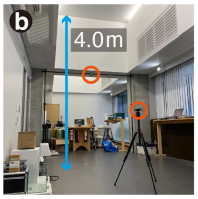
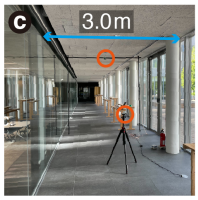
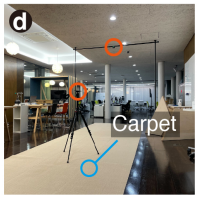
Measurement Environment					
Ceiling height	2.6 m	**4.0 m**	2.6 m	2.6 m	2.6 m
Width	6.3 m	3.8 m	**3.0 m**	6.3 m	6.3 m
Flooring material	Reflective	Reflective	Reflective	**Absorptive**	Reflective
Obstacle	Few	Few	Few	Few	**Many**

Figure 10: SyncEcho was evaluated in five different environments: (a) standard, (b) high-ceiling, (c) narrow-width, (d) carpeted, and (e) obstructed.

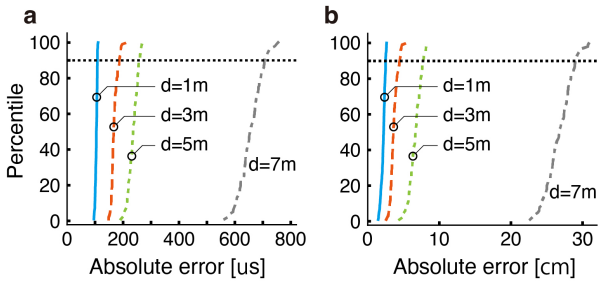


Figure 11: CDFs of (a) time offset and (b) speaker-to-microphone distance in xy plane, for different speaker-microphone distance.

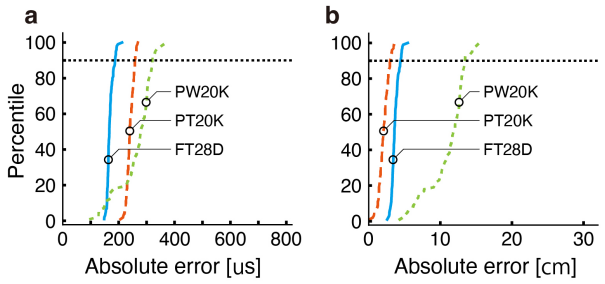


Figure 12: CDFs of (a) time offset and (b) speaker-to-microphone distance on xy plane for the three speakers.

offset. The reference speaker and the microphone were placed in close proximity to calculate the ground truth of the time offset. The signal from the reference speaker arrives at the microphone immediately, so by observing this timing, the transmission time in the microphone's audio clock can be acquired. In reality, there is a finite gap between the reference speaker and the microphone, so the propagation time corresponding to this gap (8 mm) is compensated. In the experiment, the reference speaker first transmits a 10-second signal so the receiver can calculate the true value, and then, after a time period of silence, the anchor speaker transmits a signal. The time offset estimation operation was repeated 100 times for each configuration.

The experiment was conducted in an environment with a ceiling height of 2.6 m and a width of 6.3 m. The flooring material was wood, and few obstacles existed in the space, as shown in Fig. 10a. The z coordinate of the microphone was set to 1.0 m, and the distance in the xy plane between the speaker and the microphone was set to 3.0 m. The range of z is set to z_{min} = 0.7 m and z_{max} = 1.5 m, assuming the user holds the microphone (*i.e.*, smartphone) in hand. From the following, the above settings are used unless experimental settings are explicitly mentioned in each section.

5.2 Basic Estimation Performance

This section evaluates the basic performance of SyncEcho.

5.2.1 Effect of Distance. To evaluate the estimation performance with different distances, we performed ranging at four distances between the speaker and the microphone: 1, 3, 5, and 7 m (note that this is the distance projected on the xy plane). Fig. 11a shows the cumulative distribution functions (CDFs) of the time offset for the four distances, and the 90th percentile errors were 109, 187, 259, and 706 μs, respectively. Fig. 11b is the CDFs of the distance between a speaker and a microphone on the xy plane, and the 90th percentile errors were 2.6, 4.5, 7.8, and 29.1 cm, respectively. The 90th percentile errors of the microphone's height were 1.2, 1.4, 0.9, and 2.4 cm, respectively. Generally, the estimation error increases with distance, which can be attributed to the decrease in SNR with increasing distance.

5.2.2 Speaker Variations. To evaluate how the estimation performance changes depending on the speaker, we performed ranging using three speakers: FT20D, PT20K, and PW20K, which were also used in §3.2. Fig. 12a shows the CDFs of time offset for the three speakers. The 90th percentile errors were 187, 259, and 321 μs, respectively. Fig. 12b shows the CDFs of the distance between a speaker and a microphone on the xy plane. The 90th percentile errors were 4.5, 3.0, and 13.4 cm, respectively. In addition, the 90th percentile errors of the microphone's height were 1.4, 1.2, and 3.0 cm, respectively. The results show that SyncEcho can accurately estimate time offset using various speakers, and the estimation performance varies depending on the speaker.

Table 2: Computation time

z-constraints	# of Candidates	Time [ms]
None ($0 < z < z'$)	3	191.5
0.7 m$< z <$ 1.5 m	1	155.8

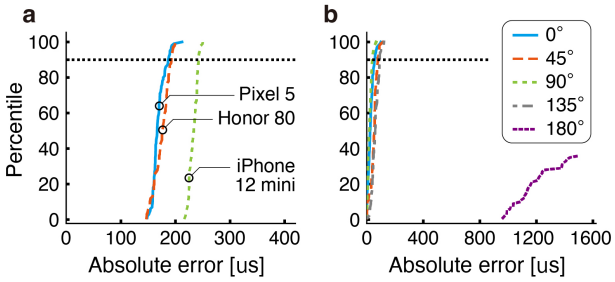


Figure 13: CDFs of time offset (a) for different smartphones and (b) for different microphone angles.

5.2.3 Computational Efficiency. We use the measurement data from §5.2.1 (100 trials at d = 5 m) to calculate the computation time per trial. We evaluate this time under two conditions: with and without constraints on the microphone's height. Table 2 displays the average computation time required per trial, along with the average number of candidate arrival time combinations computed. The acoustic signals are transmitted every 250 ms to mitigate reverberation, and the computation times fall below this 250 ms threshold, indicating SyncEcho's computational efficiency. Moreover, as Table 2 shows, constraining z successfully enhances this efficiency.

5.3 Impact of Microphones

5.3.1 Microphone Variation. To evaluate the performance differences between microphones, we performed time offset estimation using the Huawei Honor 80 and Apple iPhone 12 mini microphones as receivers, in addition to the Google Pixel 5 used in all other experiments. Each smartphone was placed so the microphone faced the speaker. Fig. 13a shows the CDF of time offsets for the different smartphones. The 90th percentile errors were 187 (Google Pixel 5), 193 (Huawei Honor 80), and 243 μs (Apple iPhone 12 mini), demonstrating that SyncEcho can accurately estimate time offset using various microphones.

5.3.2 Impact of Microphone Angle. To evaluate the impact of the microphone angle, we estimated the time offset while horizontally rotating the microphone's orientation by 45° increments. The angle when the microphone directly faces the speaker is defined as 0°. Fig. 13b shows the CDF of time offsets for different microphone angles. The 90th percentile errors were 56 (0°), 78 (45°), 40 (90°), and 91 μs (135°). At 180° (*i.e.,* the microphone facing the opposite direction of the speaker), time offsets couldn't be computed in all trials within 100 measurements; thus, the results showed large errors. We observed that the 3rd-order reflection via the floor and ceiling could not be received at 180°, likely due to the tripod blocking the signal. This could also happen when a user holds the smartphone,

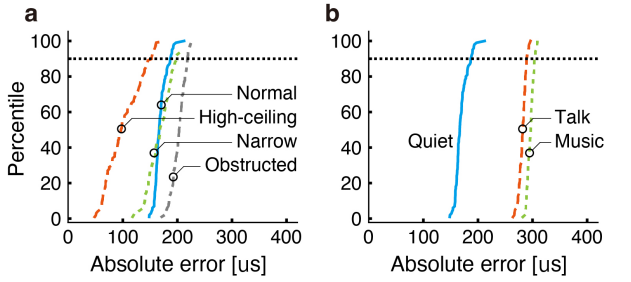


Figure 14: CDFs of time offset (a) for different spaces and (b) for different ambient noise conditions.

indicating a limitation of SyncEcho. However, it was experimentally confirmed that SyncEcho maintains high accuracy up to 135°.

5.4 Impact of Spatial Characteristics

To evaluate the impact of spatial characteristics, SyncEcho was tested in five different environments with distinct impulse responses shown in Fig. 10. The environment shown in Fig. 10a serves as a baseline, and Fig. 10b to 10e represent high-ceiling, narrow, carpeted, and obstructed environments, respectively. The speaker height was set to 2.3 m in the high-ceiling environment and 2.5 m in the other environments.

Impact of Ceiling Height. To evaluate the impact of ceiling height, experiments were conducted in a high-ceiling environment with a ceiling height of 4.0 m (see Fig.10b). In this environment, the propagation distance of reflected signals is longer, which is likely to decrease the SNR compared to a normal environment. Fig.14a shows the CDF of time offsets, with the 90th percentile error being 153 μs, indicating that SyncEcho works robustly in high-ceiling environments.

Impact of Wall Reflections. To evaluate the impact of wall reflections, experiments were conducted in a narrow corridor environment with a width of 3.0 m (see Fig.10c). In this environment, there is a higher probability of significant influence from reflected signals off walls compared to a normal environment. Fig.14a shows the CDF of time offsets, with the 90th percentile error being 199 μs. Despite the presence of wall reflections due to the short distance between the microphone, our proposed signal processing pipeline successfully removed these reflections, enabling high-accuracy time-offset estiamtion.

Impact of Floor Material. To evaluate the impact of floor material, experiments were conducted in a carpeted environment with sound-absorbing flooring (see Fig. 10d). In this environment, the high absorption rate of the floor is likely to decrease the SNR compared to a normal environment. SyncEcho failed in all measurements within this environment because the second and higher-order reflections were sufficiently attenuated, preventing their detection as peaks.

Impact of Obstacles. To evaluate the impact of obstacles, experiments were conducted in an obstructed environment (see Fig.10e). In this environment, there is a higher probability of significant influence from reflected signals off obstacles compared to a normal environment. Fig.14a shows the CDF of time offsets, with the 90th

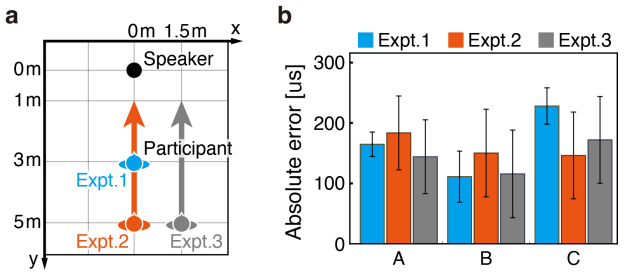


Figure 15: (a) Summary of the three experiments to evaluate the impact of user movements and (b) mean absolute error and standard deviation of time offset for each experiment. (Expt.: Experiment)

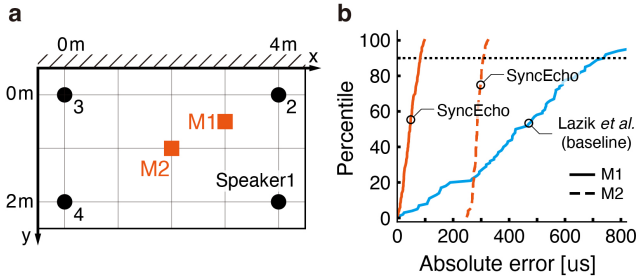


Figure 16: (a) Configuration of the speaker and microphone in the measurement, and (b) CDFs of time offset for SyncEcho and the baseline.

percentile error being 218 μs, demonstrating that SyncEcho is less affected by surrounding obstacles. This is because, compared to walls, floors, and ceilings, the reflective surfaces of many obstacles are relatively small, resulting in weaker reflected signal strength.

5.5 Impact of Ambient Noise

To evaluate the impact of ambient noise, we introduced two types of noise. The first type of noise is a human voice. We asked a participant to stand at 1 m and 90° with respect to the smartphone and read an article. The second type of noise is music that is played by an external smartphone. We placed the external smartphone near the smartphone for SyncEcho and played music. We measured the sound pressure levels by putting a sound level meter at the position of the measurement smartphone. In a quiet situation, it measured 41 dB. Additionally, human voices and music were set to be approximately 60 dB.

Fig. 14b shows the CDFs of time offset for different ambient noise conditions. The 90th percentile errors were 187 (quiet), 289 (human voice), and 304 μs (music), respectively. We observe that adequate accuracies can be achieved for different ambient noises since the frequency of human voice and music is usually below 4 kHz [36] which is much lower than the frequency band adopted for sensing (*i.e.*, 16 kHz - 19 kHz).

5.6 Impact of User Movements

To evaluate the impact of user movements, three participants (A: a 1.65 m tall male, B: a 1.63 m tall female, C: a 1.75 m tall male) conducted three experiments. In all experiments, participants were instructed to hold the smartphone in a natural position while looking at the screen. Fig. 15a shows the initial positions and movement paths of the participants in the three experiments. In experiment 1, participants were asked to be still for 5 seconds. In experiments 2 and 3, participants were asked to move for approximately 5 seconds in the directions described in Fig. 15a.

Fig. 15b compares the mean absolute error and the standard deviation of time offset among these three experiments. Fig. 15b shows that SyncEcho can accurately estimate time offset even when the user is moving. The reason for the larger standard deviation in Expt.2 and Expt.3, compared to Expt.1, is likely attributed to the variation of the distance between the speaker and microphone; the accuracy at $y = 5$ m was notably lower compared to the $y = 1$ m (refer to §5.2.1).

5.7 Comparison

We compare the performance of SyncEcho with the method proposed by Lazik *et al.* [17], which is the closest to our approach using an unmodified speaker. The four speakers required in their method are arranged as shown in Fig. 16a at a height of 2.5 m. For a fair comparison, the same chirp signal (16 kHz - 19 kHz) is transmitted from all speakers using time division multiple access. Given the spatial symmetry, microphones are positioned at M1 and M2, as shown in Fig. 16a. We conducted SyncEcho's evaluation using only one speaker.

Fig. 16b compares the CDFs of time offset for SyncEcho and the approach of Lazik *et al.* (baseline). The 90th percentile errors are 82 (SyncEcho, M1), 309 (SyncEcho, M2), and 740 μs (baseline, M1), respectively. In addition, their approach failed to accurately estimate the time offset at M2. This is attributed to their reliance on TDoA localization. When four speakers are installed on the ceiling (*i.e.*, and the height of all four speakers is the same), the estimation performance in the vertical direction significantly deteriorates in TDoA-based localization [1]. Lazik *et al.* circumvent this problem by making the height of the speakers and microphones the same, but this assumption is not usually practical.

These results reveal that our approach using a single speaker outperforms that of Lazik *et al.*, which requires four speakers.

5.8 ToF Localization

We implemented a ToF localization system based on SyncEcho's time offset estimation and evaluated the localization performance. Our localization setup consisted of two synchronized speakers, as shown in Fig. 17a, and the user holding a smartphone moved along a circular path To assess SyncEcho's pure localization performance, each speaker sent chirp signals using frequency division multiple access to minimize signal interference (Speaker 1: 16 kHz to 19 kHz, Speaker 2: 13 kHz to 16 kHz). The distance between the speakers was 4.08 m, and the time offset was estimated for 5 s using Speaker 1's signal at the starting point and leveraged for position estimation. The coordinates (x, y) of the smartphone were estimated by

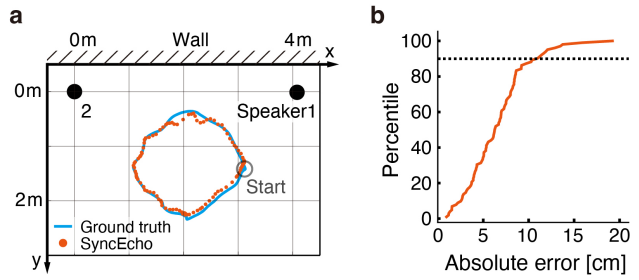


Figure 17: (a) Ground truth and SyncEcho-based trajectories as the user moved in a circular path. (b) CDF of localization errors, showing a 90th percentile error of 11.0 cm, demonstrating high accuracy of SyncEcho without dedicated hardware.

applying the least squares method to the following equation:

$$\arg\min_{x,y} \sum_{m=1}^{2} \left(\sqrt{(x - x'_m)^2 + (y - y'_m)^2 + (\hat{z} - z'_m)^2} - c(t^m - \hat{\delta}) \right)^2,$$

where $\hat{\delta}$ and $\hat{z}$ represent the time offset and the smartphone's height estimated by SyncEcho, respectively, and (x'_m, y'_m, z'_m) are the coordinates of speaker m. Additionally, t^m is the time of arrival of the direct signal from speaker m, and c is the sound speed.

As the ground truth, we used an HTC Vive tracker, which can achieve mm-level localization accuracy. Fig. 17a shows the trajectories of the ground truth and SyncEcho-based tracking, and Fig. 17b shows the CDF of the localization errors. The 90th percentile error was 11.0 cm. These results demonstrate that SyncEcho can achieve high-accuracy localization with a few speakers without requiring dedicated hardware.

6 DISCUSSION AND FUTURE WORK

6.1 Enhancing Accuracy and Robustness

Evaluation results demonstrate that our technique achieves high-accuracy time offset estimation, with a 90th percentile error of 259 μs at a 5 m distance. Naturally, this estimation performance declines as the distance between the speaker and the microphone increases, highlighting an area for future enhancement. Thus, one route for improvement is to introduce a weighing factor and emphasize the time offset acquired when the speaker-microphone distance is smaller.

In terms of robustness, a few incorrect time offset estimation results were observed when the user was in motion. This happens because the arrival time of the n-th reflected signals constantly changes with user movement, but several constraining factors can be considered for improvement. Examples of these factors are that the time offset is close to constant, the speaker-microphone distance varies continuously (*i.e.*, shouldn't "jump"), and the height of the smartphone remains relatively stable when held. It is promising to leverage these factors to remove outliers and, furthermore, enhance robustness. Moreover, enhancing localization performance by using reflections beyond the fourth order is a promising direction for future research.

6.2 Real-World Deployment

6.2.1 Implementation on Different Devices. SyncEcho can operate on any device equipped with a single microphone, making it promising for use in a variety of mobile and IoT devices beyond smartphones. In implementing SyncEcho on different devices, the most critical factor when deploying on other devices is the microphone's characteristics. To this end, we evaluated the performance of SyncEcho using different microphones in §5.3.1 and confirmed that they operate with sufficient accuracy. These experimental results suggest that SyncEcho can be deployed on a wide range of mobile and IoT devices in the future.

6.2.2 Complex and Noisy Environments. In §5.4 and §5.6, we confirmed that SyncEcho performs well in various scenarios, including those with obstacles and user movement, as long as there is a line-of-sight (LOS) between the speaker and the microphone. Furthermore, §5.5 demonstrated that time offset estimation will robustly work under noisy environments (*e.g.*, talking, playing music) with only a minor decay in accuracy. These results suggest that SyncEcho will perform robustly in various real-world scenarios.

Although further non-ideal situations such as non-line-of-sight (NLOS) conditions and extreme noise can occasionally occur in practice, the advantage of estimating time offsets for localization is that it can continuously support localization even if the time offset is estimated periodically. Thus, exploring system-level designs to leverage higher-level contexts could improve system robustness. For example, developing algorithms to seamlessly handle frequent transitions between LOS and NLOS, or classifying external noise to apply appropriate denoising algorithms, are potential avenues for future work.

6.3 Limitations

§5.4 demonstrates that SyncEcho works well in many situations, such as in high-ceiling rooms, near walls, and around obstacles. There are, however, specific situations where it may not perform as effectively, like in rooms made with sound-absorbing materials or curved ceilings. These are common challenges for all systems that use reflection-based positioning.

Despite this, one great benefit of SyncEcho is that it can keep doing ToF localization for tens of minutes without multipath reflections after it calculates the time offset using a single speaker. This means that if the device goes into a multipath-rich zone with a single speaker once during this time, it can continuously perform ToF localization regardless of the multipath environment. This benefit significantly reduces the impact of the above challenges, making SyncEcho useful in many more conditions.

7 CONCLUSION

This paper proposed SyncEcho, the first approach to estimate the time offset with a single speaker, by using higher-order reflections. SyncEcho operates under various conditions, enabling ToF-based localization. A notable benefit of SyncEcho is that it only requires a single, unmodified speaker for localization. This feature significantly lowers the threshold for incorporating location awareness into computer systems and will support the emergence of novel information systems and user interfaces.

ACKNOWLEDGMENTS

This work was supported by a joint research program with Daikin Industries, Ltd. (Research Institute for an Inclusive Society through Engineering, Space Design for Co-creation).

REFERENCES

[1] Takayuki Akiyama, Masanori Sugimoto, and Hiromichi Hashizume. 2017. Time-of-arrival-based indoor smartphone localization using light-synchronized acoustic waves. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* 100, 9 (2017), 2001–2012.

[2] Joan Bordoy, Christian Schindelhauer, Fabian Höflinger, and Leonhard M Reindl. 2019. Exploiting acoustic echoes for smartphone localization and microphone self-calibration. *IEEE Transactions on Instrumentation and Measurement* 69, 4 (2019), 1484–1492.

[3] Chao Cai, Henglin Pu, Peng Wang, Zhe Chen, and Jun Luo. 2021. We hear your pace: Passive acoustic localization of multiple walking persons. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–24.

[4] Haige Chen and Ashutosh Dhekne. 2022. Pnploc: Uwb based plug & play indoor localization. In *2022 IEEE 12th International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. IEEE, 1–8.

[5] Zhe Chen, Guorong Zhu, Sulei Wang, Yuedong Xu, Jie Xiong, Jin Zhao, Jun Luo, and Xin Wang. 2019. M3: Multipath assisted Wi-Fi localization with a single access point. *IEEE Transactions on Mobile Computing* 20, 2 (2019), 588–602.

[6] Giorgio Conte, Massimo De Marchi, Alessandro Antonio Nacci, Vincenzo Rana, Donatella Sciuto, et al. 2014. BlueSentinel: a first approach using iBeacon for an energy efficient occupancy detection system.. In *Proceedings of the 5th ACM Workshop On Embedded Systems For Energy-Efficient Buildings*. Citeseer, 11–19.

[7] Francesca De Cillis, Luca Faramondi, Federica Inderst, Stefano Marsella, Marcello Marzoli, Federica Pascucci, and Roberto Setola. 2017. Hybrid indoor positioning system for first responders. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 50, 2 (2017), 468–479.

[8] Viktor Erdélyi, Trung-Kien Le, Bobby Bhattacharjee, Peter Druschel, and Nobutaka Ono. 2018. Sonoloc: Scalable positioning of commodity mobile devices. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. 136–149.

[9] Bernhard Großwindhager, Michael Rath, Josef Kulmer, Mustafa S Bakr, Carlo Alberto Boano, Klaus Witrisal, and Kay Römer. 2018. SALMA: UWB-based single-anchor localization system using multipath assistance. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*. 132–144.

[10] Bernhard Großwindhager, Michael Stocker, Michael Rath, Carlo Alberto Boano, and Kay Römer. 2019. SnapLoc: An ultra-fast UWB-based indoor localization system for an unlimited number of tags. In *Proceedings of the 18th International Conference on Information Processing in Sensor Networks*. 61–72.

[11] Andy Harter, Andy Hopper, Pete Steggles, Andy Ward, and Paul Webster. 1999. The anatomy of a context-aware application. In *Proceedings of the 5th annual ACM/IEEE international conference on Mobile computing and networking*. 59–68.

[12] Fabian Höflinger, Rui Zhang, Joachim Hoppe, Amir Bannoura, Leonhard M Reindl, Johannes Wendeberg, Manuel Bührer, and Christian Schindelhauer. 2012. Acoustic self-calibrating system for indoor smartphone tracking (assist). In *2012 international conference on indoor positioning and indoor navigation*. IEEE, 1–9.

[13] Yuming Hu, Feng Qian, Zhimeng Yin, Zhenhua Li, Zhe Ji, Yeqiang Han, Qiang Xu, and Wei Jiang. 2022. Experience: Practical indoor localization for malls. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*. 82–93.

[14] Yifei Jiang, Xin Pan, Kun Li, Qin Lv, Robert P Dick, Michael Hannigan, and Li Shang. 2012. Ariel: Automatic wi-fi based room fingerprinting for indoor localization. In *Proceedings of the 2012 ACM conference on ubiquitous computing*. 441–450.

[15] Manikanta Kotaru, Kiran Joshi, Dinesh Bharadia, and Sachin Katti. 2015. Spotfi: Decimeter level localization using wifi. In *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*. 269–282.

[16] Ye-Sheng Kuo, Pat Pannuto, Ko-Jen Hsiao, and Prabal Dutta. 2014. Luxapose: Indoor positioning with mobile phones and visible light. In *Proceedings of the 20th annual international conference on Mobile computing and networking*. 447–458.

[17] Patrick Lazik, Niranjini Rajagopal, Bruno Sinopoli, and Anthony Rowe. 2015. Ultrasonic time synchronization and ranging on smartphones. In *21st IEEE Real-Time and Embedded Technology and Applications Symposium*. IEEE, 108–118.

[18] Patrick Lazik and Anthony Rowe. 2012. Indoor pseudo-ranging of mobile devices using ultrasonic chirps. In *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems*. 99–112.

[19] Dong Li, Jialin Liu, Sunghoon Ivan Lee, and Jie Xiong. 2020. FM-track: pushing the limits of contactless multi-target tracking using acoustic signals. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 150–163.

[20] Jie Lian, Jiadong Lou, Li Chen, and Xu Yuan. 2021. Echospot: Spotting your locations via acoustic sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–21.

[21] Wenguang Mao, Wei Sun, Mei Wang, and Lili Qiu. 2020. Deeprange: Acoustic ranging via deep learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–23.

[22] MarketsandMarkets. 2023. Indoor Location Market by Component (Hardware, Solutions, and Services), Technology (BLE, UWB, Wi-Fi, RFID), Application (Emergency Response Management, Remote Monitoring, Predictive Asset Maintenance), Vertical and Region - Global Forecast to 2028. https://www.marketsandmarkets.com/Market-Reports/indoor-location-market-989.html. Accessed: 2024-06-21.

[23] Hiroaki Murakami, Takumi Suzaki, Masanari Nakamura, Hiromichi Hashizume, and Masanori Sugimoto. 2020. Five degrees-of-freedom pose-estimation method for smartphones using a single acoustic anchor. *IEEE Sensors Journal* 21, 6 (2020), 8030–8044.

[24] Bradford W Parkinson and Stephen W Gilbert. 1983. NAVSTAR: Global positioning system—Ten years later. *Proc. IEEE* 71, 10 (1983), 1177–1186.

[25] Chunyi Peng, Guobin Shen, Yongguang Zhang, Yanlin Li, and Kun Tan. 2007. Beepbeep: a high accuracy acoustic ranging system using cots mobile devices. In *Proceedings of the 5th international conference on Embedded networked sensor systems*. 1–14.

[26] Swadhin Pradhan, Ghufran Baig, Wenguang Mao, Lili Qiu, Guohai Chen, and Bo Yang. 2018. Smartphone-based acoustic indoor space mapping. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–26.

[27] 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.

[28] Sheng Shen, Daguan Chen, Yu-Lin Wei, Zhijian Yang, and Romit Roy Choudhury. 2020. Voice localization using nearby wall reflections. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–14.

[29] Merrill I Skolnik. 1962. Introduction to radar. *Radar handbook* 2 (1962), 21.

[30] Masanori Sugimoto, Hayato Kumaki, Takayuki Akiyama, and Hiromichi Hashizume. 2016. Optimally modulated illumination for rapid and accurate time synchronization. *IEEE Transactions on Signal Processing* 65, 2 (2016), 505–516.

[31] Verified Market Research. 2024. Global In-Ceiling Speaker Market Size By Type (Active In-Ceiling Speaker, Passive In-Ceiling Speaker), By Application (Household, Commercial), By Geographic Scope And Forecast. https://www.verifiedmarketresearch.com/product/in-ceiling-speaker-market/ Accessed: 2024-06-26.

[32] Anran Wang, Jacob E Sunshine, and Shyamnath Gollakota. 2019. Contactless infant monitoring using white noise. In *The 25th Annual International Conference on Mobile Computing and Networking*. 1–16.

[33] Lei Wang, Haoran Wan, Ting Zhao, Ke Sun, Shuyu Shi, Haipeng Dai, Guihai Chen, Haodong Liu, and Wei Wang. 2023. Scalar: Self-calibrated acoustic ranging for distributed mobile devices. *IEEE Transactions on Mobile Computing* (2023).

[34] Weiguo Wang, Jinming Li, Yuan He, and Yunhao Liu. 2020. Symphony: localizing multiple acoustic sources with a single microphone array. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 82–94.

[35] Jie Yang, Simon Sidhom, Gayathri Chandrasekaran, Tam Vu, Hongbo Liu, Nicolae Cecan, Yingying Chen, Marco Gruteser, and Richard P Martin. 2011. Detecting driver phone use leveraging car speakers. In *Proceedings of the 17th annual international conference on Mobile computing and networking*. 97–108.

[36] Fusang Zhang, Zhi Wang, Beihong Jin, Jie Xiong, and Daqing Zhang. 2020. Your Smart Speaker Can" Hear" Your Heartbeat! *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–24.

[37] Zengbin Zhang, David Chu, Xiaomeng Chen, and Thomas Moscibroda. 2012. Swordfight: Enabling a new class of phone-to-phone action games on commodity phones. In *Proceedings of the 10th international conference on Mobile systems, applications, and services*. 1–14.

[38] Bing Zhou, Mohammed Elbadry, Ruipeng Gao, and Fan Ye. 2017. BatMapper: Acoustic sensing based indoor floor plan construction using smartphones. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. 42–55.

[39] Bing Zhou, Mohammed Elbadry, Ruipeng Gao, and Fan Ye. 2017. Battracker: High precision infrastructure-free mobile device tracking in indoor environments. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*. 1–14.

[40] Yuzhou Zhuang, Yuntao Wang, Yukang Yan, Xuhai Xu, and Yuanchun Shi. 2021. ReflecTrack: Enabling 3D Acoustic Position Tracking Using Commodity Dual-Microphone Smartphones. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. 1050–1062.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009