Scribe: Simultaneous Voice and Handwriting Interface

YANG BAI, University of Maryland College Park, USA
IRTAZA SHAHID, University of Maryland College Park, USA
HARSHVARDHAN TAKAWALE, University of Maryland College Park, USA
NIRUPAM ROY, University of Maryland College Park, USA

This paper presents the design and implementation of *Scribe*, a comprehensive voice processing and handwriting interface for voice assistants. Distinct from prior works, *Scribe* is a precise tracking interface that can co-exist with the voice interface on low sampling rate voice assistants. *Scribe* can be used for 3D free-form drawing, writing, and motion tracking for gaming. Taking handwriting as a specific application, it can also capture natural strokes and the individualized style of writing while occupying only a single frequency. The core technique includes an accurate acoustic ranging method called Cross Frequency Continuous Wave (CFCW) sonar, enabling voice assistants to use ultrasound as a ranging signal while using the regular microphone system of voice assistants as a receiver. We also design a new optimization algorithm that only requires a single frequency for time difference of arrival. *Scribe* prototype achieves 73 μm of median error for 1D ranging and 1.4 mm of median error in 3D tracking of an acoustic beacon using the microphone array used in voice assistants. Our implementation of an in-air handwriting interface achieves 94.1% accuracy with automatic handwriting-to-text software, similar to writing on paper (96.6%). At the same time, the error rate of voice-based user authentication only increases from 6.26% to 8.28%.

CCS Concepts: • **Human-centered computing → Interaction design process and methods**.

Additional Key Words and Phrases: Handwriting tracking, voice assistants, cross frequency continuous wave sonar

ACM Reference Format:
Yang Bai, Irtaza Shahid, Harshvardhan Takawale, and Nirupam Roy. 2023. *Scribe*: Simultaneous Voice and Handwriting Interface. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 4, Article 148 (December 2023), 31 pages. https://doi.org/10.1145/3631411

1 INTRODUCTION

Voice assistants, such as Amazon Echo, Google Home, and Apple Siri, have become the fastest-growing consumer technology ever, with growth faster than smartphones [7, 54]. Nearly 50% of U.S. households have voice-enabled smart speakers, and more than 157 million such devices are sold in the US alone to date [1]. Naturally, voice interfaces are evolving to become the default interface to smart environments and home automation. This gives rise to an ecosystem of innovative acoustic sensing and perception capabilities around these devices beyond traditional speech recognition. A predominant set of these techniques repurposes this acoustic interface for various motion tracking with an active sound source or its reflections [33, 51, 60, 71, 88]. However, these techniques fail to coexist with the primary functionality of the voice interfaces – voice activation. They require exclusive access to the microphones and therefore disrupt voice assistants' default behavior of continuous wake-word detection and

Authors' addresses: Yang Bai, University of Maryland College Park, Maryland, USA, yangbai8@umd.edu; Irtaza Shahid, University of Maryland College Park, Maryland, USA, irtaza@umd.edu; Harshvardhan Takawale, University of Maryland College Park, Maryland, USA, htakawal@umd.edu; Nirupam Roy, University of Maryland College Park, Maryland, USA, niruroy@umd.edu.

2474-9567/2023/12-ART148
https://doi.org/10.1145/3631411

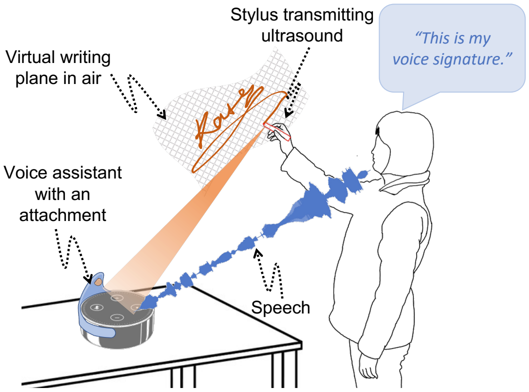

Stylus transmitting ultrasound
"This is my voice signature."
Virtual writing plane in air
Voice assistant with an attachment
Speech

Fig. 1. *Scribe* enables voice assistants to recover handwritten notes, signatures, and trajectories using an acoustic stylus. It can be used to sign a payment using two-factor authentication.

subsequent speech recognition. To be adopted by the existing infrastructure of voice-first devices, motion-based applications should be transparent to voice processing on these devices. This paper aims to introduce this novel capability of simultaneous voice and motion interfacing on voice assistants. Our technique, however, does not compromise on the state-of-the-art motion tracking accuracy while keeping the speech recognition performance unaltered. We demonstrate this dual functionality through a full-fledged natural handwriting interface on voice assistants, called *Scribe*.

A natural handwriting and sketching interface on voice assistants opens up a new range of applications that was difficult to imagine before. As shown in Figure 1, possible applications will include taking notes during a phone conversation over voice assistants or draw simple sketches. If executed well, a handwriting interface can capture features of a personalized writing and even user's physical signature for online payments. It can also enable search with a combination of verbal description and sketch (e.g. "Alexa, which North American tree has dark green leaf like this ⟨draws a rough sketch of the leaf ⟩"). In security applications, the simultaneous voice and handwriting interface can be a tool for two factor authentication and challenge-response based user verification and *liveness* detection. While a wide variety of novel applications are possible, a practical implementation of such a system needs to meet several challenges.

With the advent of smart devices, several techniques have been proposed to enable a virtual writing interface in the air [48, 53, 70, 87]. While writing recognizable words is possible with these techniques, they are far from being a natural handwriting interface. Unlike written or drawn on paper or a tablet screen, these in-air writing interfaces can not capture the individualized style of writing or replicate the precise strokes of a drawing. We ask the question, *Is it possible to develop a virtual interface in-air to write effortlessly using natural handwriting?* The voice assistants come with a well-designed acoustic frontend including a planar microphone array ideal for localization and spatial analysis of sound. Still, it is challenging to capture the natural strokes of spontaneous handwriting. The personalized style of handwriting includes subtle sub-mm movements of the pen tip continuously moving in the air. Moreover, handwriting recovery will need to translate the arbitrary 3D trajectory of the pen to a 2D writing recognizable by humans and machines. We address the first challenge in our system by developing a novel motion-tracking technique and the second challenge using a sequence of post-processing methods customized for the handwriting interface.

The latest techniques for contactless motion detection build on the principles of active sonar. The Frequency Modulated Continuous Wave (FMCW) sonar technique is particularly effective for detecting small motions due to its high sensitivity and range resolution. Here an acoustic speaker generates a chirp signal with continuously varying frequency within a fixed bandwidth and the receiving microphone detects the distance of the source from the phase delay of the received chirp. However, the resolution of the FMCW sonar depends on the slope of the chirp which is limited by its bandwidth. The speaker-microphone system of smart speakers has an overall bandwidth of around 24 kHz [71] which can potentially lead to a fine resolution for gesture recognition. Unfortunately, the vast majority of this frequency range falls within the human audible range, and thus sending a probing signal in this frequency range leads to loud and disturbing noise. Therefore, past works on acoustic motion detection systems limit the probing signal to use only frequencies above 18 kHz, which is partially *inaudible* to adults. While it drastically narrows down the available bandwidth to only 6 kHz (i.e., frequencies between 18 to 24 kHz [89]) and degrades range resolution, younger people still show sensitivity to sound in this range — ultimately defeating the purpose of both high resolution and inaudibility.

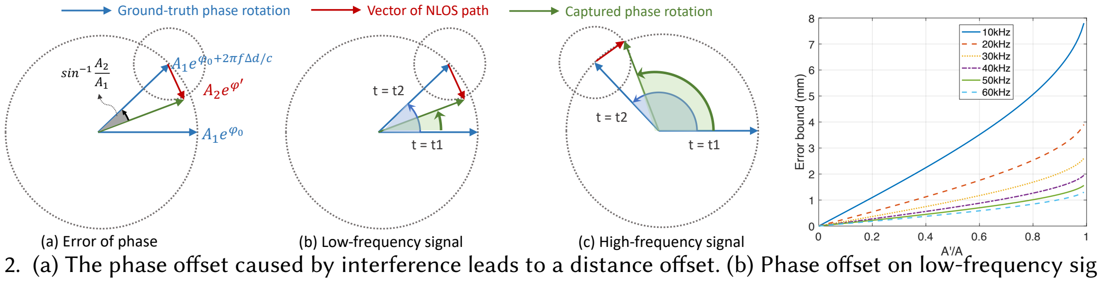
Scribe addresses this limitation by using a very high-frequency ultrasound signal of unrestricted bandwidth for ranging. Although regular voice assistants are not capable of receiving ultrasound signals, we show that the inherent non-linearity present in the off-the-shelf microphone systems provides a natural multiplication operation in the acoustic signal path. If signals are carefully designed, they can leverage this implicit multiplication operation to shift down a high-frequency signal to the low-frequency band, which the microphone can readily record without any modification to the hardware. We build on this intuition to develop our *cross-frequency continuous wave* sonar that generates ultrasound for ranging but receives and processes the signal at the regular audible range in the off-the-shelf voice assistants. Note that the built-in speakers of a voice assistant can not produce these high-frequency ultrasound signals. We design an external add-on speaker module capable of producing ultrasound signals. This speaker module operates as a stand-alone external speaker connected through the audio port of the smart speaker. Along with the low-latency high-resolution 3D trajectory tracking, *Scribe* applies geometric mesh parameterization technique to recover the handwriting from an arbitrary virtual 3D surface to a flattened projection on a 2D plane. The post-processing eliminates any stray trajectory due to pen-lift and unwanted movements of the pen.

In addition to handwriting detection, the core ranging and localization with the cross-frequency sonar can lead to applications that require high-resolution tracking. Examples include teleoperation and remote surgery with precise hand motion tracking, real-time orientation estimations for AR/VR systems, and bodily vibration and tremor detection. While several opportunities for applications open up, this paper focuses on developing the core capabilities and assessing the limits of CFCW sonar and *Scribe* systems. We have made the following three specific contributions at the current stage of this project:

- Designed, implemented, and evaluated the cross-frequency continuous wave sonar that enables unmodified acoustic devices to receive ultrasound ranging signals while operating on the audible frequency range.
- Developed a sequence of processing methods to recover handwriting from 3D trajectory drawn in the air. The output produces handwriting that retains the individualized writing styles and legible to humans as well as machines.
- Implemented a hardware/software prototype of the end-to-end system for the community to reproduce, evaluate, and build on the *Scribe* system.

2 PRIMER: LOCATION FROM PHASE

Distance estimation is a fundamental building block of localization, and its accuracy depends on the Time of Flight (ToF) estimation. The phase of a coherent signal is a measurable parameter that continuously changes



(a) Error of phase (b) Low-frequency signal (c) High-frequency signal

Fig. 2. (a) The phase offset caused by interference leads to a distance offset. (b) Phase offset on low-frequency signal. (c) Phase offset on high-frequency signal. (d) Distance offset caused by interference is proportional to the inverse of frequency. High-frequency signal is more robust to interference in distance tracking.

over time. Therefore, if the phase of the received signal is estimated correctly, it can tell the time delay of signal propagation or the ToF from its origin at the beacon. When transmitting a pure tone, the received signal can be represented as $S_R = sin(2\pi f(t - t_p))$. Here f is the frequency of the signal, and t_p is the time delay of propagation. Through analyzing the phase of recorded signal $2\pi f t_p$, we can calculate the time of arrival t_p. The phase shift of two adjacent samples can be formulated as $\Delta\phi = 2\pi f \Delta d/c$, where Δd is the distance change within the time period and c is the speed of sound.

2.1 Advantage of Pure-Tone Based Ranging

FMCW is a widely accepted technique for ranging and localization. However, it has two limitations. FMCW transmits a chirp that occupies a wide frequency band, and the resolution of distance tracking is proportional to $\frac{c}{B}$, where c is speed of sound and B is the bandwidth of chirp. To achieve higher resolution, a wide frequency band is occupied. If the band overlaps with the 0-8kHz frequency band of voice, it disables the voice interface. Existing works use near-ultrasound frequency band for the chirp, but not all the devices support near-ultrasound frequency band. Voice assistants such as Amazon Alexa and Amazon Echo only support 0-8kHz frequency band. Similar to FMCW, Time-of-Arrival (ToA) and Time-Difference-of-Arrival (TDoA) also require a wide frequency band for precise correlation. The advantage of pure-tone based ranging is that it only requires one frequency, minimizing the interference with voice interface and other applications requiring a wide frequency band.

2.2 Advantages of High-Frequency Signal

2.2.1 Distance resolution is proportional to the frequency. The sensitivity of distance measurement using phase depends on the wavelength of the signal as the phase changes 2π radian per wavelength. With same distance change, a larger phase change is shown in higher frequency signal. The microphones of the voice assistants operate at the low-frequency audible signals of longer wavelength leading to poor ranging and localization performance.

2.2.2 High-frequency signal leads to smaller distance offset under interference. Let us assume that the interference caused by environmental noise or multipath creates an indirect path with a lower amplitude than the direct path in phase domain. With same distance change, Figure 2 shows the complex representation of phase change for both high-frequency and low-frequency signal. The line-of-sight path is shown in blue and the interference is shown in red. The sum of the signal in green is the captured phase at the receiver side. Suppose the object moves in distance d, the theoretical phase shift is $e^{2\pi f d/c}$. The maximum phase error $\Delta\phi$ caused by the interference is $sin^{-1}(\frac{A_2}{A_1})$ [70]. Figure 2(b) and (c) shows the phase offsets on low and high frequencies. Although the largest phase offsets in low frequency and high frequency are the same, the distance offset caused by the phase offset is proportional to the inverse of frequency. If we convert the phase offset to distance offset, $\Delta d = \frac{\Delta\phi c}{2\pi f}$. Figure 2 (d)

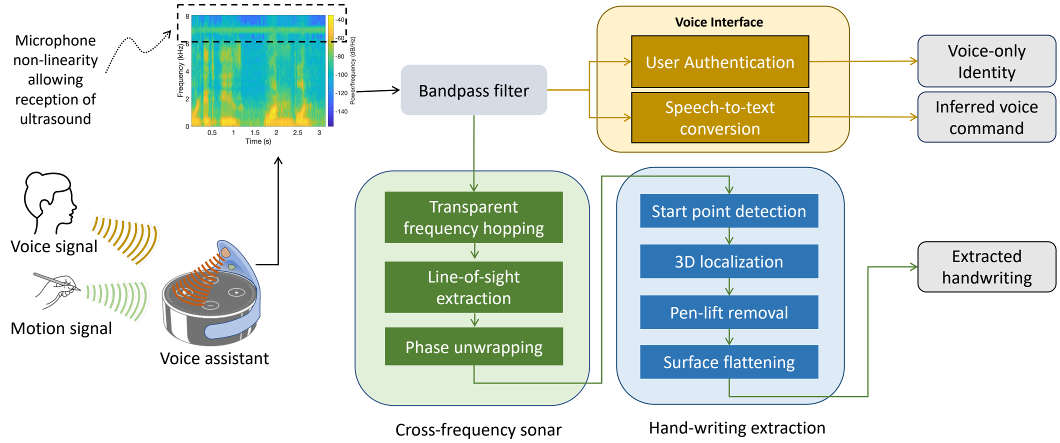


Fig. 3. System overview of *Scribe*.

shows the distance error as a function of A_2/A_1 and frequency. The key observation is the distance offset caused by interference is proportional to the inverse of frequency. In other words, the high-frequency signal is more robust to interference in distance tracking.

2.3 Challenges of Pure Tone-based Ranging

2.3.1 Vulnerable to multipath distortion. Unmodulated pure tone-based systems suffer from an environmental multipath effect that degrades the phase accuracy of the signal. Although pure tone ranging is accurate, multipath fading severely impacts the accuracy. The non-line-of-sight signals fall in the same frequency. This inseparable overlap interferes with the phase values, leading to inaccurate distance estimation.

2.3.2 Inability to obtain initial distance. While pure-tone based ranging can capture precise distance change, it cannot capture the absolute initial distance due to phase wrapping. The phase of absolute distance is $2n\pi + \phi$, phase tracking can only capture ϕ, without information of the number of wrapping n. We address these limitations in our novel ranging technique explained next.

3 CROSS-FREQUENCY SONAR DESIGN

Scribe develops an acoustic ranging method that uses two completely separate bands for the transmit signal and received signal – leading to the ranging technique we call Cross-Frequency Continuous Wave (CFCW) sonar. If works, CFCW sonar can introduce several crucial advantages to sensing applications with these household devices. **(a) (Accuracy)** Given it can leverage high-frequency ultrasound signals for ranging, it can have orders of magnitude higher accuracy compared to audible frequency-based techniques. **(b) (Inaudibility)** It allows accessing a wide band of inaudible frequencies with regular microphones, while existing techniques can only use a small band of 'near-ultrasound' frequencies between 18-24 kHz. **(c) (Power and hardware simplicity)** CFCW sonar operates at a low frequency which requires lower sampling rates. It offers similar accuracy to ultrasound with low complexity hardware as the power consumption and processing and storage requirements of the platform increase with the frequency of operation.

The CFCW sonar leverages implicit frequency translation possible in regular microphones. It carefully designs its ultra-sound transmit signals such that they can leverage the fundamental nonlinearity in microphones for automatic down conversion of the signals to the low-frequency recording range of the microphones. While

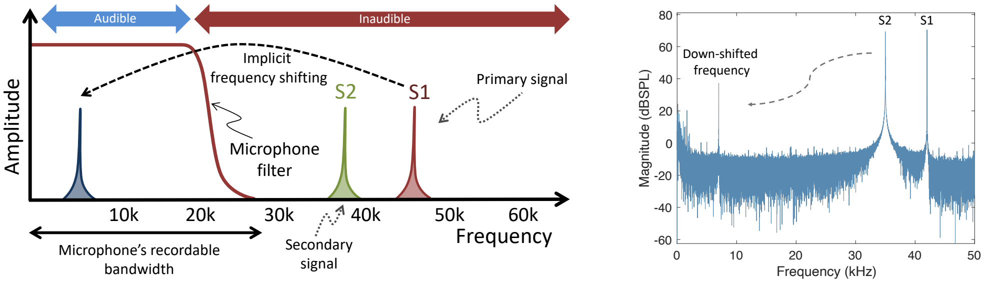


Fig. 4. (a) The non-linearity of the microphone multiplies the signals and implicitly shifts the frequency to the audible band. The figure is reproduced from [58]. (b) This figure shows a real-world experiment of frequency shifting with non-linearity of microphone. With ultrasound frequencies S1 and S2 as 42kHz and 35kHz, the down-shifted frequency is 7kHz.

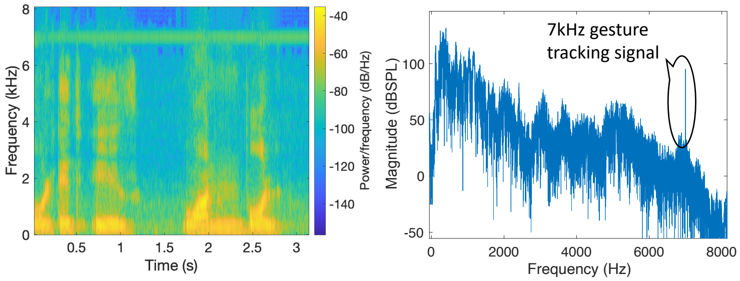


Fig. 5. (a) Spectrogram and (b) FFT of the signal captured by Amazon Echo Dot voice assistant.

nonlinear frequency conversion enables the core CFCW technique, it requires further signal design to eliminate environmental effects for effective distance tracking and start point detection for *Scribe*. The system is illustrated as Figure 3. We explain the step-by-step signal model next.

3.1 Implicit Frequency and Phase Translation

3.1.1 Microphone non-linearity primer. Commonly microphone is a linear system, which means the signal recorded by the microphone is a linear combination of the input signals. If the input signal is S, the recorded signal S_{out} is $S_{out} = A_1 S$, where A_1 is the complex gain. However, as introduced in Backdoor [59], the microphone shows non-linearity when the frequency of the input signal is above $25kHz$. The recorded signal S_{out} can be modeled as $S_{out} = \sum_{n=1}^{N} A_n S^n$. Since the third and upper terms are too weak and thus can be ignored, the S_{out} can be represented as $S_{out} = A_1 S + A_2 S^2$. As shown in Figure 4, we transmit two ultrasound pure tones together as $S = sin(2\pi f_1 t) + sin(2\pi f_2 t)$, where f_1 and f_2 are 45kHz and 38kHz. With nonlinearity of microphone, the recorded signal also includes the components produced by S^2, which are $f_1 + f_2$ and $f_1 - f_2$. While $f_1 + f_2$ is higher than the threshold that can be recorded by the off-the-shelf microphones, $f_1 - f_2$ can be recorded with proper selection of frequencies, as shown in Figure 4(a). We further measure the non-linearity of omnidirectional ADMP401 MEMS microphones [24] with a sampling rate of 100kHz. A down-shifted frequency in audible band is created with non-linearity of microphone which aligns with the theory. Next, we experimentally verify whether an off-the-shelf device can show such implicit frequency shift. We exposed an Amazon Echo Dot voice assistant with the dual high-frequency signals and recorded the shifted signal. Frequency-domain analysis of the recorded signal is shown in Figure 4(b). We also verify that the Amazon Echo Dot device can simultaneously record the human voice along with the shifted tracking signal as shown in Figure 5.

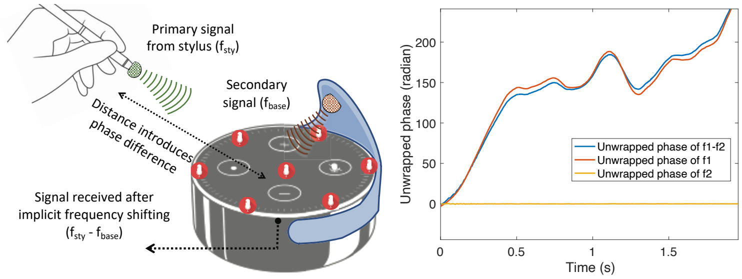


Fig. 6. (a) The principle for distance estimation using phase from shifted frequency. (b) The unwrapped phases of frequencies f_1 and $f_1 - f_2$.

Interestingly, if phase change is happened in frequency f_1, the phase change can be mapped to the down-converted frequency $f_1 - f_2$. As $S = sin(2\pi f_1 t + \phi) + sin(2\pi f_2 t)$, the received signal at frequency $f_1 - f_2$ is $sin(2\pi(f_1 - f_2)t + \phi)$. Therefore, we map the phase of frequency f_1 on $f_1 - f_2$ by down-converting the frequency using non-linearity of microphone. Here we call f_1 as primary channel and f_2 as secondary channel. The primary signal source acts as an acoustic beacon attached to the target object. The secondary signal source is fixed using an add-on module to the voice assistant. The secondary source emits a low-power signal that illuminates only the microphone array of the voice assistant. These signals combine inside the microphones to create a different signal that carries the phase offset between the primary and secondary tones. The phase difference is due to the distance of the target from the voice assistant. Figure 6(a) shows the working principle for distance estimation using the phase-based method on the implicitly shifted low frequency. In Figure 6(b), we show the unwrapped phases of primary ultrasound signal, secondary ultrasound signal, and implicitly produced low-frequency signal. The frequency of the secondary signal is almost constant with a only small variation within 0.02 radian due to the environmental interference. The phases of primary ultrasound signal and the down-converted signal follow the same trends, with negligible difference because of the variation of phase from secondary signal as well as the different noise level. Figure 7(a) shows a physical setup to leverage the dual-frequency model of the CFCW sonar. Note that the location of the secondary signal is crucial. Although we assume the secondary signal only illuminates the microphone array, the microphone array itself also acts as a reflector to induce multipath, as well as the objects around it, such as the surface of the table. Therefore, we need to select the location that has most negligible multipath. We simulate the ray tracing from secondary speaker to the microphones on the smart speaker in COMSOL, when the secondary speaker faces in parallel with the surface of smart speaker or faces towards it vertically. In the scene, the smart speaker is put on a table, and human's hand is writing on the top of the smart speaker. Figure 7(b) shows the impulse response from two situations. From the simulation, we found when the secondary signal is in parallel with the surface of smart speaker, the reflection from the objects is smaller. This is because the surface of the table acts as a reflector and induces more multipath when the secondary speaker faces to the smart speaker vertically. Based on this result, we select the location of the secondary speaker as parallel to the smart speaker.

3.2 Multipath Avoidance

3.2.1 Line-of-sight extraction. *Scribe* relies on the accurate phase estimation of the signal for ToF measurement, however phase of the received signal can be severely affected by environmental effects of which multipath signal propagation is the most significant. Multipath is a natural phenomenon where a signal, after leaving the transmitter, reflects off objects in the environment to create replicas and the replicas propagate through paths of different delays before combining at the receiver. The lengths of these individual paths decide the phase delays of the replicas and therefore their superimposition leads to an unknown amplitude and phase of the received signal. For location tracking, *Scribe* requires the phase of the direct line-of-sight (LOS) path, but any non-line-of-sight

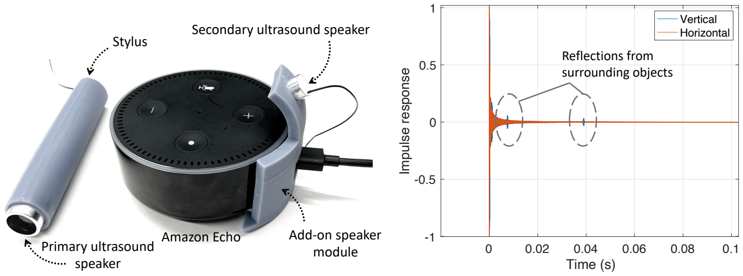


Fig. 7. (a) The model for the dual-frequency add-on module for voice assistant. (b) Simulated impulse responses of the secondary signal when the secondary speaker faces vertically and horizontally to the device.

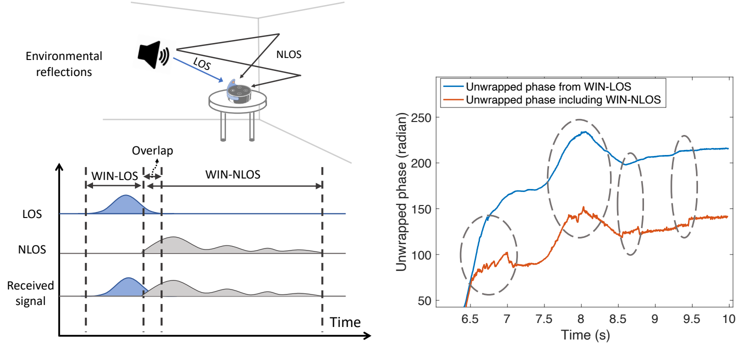


Fig. 8. (a) In the combined signal of LOS and NLOS channels, there exists a window in which only LOS signal is present. (b) Comparison of the extracted unwrapped phases from only WIN-LOS and including WIN-NLOS. Interference causes phase jitters.

(NLOS) path can introduce an unpredictable error to the phase. However, note that there is only one LOS path, and its length is shorter than all NLOS paths. NLOS paths are indirect paths and therefore arrive at the receiver later than the LOS signal. This provides a window of opportunity to extract the undistorted phase from the LOS path before the multipath replicas superimpose on this signal as explained next. Potentially the phase estimation system can use a part of the signal clear from multipath superposition to get an accurate phase of the transmit signal.

Theoretically, in a pulse-based probing signal, there exists a time window when only the LOS signal is present at the receiving sensor. Figure 8(b) compares the unwrapped phases from only WIN-LOS and with WIN-NLOS. Unexpected jitters in phase happen when extracting phases including the NLOS. The length of this time window, shown as WIN-LOS in Figure 8(a), is equal to the delay of the first NLOS path which is an unknown environmental parameter. The smallest delay of the NLOS path, which leads to the smallest size of the time window, is related to the distance of the closest large objects to the device. Another parameter that impacts the worst-case estimate of the time-window WIN-LOS is the FFT resolution. *Scribe* estimates the phase by first calculating the FFT coefficients of the time series data of length WIN-LOS seconds. This signal is sampled by the microphone at 16 kHz standard sample rate of Alexa) after the nonlinear conversion of the ultrasound signal. A too-short WIN-LOS will lead to wider FFT bins combing a wider band of frequencies to the same bin. At the baseband, the microphone also records ambient sounds, such as human voices and other household noise. The energy of these sounds is mostly limited within 4 kHz [76] and the CFCW signal is mapped to 7 kHz at the baseband. Therefore, we limit the FFT bin width to a maximum of 2 kHz, which in turn limits the minimum size of WIN-LOS to 0.5 ms. In other words, in an ideal case, there should not be a reflector that can create a NLOS signal that has a delay of within 0.5ms relative to the LOS signal. With a speed of around 340m/s, 0.5ms is mapped to 17cm of distance for sound transmitting in the air, which indicates the distance of the NLOS path should be at least 17cm longer than the LOS path. Within 17cm of distance difference, the most possible and strongest reflector is the hand holding the stylus. Note that we do not need to eliminate the reflection from the hand, as it follows the same trajectory as the stylus, thus causing a similar phase change. Although we can extract an accurate phase by only using WIN-LOS for phase extraction, the lasting WIN-NLOS will pollute the following signals, stopping us to find the next WIN-LOS for accurate phase extraction. To solve this problem, we design a pulsed signal that can hop to a different frequency to avoid delayed NLOS signals. As explained next, the CFCW receiver can continuously operate on a constant frequency while the ultrasound-ranging signals hop.

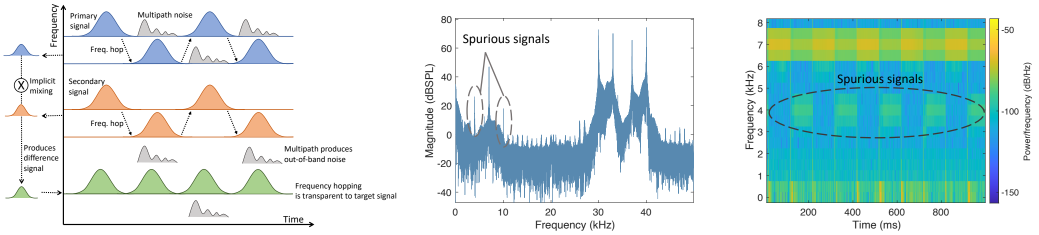


Fig. 9. (a) An illustration of transparent frequency hopping. Although the transmitted frequencies keep changing, the frequency of the received signal is constant. The spurious signal does not interfere with the frequency of tracking. (b) FFT of the signal captured by microphone with a sampling rate of 100kHz. (c) Spectrogram of the captured signal below 8kHz.

3.2.2 Transparent frequency hopping. In CFCW sonar, the received frequency f_{rcv} is different from the dual frequencies This spurious frequencyf_{sty} and f_{base}) transmitted in the air and it is equal to the difference between the two transmitted frequencies ($f_{rcv} = f_{sty} - f_{base}$). Therefore, if the transmitted frequencies change in a synchronized way to maintain a constant difference between them, the received signal will remain constant. In other words, we can design a frequency hopping for the transmitted signal pair while its effect is transparent to the received signal. We used this observation to avoid multipath in the CFCW sonar.

The dynamic multipath distortion is primarily caused by the f_{sty} signal source, which is facing towards the voice interface, changes location over time, and is of relatively higher power to have stronger NLOS reflections. The f_{base} signal source, on the other hand, is placed in a static location only a few centimeters away from the microphone array. Moreover, the power of the f_{base} signal is kept low – just enough to reach all the microphones in the array – making NLOS multipath from this signal too weak to cause a significant impact on the received difference signal. During the signal transmission process, f_{sty} and f_{base} signals hop to new values $\hat{f}_{sty}$ and $\hat{f}_{base}$ respectively. Although the frequency of transmitted signals keep changing, the received CFCW frequency f_{rcv} is constant as long as $f_{sty} - f_{base}$ does not change. More importantly, during frequency hopping, the multipath also only produces out-of-band noise, without interfere with frequency for tracking. As shown in Figure 9(a), the delayed NLOS components of f_{sty} mixes only with the current $\hat{f}_{base}$ to create a spurious signal at frequency ($f_{sty} - \hat{f}_{base}$). This spurious frequency has a clear gap between the frequency for tracking, leaving tracking unaffected by the multipath effect. As an experimental validation, we record the received signal using 100kHz sampling rate to capture both ultrasound and audible components. As shown in Figure 9(b), the spurious frequencies caused by multipath are on 4kHz and 10kHz, with a clear gap between the 7kHz component for tracking. Figure 9(c) shows the spectrogram of the frequency band below 8kHz. This transparent frequency hopping-based multipath avoidance hops transmitter's frequencies to new values together every 3 ms. Since we are transmitting two ultrasound signals in the air, only when the microphone captures the two signals simultaneously, a down-converted audible band signal can be created. To make sure there is an overlap between the two transmitted signals with a supported distance of 70cm, we set a rate of frames 333 per second. It takes 2ms for the primary signal to arrive at the microphone in 70cm of distance. We leave another 1 ms for phase extraction, resulting in a frame rate of 333. In summary, different from pure tone-based tracking that has NLOS signals fall in the same frequency, CFCW shifts NLOS signals to different frequencies to mitigate the NLOS interference. At the same time, CFCW built in with the advantages of high-frequency pure tone-based tracking, including only occupying one frequency, having high distance resolution, and more resilient to interference.

3.2.3 Phase unwrapping. Phase unwrapping is a classical approach to recover the original phase value from the wrapped phase value. The phase value is wrapped within $[-\pi, \pi]$, and the goal of phase unwrapping is to reconstruct the continuous phase by removing the "sudden jumps". The true phase (i.e., true phase without wrapping) can be reconstructed as long as the difference between the subsequent phases is less than π. To satisfy this condition, we need to guarantee the speed is less than $\frac{c}{2f_1\Delta T}$. With a f_{sty} 40kHz and ΔT 3ms, the maximum speed is 1.41m/s. It is revealed that the peak velocity of hand movement is 2.7m/s, and the peak velocity of hand gesture is 1.8m/s [23]. Meaning our system can result in an error when the moving speed is high, limiting us to achieve high accuracy with even higher frequency. To solve this problem, we induce speed to estimate the direction of phase rotation. Since the hand movement is continuous, the speed of hand movement also varies continuously, without a sudden jump. For example, the speed of hand movement cannot have a sudden change from 1.5m/s to -1.5m/s within milliseconds. Therefore, when we find a sudden speed change from above 1m/s to a negative speed from the unwrapped phases, it indicates that the speed exceeds 1.41m/s and the phase change is larger than π. Based on the speed information, we identify the direction of phase unwrapping between clockwise and anti-clockwise. Ideally, when the speed was over 1m/s in the last samples and was not decreasing dramatically, it is more likely the phase is rotating anti-clockwise. Otherwise, when the speed was over 1m/s in the last samples and suddenly decrease to a negative speed, it is rotating clockwise. By inferring the direction of phase change using speed, we only require the absolute difference between subsequent phases to be within 2π, thus the highest speed we can support is 2.83m/s. As shown in Figure 10(a), when the speed of movement exceeds the threshold, simply unwrapping the phases results in a sudden speed change. We detect the sudden speed changes and correct the direction of phase unwrapping continuously. Figure 10(b) shows a comparison between simply unwrapping the phase and applying speed-based phase unwrapping.

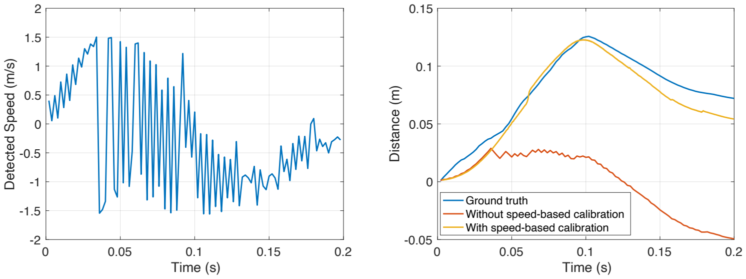


Fig. 10. (a) Traditional phase unwrapping results in sudden speed changes. Phase changes more than π when the speed exceeds 1.41m/s. (b) A comparison of distance tracking with and without speed-based phase unwrapping when the speed of movement exceeds 1.41m/s.

3.2.4 Why synchronization is not required? We do not require synchronization between transmitters and receivers. First of all, even with synchronization, we still cannot find the initial distance by using a single frequency. A precise correlation for ToF requires a wide band of frequency. To get the absolute distance for localization, we design an algorithm to detect the distance at the start point without the requirement of synchronization or a wide band signal as we explain next.

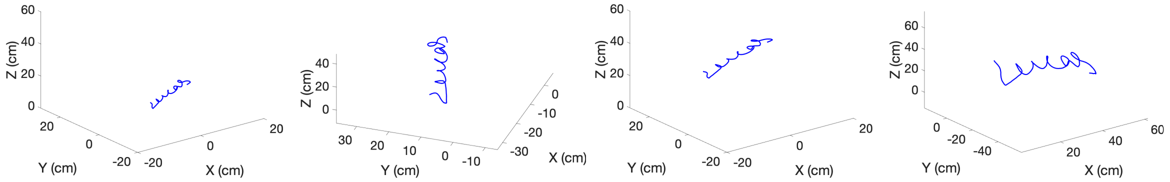


Fig. 11. Signature with (a) correct start point and (b,c,d) with start point offsets up to 45cm. The handwriting only rotates and shrinks, with the correct shape of handwriting.

3.3 Start Point Detection

To get the 3D trajectory matches with the ground truth, we need an accurate start point for the initial distance. However, as an application of handwriting system, we are concerned with the shape of the writing, which mainly profiles the relative instead of absolute spatial distances between samples. As shown in Figure 11, we experimentally give the wrong start points with an offset of up to 45cm. Although the handwriting rotates and shrinks, the overall shape of the writing remains the same. However, to better evaluate the 3D tracking accuracy, we design a start point detection algorithm for 3D tracking applications that require absolute locations.

Without synchronization between transmitters and receivers, our start point detection algorithm builds on a traditional approach called Time Difference of Arrival (TDoA). TDoA approach locates a target at intersections of hyperbolas or hyperboloids that are generated with foci at each fixed receiver of a pair. Existing algorithm of estimating the time delay between two receivers is to find the time samples that have highest correlation between the transmitted and received signals. However, the time delay is not precise enough due to two reasons: (1) The time resolution of the correlation is limited by the sampling rate of received signal. (2) A precise correlation requires a sharp pulse as transmit signal, which requires a wide frequency band. Unfortunately, using wide frequency band hinders the co-existence between voice interface and tracking for low sampling rate devices like voice assistants. Since the distance between the microphones is only 3.6cm on voice assistants, the distance drift causes even larger error on location estimation.

To address the issue mentioned above, we take advantage of the precise phase from CFCW sonar. Suppose the wrapped phase difference captured by a microphone pair is θ_{ij}, and the true phase difference is $2n\pi + \theta_{ij}$, where n is the number of wrapping. Based on the theory of triangle, the maximum distance difference from speaker to two microphones is the distance between two microphones. Thus, suppose the range between the pair of microphones is d and the wave length of signal is λ, we can define the range of n as $n \in [-\frac{d}{\lambda}, \frac{d}{\lambda}]$, where n is an integer. We find that with multiple pairs of microphones, only when the vector N map with the ground truth, the hyperbolas can cross on the same point, i.e., the start point. In simulation, we loop through every possible N for microphone pairs to optimize the location P and the corresponding error $E(P)$. The optimized location $\hat{P}$ is

$$\hat{P} = \arg\min_{P} \sum_{i \neq j} \left\| \lambda(N_{ij} + \frac{\theta_{ij}}{2\pi}) - c\tau_{ij}(P) \right\|^2 \tag{1}$$

where i and j represents the ids of microphones, λ is the wave length, N_{ij} is the number of phase wrapping for time difference of arrival τ_{ij}, c is speed of sound, and θ_{ij} is the captured phase difference between two microphones, i.e., $\theta_i - \theta_j$. As shown in Figure 12, although there are many other replica peaks, only the true start point has the minimum optimization error. Therefore, our optimization problem aims to find $\hat{N}$ and $\hat{P}$ that gives the minimum error. Since bruteforcely looping all possible N is computationally heavy, the optimization boils

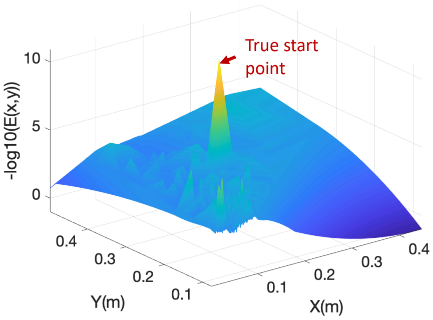


Fig. 12. The optimization error.

down to the minimization of

$$\hat{P}, \hat{N} = \underset{P,N}{\arg\min} \sum_{i \neq j} \left\| \lambda(N_{ij} + \frac{\theta_{ij}}{2\pi}) - c\tau_{ij}(P) \right\|^2 \tag{2}$$

where $N \in [\frac{-D}{\lambda}, \frac{D}{\lambda}]$ and D is the vector of distance between each pair of microphones. As shown in Figure 12, this objective function is non-convex, with many local minima. Moreover, this function is discrete, with N as integers. To solve this challenge, we apply a non-convex optimization algorithm called genetic algorithm [50] that also accept integer parameters. By minimizing the loss, we optimize the N and P simultaneously to find the true start point with global minima. This approach frees up the requirement of using a wide band signal to capture the time delay for TDoA. We translate the high-precision acoustic tracking method mentioned in this section to a handwriting interface prototype through a set of application specific techniques. Before elaborating on these techniques in Section 5, we discuss our user study to understand the scope of the application space and suitable form of the input device for such applications in the next section.

4 USER STUDY: FORM OF THE INPUT DEVICE AND TARGET APPLICATIONS

Scribe aims to develop an interface with the voice-first devices to capture strokes drawn or written in the air. The system has opted for an active stylus as the input device over the existing alternative forms including hand-worn wearable or simply using the fingers to write. The question of a suitable input device for writing or drawing in general has been explored in the existing literature and a significant number of past studies concluded that input by stylus is more accurate compared to that by finger [22, 47, 55, 66]. A comparative empirical analysis of finger and pen strokes [66] shows the stylus-drawn gestures are closer to reference shapes for parameters such as the size ratio, aperture between start point and end point, corner shape distance and intersecting points deviation, and axial symmetry. Another paper [55], uses a 20-DoF hand kinematic model and performs an analytical study. It shows that the stylus leads to higher precision and more isotropic motion performance. The advantage of a stylus lies in the additional support provided with multiple fingers [9], the smaller area of the tip of the stylus compared to the tip of a finger [66], ease to observe the motion of the tip compared to occlusion by finger [69], and most importantly the familiarity with this mode of input as humans are taught to write with a stylus-like devices (e.g., pens and pencils) since childhood. Interacting with smart devices using a stylus is adapted as a natural mode compared to the index finger which has a steeper learning curve.

4.1 Process

We divided the overall user study in two parts, as the flow shown in Figure 13. First we conducted an online survey with 100 participants on the Amazon Mechanical Turk (AMT) to understand the general opinion on

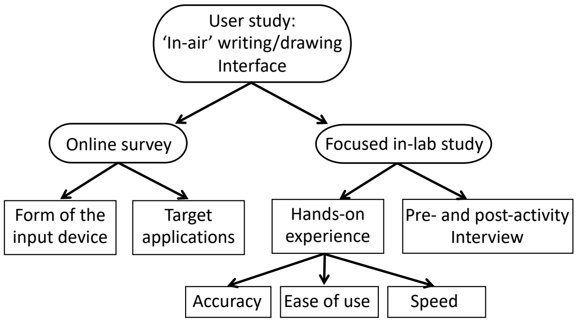


Fig. 13. Different components of the user study.

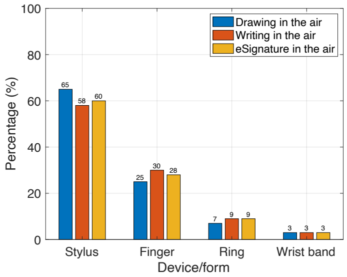


Fig. 14. Preference of input device for drawing, writing, and signing documents in the air.

possible application of the 'in-air' writing and drawing techniques with voice assistants and to learn the opinions from general public. On a scale of 1-5, 86% of the participants have a familiarity with the latest mobile and wearable technologies including or above 4, and 96% of them use voice assistants, with 74% use voice assistants at least once a day. 87% of the participants have the experience of using stylus with a tablet. The online survey shows a clear inclination toward the stylus as an input device as elaborated later in this section. However, the online survey participants most likely do not have first-hand experience with an 'in-air' writing interface. To bridge this gap, we recruited 10 participants for a separate in-lab, hands-on study and interview to learn specific requirements from the 'in-air' input device and opinions on the modality of the input device. For this in-lab study and interview we selected candidates who are familiar with latest smart technologies and experienced in regularly using multiple input devices (e.g., touch display, stylus, smart rings, computer mouse, joysticks) for writing, drawing, and gesture-based control. Participants took part in an 'in-air' writing and drawing activities using a stylus and the index finger while the system captured the strokes along with a camera-based tracking system for reference. The participants were interviewed before and after the activities.

4.2 Findings: Input Device Preference

4.2.1 Online survey. In general, 92% of the participants feel writing or drawing in the air can be a useful technique, and 78% feel moderately likely or very likely use this technique regularly. Figure 14 shows the distribution of users' preferred input device among the choices of stylus, finger, smart ring, and smart wrist-band for writing in the air.
When asked the reason for selecting a specific input device for in-air writing, drawing, or signing documents, participants who opted for a pen-like device (stylus) primarily mentioned the ease of use and familiarity of such a device.

> "I think it's the most natural version and the one that I am probably most accustomed to." [P1]
>
> "Pen-like device will be much better than finger. Because we are used to that." [P2]

A significant number of participants mentioned accuracy as the reason to favor stylus over any other mode of input.

> “I would prefer a stylus-based device for drawing/writing/eSignature in the air because it is highly accurate, quick, and easy to use. The stylus allows for precision and smooth strokes while writing or drawing, allowing for accurate representations of shapes and lines. The device also feels natural to use...” [P4]
>
> “I think a stylus is much more accurate than anything else and I’m used to using it.” [P5]

A group of participants favored stylus for writing which needs more control over the strokes than sketches.

> “Drawing freeform would be easier with just my finger, but for actual writing, I would prefer to use a stylus, so that I can more precisely control exactly where the device is reading from.” [P6]

The 28% participants who preferred to use a finger for ‘in-air’ input device in the online study, predominantly mentioned the ease of device-free interaction as primary reason.

> “Finger is the only option that doesn’t require another object to work so it would be ideal for everyone, and at any time. Pens and rings would get lost and a wrist band feels like it would be inaccurate for tracking smaller movements.” [P7]

4.2.2 In-lab study and interview. According to our survey as well as past studies on input modalities [22, 47, 55, 66], a stylus or a pen-like device is preferred primarily due to its accuracy, ease of use, and speed. Our in-lab study focuses on these parameters in understanding users’ opinion. In this in-lab study, we asked 10 participants to perform ‘in-air’ writing and drawing activities using both stylus and finger. We categorize the tasks as ‘simple’ and ‘complex’. The simple activities involved writing short words and copying simple sketches, while the complex tasks included writing long words and phrases and drawing a given object from imagination.

(A) Accuracy of the strokes. Figure 15 shows some example outputs that indicates better control, maneuver, and accuracy with a stylus, particularly for complex writing and drawing. In the post-activity interview, 7 out of 10 participants confirms that the stylus’s outputs are closer to their “own handwriting on a paper”.

We asked the participants about their preferences for both simple and complex tasks before using the systems, after using the systems but before viewing the results, and after viewing the results. As shown in Figure 16(a,b), in general for simple writing/drawings, the preference between stylus and finger is evenly distributed, while for complex movements, 9 out of 10 participants prefer stylus before usage and the preferences remain the same after usage in favor of stylus. However, one person switched to finger after viewing the outputs for the following reason. Note that in some cases, the user’s preference is undecided.

> “if there is no tool needed, its easier to just write whenever you feel like instead of looking for a tool” [L5]

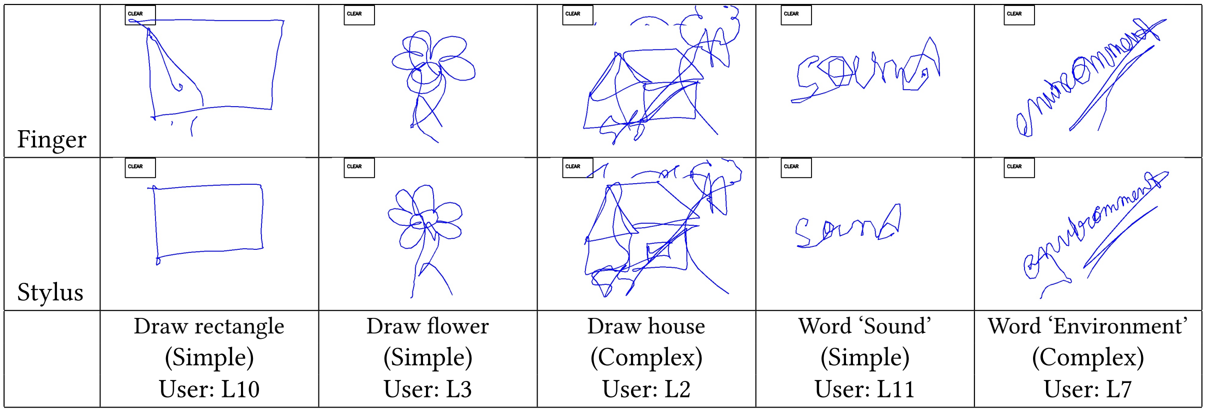

CLEAR
Finger
Stylus
Draw rectangle (Simple) User: L10
Draw flower (Simple) User: L3
Draw house (Complex) User: L2
Word 'Sound' (Simple) User: L11
Word 'Environment' (Complex) User: L7

Fig. 15. Comparison of outputs using stylus and finger. The participants confirms that the outputs from the stylus are more accurate to their intended strokes. Note: The strokes are not post-processed (e.g., pen-lift elimination).

However, the participant also mentioned that "its just more comfortable to write longer or more complex stuff with the stylus" that supports our motivation.

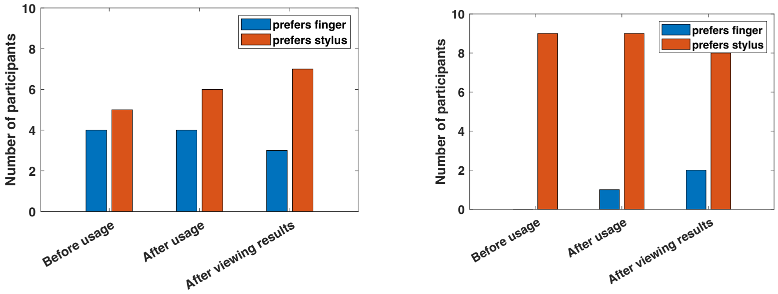


Fig. 16. (a) Preference of stylus and finger for simple drawings and writings and (b) for complex drawings and writings.

Five out of 8 participants who selected stylus as preferred device after comparing the outputs mentioned accuracy as one of the factors to opt for this input device. Following are some notable comments.

> "It depicted what I was trying to write more accurately." [L1]
>
> "Easier, output of stylus writing is actually closer to my actual writing style." [L3]
>
> "it gives better control over the details and continuity." [L5]

(B) Ease of use. Besides accuracy, comfort or convenience of use is another important fact for an input device. Existing work [55] shows that writing with finger results in greater displacement of wrist compared to writing with stylus. Four out of 8 participants who selected stylus as preferred device for complex tasks mentioned comfort or ease of use as the factor.

> “its just more comfortable to write longer or more complex stuff with the stylus” [L5]
>
> “I’m used to drawing/writing with a pen, the output with a pen will be better as I can rely more on muscle memory” [L8]

Wrist motion is related to energy expenditure during the task of writing and thus writing with fingertips induces higher effort for the user [62]. A stylus in some other form factor such as a ring [84] or wrist-band [82] show similar problem. Another work [34] shows even professionals that regularly used digital screens preferred the stylus as an input method over finger as it proved to be easier, faster and more accurate. Our findings are aligned with this study regarding the ease of use.

(C) Speed to writing or drawing. There has been inconclusive debate about stylus or finger is faster in performing writing tasks where [55] claims finger is faster but on taking a deeper dive [66] claims that stylus is actually faster when drawing complex shapes. In our in-lab study and interviews, none of the participants was confident if the speed of writing or drawing was a decisive factor to select the stylus or the finger as an input device.

4.3 Findings: Target Applications and Scopes of Improvement

We asked both online survey and in-lab study participants about possible utilities of the simultaneous voice and handwriting interface on voice-first devices. All the responses were positive about this additional interface and added to our understand of the possible application space. A majority of the responses mentioned taking quick and time-sensitive notes, including grocery lists with specific information, phone numbers, appointment reminders, or a mathematical formulae of a measurement. A participant mentioned taking notes of a specific shape and being able to interact with the voice-assistant with both speech and approximate drawings. One of the in-lab study participants mentioned that such an interface can encourage journal keeping by allowing to add a thought easily as it comes to mind. A significant number of participants expressed excitement in the possibility of electronically signing a document on voice assistants. The major concerns about the stylus-type input device and direction for improvement was about making the device lightweight and comfortable, identifying and eliminating pen lifts between two segments of letters, alignment of the virtual writing plane, and a way to provide haptic feedback to emulate a writing surface in the air.

Along with physical measurements of the movement of the active stylus in air during writing, we incorporate the application-specific feedback collected during the user studies in our prototype design.

5 APPLICATION-SPECIFIC DESIGN

Precise motion tracking on voice assistants can benefit a wide range of existing applications and open up possibilities for new ones. Several non-verbal human-machine interfaces can be developed with gesture and motion tracking. We have developed natural handwriting detection as the representative application. Detection of words written in the air is a classic application explored with a variety of sensing modalities. We pushed the boundary of this application by not only enabling fine-grained motion tracking required for subtle strokes created during sub-mm movements of the fingers, but also developed techniques for post-processing the data so that even the handwriting can be verified from the notes. Interestingly, during the process of motion detection, *Scribe*’s signal does not interfere with the voice signal and therefore the voice assistant can perform its regular operations while simultaneously capturing the writing.

Scribe detects the pen strokes by localizing the acoustic source (the primary signal). However, recovering meaningful writing from the continuous trajectory of the pen tip will require addressing two challenges. **(a)**

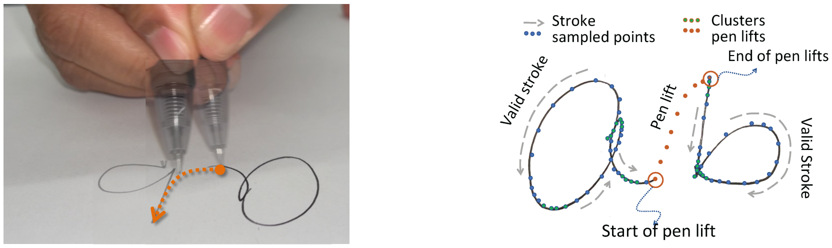


Fig. 17. An analysis of the handwriting process. Each character has two to four clusters that have slower moving speed than the strokes and pen lifts.

Pen-lift elimination: While writing on paper or other solid surfaces, we often lift the tip to jump to the next stroke which is disconnected from the current one, as shown in Figure 17. This event is transparent while writing on a solid surface like a paper or a tablet screen as it does not generate any strokes. However, 3D tracking-based writing detection cannot naturally distinguish a pen-lift event from the actual strokes. **(b) Flattening virtual surface:** Given the user can write at any angle on her assumed surface in the air and the surface is not flat, the projection of the 3D trajectory on the horizontal plane leads to distortion of the strokes. The distortion often makes the writing illegible to both humans and automatic text recognizing software.

5.1 Localization in 3D Space

Distances from at least three spatially separated microphones enable us to apply the traditional trilateration method [27] to find the 3D location of the target in space. We use distance measurements from all the available microphones for location estimation using a multi-lateration technique. Therefore, the location of the target should be somewhere on a sphere of radius d_m, centered at the microphone's location. Theoretically, we should be able to find the location of the target by solving the system of N equations. However, in practical systems, the individual distance measurements are not perfect due to noise, and therefore, the spheres do not have a common solution. We use the following optimization to find the estimated location of the target in 3D space. It essentially minimizes the distance of the estimated location from all the spheres.

$$\arg\min_{x,y,z} \sum_{i=1}^{N} \sqrt{(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2 - d_i^2} \tag{3}$$

5.2 Removing Pen-Lift From Virtual Surface

The 3D trajectory contains both the actual pen strokes, which are a part of the writing, as well as the stray movements, like pen-lifts. As shown in Figure 18(b), the written word 'fit' is unrecognizable from the raw trajectory of the pen because of unwanted marks due to pen-lifts in between two disconnected strokes. To avoid this error, we will first need to identify the accurate start and end of the pen-lift trajectories.

Handwritten letters and drawings can be viewed as a collection of stroke segments of different curvatures and lengths. Each of these segments starts and ends with a change in direction of the stroke. While it is difficult to quantify an angle or sharpness in the change of directions of segments due to various writing styles and user-dependent features, the velocity of the pen movement offers a reliable feature. In Figure 18(a), we show the speed change when writing the word "fit". We detect the clusters where the speed arrives at a local minimum. Then we show the detected clusters in Figure 18(b). The red points are those mapped to the local minima in speed, which are the turning points of handwriting. The segment-by-segment annotated handwriting reveals that the speed of the pen comes to a low value at the end of a segment and just before the beginning of the next,

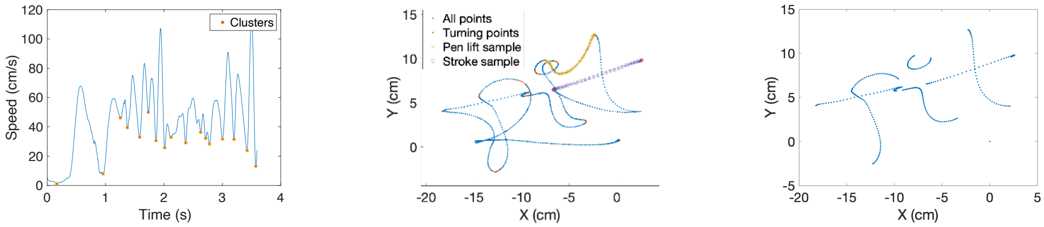


Fig. 18. An example of pen lift removal of the word "fit". (a) The variation of speed during handwriting. We detect the clusters of local minimas. (b) When mapping the detected clusters with local minimum of speed, we find the clusters are the turning points of writing. (c) Processed "fit" with all pen lifts removed.

as shown in Figure 17(b). We adopted this intuition by first calculating the velocity of the pen between every two samples of location on the trajectory.

As shown in Figure 18(a) and (b), a cluster of 3D location points that corresponds to the local minima of the time-varying velocity indicates the start or end of a segment. Note that these clusters i.e., the ends of the segments necessarily belong to the imaginary writing plane in the air. Now it is left to identify a pair of such clusters indicating the start/end of a pen-lift. To this end, first, we fit a triangular plane to each three of the clusters. A stroke segment joining two clusters of this triangle is either entirely a valid stroke or a pen-lift. We calculate the orthogonal distance of all points of a segment from this plane to identify a pen-lift segment by its off-plane movement. Figure 18(c) is the processed handwriting after removing pen lifts. In a few rare cases, the user slows down during a pen-lift movement which erroneously produces a cluster on a pen-lift trajectory. Our algorithm mistakenly includes these pen-lift trajectories as valid strokes. We correct this by fitting a third-order plane through all the clusters. Given the variations in the imaginary writing, a plane is smooth corresponding to sudden movements for pen-lifts, the spurious clusters on the pen-lift trajectory stand out from this fitted plane. These clusters are then pruned before applying the triangle fitting mentioned above.

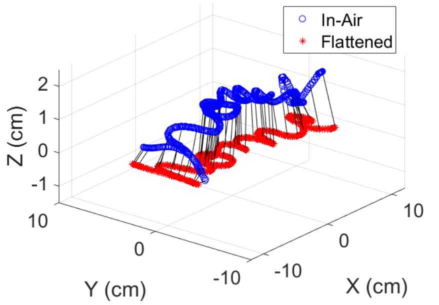


Fig. 19. An example of plane flattening of the word "home".

5.3 Flattening Writing Surface

Scribe aims to develop a natural writing interface on the voice assistants and therefore, does not restrict the orientation of the writing plane or the writing style. The user can simply write in the air near the voice assistant and *Scribe* attempts to produce the writing (or drawing) in a human and machine-readable form. A simple projection of the trajectory on to the horizontal plane can severely squeeze and distort the writing.

For the conversion of the 3D trajectory to 2D writing, we first identify the writing surface from the 3D point cloud of the trajectory points by fitting a third-order surface to them. This smooth surface is the assumed virtual

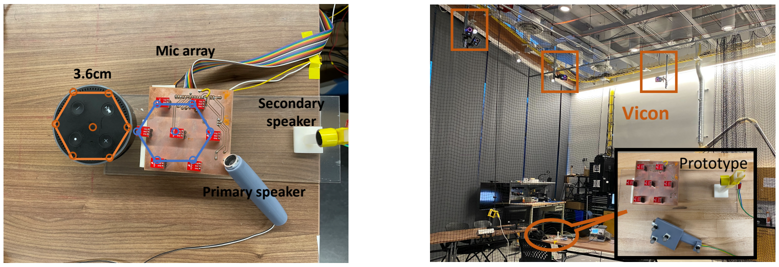


Fig. 20. (a) Hardware setup for system evaluation and (b) Vicon system for ground truth of 3D accuracy.

plane on which the user has written. Next, we take an orthogonal projection of the points from the 3D pen trajectory on this surface. Note that the pen-lift and other stray marks are already removed from this trajectory data in the previous step. The next step is to project this arbitrary writing surface, along with the writings, on the 2D horizontal plane. The curves on the writing surface will require careful flattening to avoid any distortion of the written strokes during projection. We refer to the techniques used in mesh parameterization for this purpose. While there are several techniques for curved surface flattening and transformation [29, 61], we adopted an isometric mapping (Isomap) [65] based approach. Isomap is a non-linear dimensionality reduction technique and several recent techniques [8, 15, 85, 90] build on this core idea to flatten triangularized mesh surfaces. This approach particularly suits our purpose as the algorithm keeps the geodesic distance between the points of the writing unchanged. Although Isomap is not the fastest algorithm for surface flattening, it provides robustness and applies to surfaces with sharp curves that occasionally appear on our unsupported writing surface. Figure 19 shows a flattened surface of word “home” from the original 3-dimensional space to the 2-dimensional plane.

6 IMPLEMENTATION

To validate our implementation of *Scribe*, we build a hardware prototype and data processing pipeline. As shown in Figure 20(a), we use a 7-microphone array placed horizontally on the table and two low-cost($0.5) ultrasound speakers [21]. The size of the microphone array is the same as Amazon Echo, with distances between microphones as 3.6cm. We used omnidirectional ADMP401 MEMS microphones [24] sampled at 16kHz, similar to the components used in the Amazon device [6]. The microphones are sampled simultaneously using a multi-channel data acquisition system [39]. The secondary ultrasound speaker is placed at the same plane as the microphone array, 15cm away from the center. The primary speaker is equipped with a pen-shape stylus or act as a transmitter to project signal in the air. The speakers are driven and synchronized by a Keysight 33500B function generator. The collected data is processed offline using Matlab scripts on a computer. For 3D accuracy evaluation, we use Vicon [67] system as our ground truth, as shown in Figure 20(b).

Table 1. Comparison with prior works on acoustic motion tracking.

System	Setup	Tech.	Audible	1D/3D Accu.	Refresh rate	Range	Mic sep.	Compat.*
CAT [48]	Speaker-Phone	FMCW	N	4mm/9mm	25Hz	7m	90cm	No
SoundTrak [87]	Speaker-Watch	Phase	Y	−/13mm	86Hz	20cm	4cm	No
MilliSonic [70]	Microphone-Phone	FMCW + phase	N	0.7mm/2.6mm	40Hz	3m	6-15cm	No
Scribe	**Alexa-speaker**	**CFCW + phase**	**N**	**0.07mm/1.4mm**	**333Hz**	**70cm**	**3.6cm**	**Yes**

* Column ‘Compat.’ indicates if the solution is compatible with voice interfaces on low sampling rate devices like Amazon Alexa.

7 EVALUATION

Table 1 summarizes the performance of *Scribe* and compares with existing techniques.

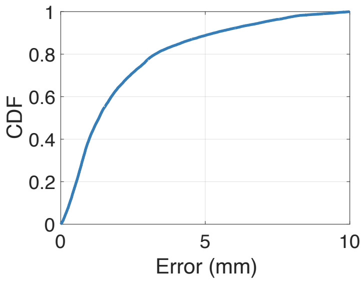


Fig. 21. Location error.

7.1 Location Tracking Performance

7.1.1 3D location tracking accuracy. We measure the 3D localization and tracking accuracy with the distance estimations from 7 microphones in the array and the 3D localization algorithm described in Section 5.1. To get the ground truth, we use the optoelectronic motion capture system of Vicon with 0.2mm tracking accuracy with a frame rate of 250Hz. We also use the same frame rate so that the distance estimations can be matched. The 3D region of detection is $0.5m \times 0.5m \times 0.4m$. The frequency of the primary speaker is 40/42kHz. Figure 21 shows the CDF of 3D localization for *Scribe*. The median error is 1.4mm with a 3.6cm distance between the microphones. We evaluated *Scribe* with the similar microphone array configuration available in off-the-shelf voice assistants. Figure 22 shows some qualitative examples of 3D tracking. Here the red lines indicate the ground truth trajectory captured by the Vicon system, and the blue line is the trajectory estimated with *Scribe*.

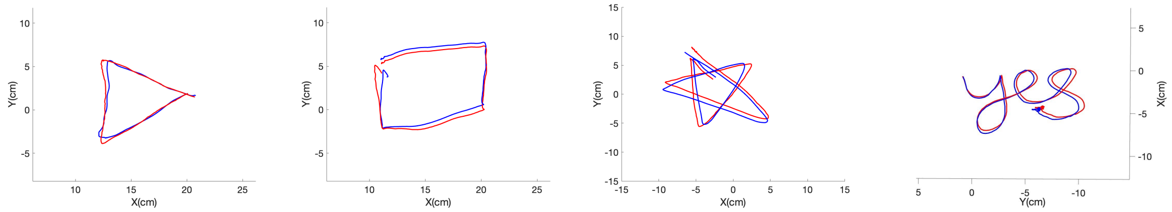


Fig. 22. 3D tracking examples. The red and blue lines represent estimates from Vicon and *Scribe* respectively.

7.2 Performance of Voice Recovery

Scribe is capable of simultaneously recovering voice and localization signals since *Scribe* only occupies 7kHz frequency, minimizing interference to the frequency band of voice. In this section, we evaluate the quality of the recovered voice in the presence of a ranging signal. We apply an existing i-vector-based speech-independent speaker authentication [30] and Google speech-to-text conversion [5] techniques for the evaluation of voice recovery. We first train the speaker authentication model with 9 males and 9 females in the Pitch Tracking Database from Graz University of Technology (PTDB-TUG) [56], where each subject has 236 8-second voice pieces. Then we enroll one male and one female using three voice samples. We use the rest 233 samples for testing. To get the ground truth of the system, we first use the original voice pieces for testing. After that, we play the samples while collecting the 7kHz handwriting tracking signal, and save them for testing. We show the accuracy with 7kHz in Figure 23, the error rate only increases from 6.26% to 8.28%, which indicates the speaker authentication system is robust despite simultaneous reception of handwriting tracking signal. For speech-to-text conversion, we have one male and one female speaking 5 sentences each during the hand movement with and without *Scribe*. We feed the recorded signals to the speech-to-text converter Google Speech Recognition without any change. The word recognition accuracies of converted text messages are both 96.0%. One example of the sentence is "Hi Alexa, I am using air writing to draw a star". The recognized text is "Hi Alexa, I am using hair writing to draw a star".

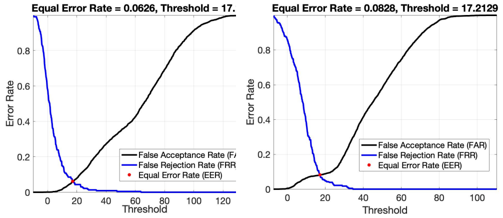


Fig. 23. Speaker authentication without (left) and with (right) the 7kHz handwriting tracking signal.

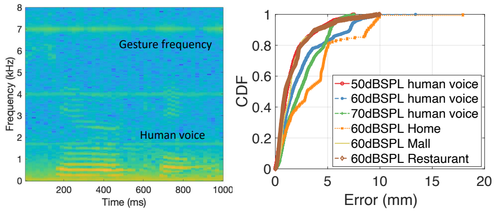


Fig. 24. Spectrogram of the recorded signal under human voice (left) and 3D results under different types of human voices and environmental noises (right).

7.3 Performance Under Human Voice and Environmental Noise

To evaluate the performance with human speaking while writing in the air, we perform the same process of evaluation as the 3D localization. The only difference is we play the noises at the same time. We test the robustness under 50/60/70dBSPL of human voices. Note that the highest strength of human speaking in daily life is 50dBSPL. Even when the sound level is 70dBSPL, the median error of *Scribe* is 2.4mm, as shown in Figure 24(b), meaning the human voice does not affect its accuracy with a 16kHz sampling rate. Figure 24(a) shows an example of the spectrogram of the recorded signal. There is a clear separation between the frequencies of a human voice, environmental noise, and the 7kHz down-converted signal for tracking. We also play three representative types of environmental noises in the home, mall, and restaurant. The sound levels are set as 60dBSPL. The median error of *Scribe* with noises in malls and restaurants are 1.4mm and 1.6mm, respectively. The error under home noise is 3.4mm which is relatively high. The reason is that the frequency of running water has a component of 7kHz.

7.4 Performance of Handwriting Recovery

To verify our accuracy of in-air writing-to-text conversion, we recruit ten participants (2 female and 8 male) to do a small-scale user study. The users were asked to draw 5 words each both on paper and in air separately. Our objective is to compare the accuracy of writing to text conversion with handwriting on paper and in the air. To enable a fair comparison, we attach the primary speaker to an apple pencil. After capturing both the on-paper and in-air writings, we feed them into Google Keep for writing-to-text conversion and compare the accuracy. As shown in Figure 25, we compare the accuracy of each word from the ten users, which are "alexa", "home", "okay", "sign", and "word". The accuracy is the number of correctly detected characters over the number of characters. We find the overall accuracy of *Scribe* and on-paper writing are 94.1% and 96.6%, which means *Scribe* shows comparable accuracy with writing on paper. The accuracy of recognizing Alexa is relatively lower than other words since the way users are writing "x" varies a lot. Some of the "x" is recognized as "n" for some users. To test our performance of handwriting recovery, we write both paragraphs and single words in the air. For the paragraph recovery, we write six representative paragraphs in the air and use Google Keep to convert images from text with an accuracy of 92.4%. In Figure 26, we show two representative examples.

7.5 Performance of Signature Recovery

To test the performance of signature recovery, we apply both human study and existing deep learning-based signature verification models. Our objective is to compare the similarity of the handwritten signatures and in-the-air recovered signatures. Not only the written text can be recognized, but the biometric characteristics in handwritten signature actions can be accurately captured, so that signature verification is possible.

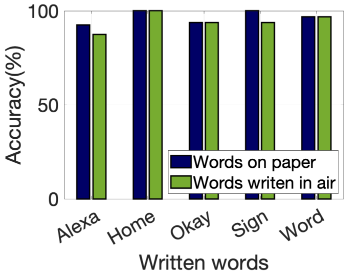


Fig. 25. The accuracy comparison between the image-to-text conversion of on-paper and in-air handwriting.

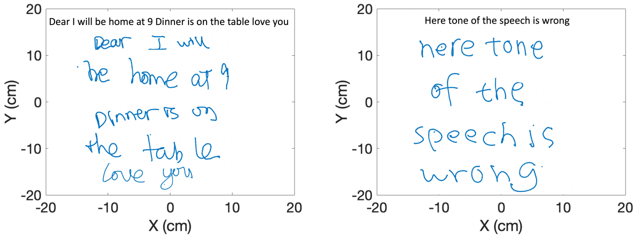


Fig. 26. Examples of paragraph recovery.

For human study verification, we first recruit 20 participants (4 female and 16 male) to sign their signatures both on an iPad using an apple pencil, and in the air. They were asked to sign in a $10 \times 5cm$ block, as small as the electronic signature pads used in stores. The distance from the microphone array to the block is 10cm. Note that these signatures are not their real signatures. To avoid the inconsistency between the two signatures, we attach the primary speaker to the apple pencil. In theory, the primary speaker will move in the air with the same trajectory as on the iPad. In practice, interference also happens since the users may slightly rotate their apple pencil while writing. We treat this interference as unavoidable. After capturing in-air signatures, we recover the raw trajectory, apply pen-lift removal and surface flattening, and save them as images.

7.5.1 Professional signature verification. After capturing both on-tablet and in-air writings, we first ask a signature verification professional to give his opinion on if the 20 pairs of signatures are matched or not. The professional treats 18 out of 20 pairs are matched, resulting in an accuracy of 90%. The strategy of signature matching follows two rules: (1) the structure of the signatures should be the same, including the shapes of the whole signature and each character. (2) The angles of the pen touching and lifting the surface should follow the same trend. We show two examples in Figure 27, including one pair of matched and one pair of not matched signatures. The explanation from the professional for the unmatched pair is that the shape of 'd' is different, and the angle of the pen lift when finishing the signature is not the same.

7.5.2 Opinion of common users. After that, we select five pairs of signatures and post a survey on Amazon Mechanical Turk. 30 people participated in the survey to share their opinions on if the in-air signatures are real or forged. The participants selected their confidence scores ranging from 0 to 10. Here, 0 means they have very low confidence that the pair of signatures are from the same person and 10 means they have very high confidence that the pair of signatures are from the same person. A score of 5 will show that they are 50% sure (in the middle) that the pair of signatures are from the same person. We also give the participants a list of features that are commonly used for signature verification: signature shape and dimensions, pen lifts, shaky handwriting, signs of retouching, letter proportions, letter slants, and smoothness of curves. Figure 28 shows the distribution of the confidence scores from 5 pairs of signatures and 30 participants. Among the 150 ratings, 85.3% of them are more than 50% sure that the pair of signatures are from the same person. This result means *Scribe* not only can capture the recognizable written characters but also can capture precise biometric characteristics of handwritten signature actions, which can be used for signature verification. Several examples of signatures are shown in Figure 29.

7.5.3 Automatic signature verification. Besides human study, we also apply two existing deep learning-based offline signature verification models to compare the similarity of the handwritten and recovered signatures. Signver [26] is a deep learning library for signature verification. This project won the Best Computer Vision

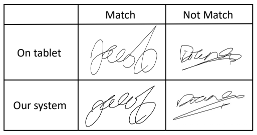


Fig. 27. Examples of matched and not matched signatures.

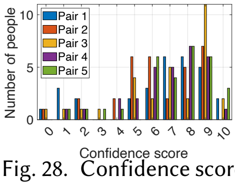


Fig. 28. Confidence scores of the survey with the signatures.

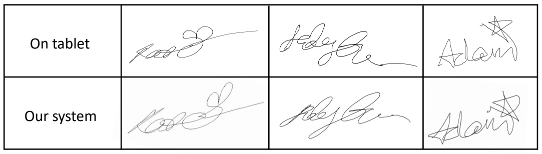


Fig. 29. Examples of signature pairs on iPad tablet (top) and recovered in-air signature (bottom).

Application Award in AWS Deep Learning Challenge. It measures the cosine distance between the extracted features of two signatures to verify if they match or not. The smaller the distance, the more confident the signatures are from the same person. Signet [25] is a well-known deep learning model for offline signature verification. This model uses a convolutional siamese network to envelop minute inconsistency in signatures into the extracted features, and compute the euclidean distance between the features. For both models, the smaller distance between extracted features represents the higher similarity between signatures. We used the 20 pairs of on-paper and in-air signatures from different people as our dataset. After that, we compute the distances between pairs of signatures from the same and different people. As shown in Figure 30, the equal error rates are 0.14 and 0.29, respectively. This result shows the on-paper and in-air signature pairs share a very high similarity. Moreover, we tested the thresholds for signature matching in both the models using their training datasets, which are 0.3 and 28 with equal error rates of 0.12 and 0.23, comparable with the thresholds and equal error rates using our dataset. This result shows that on-paper and in-air signatures have reasonable similarities for signature verification.

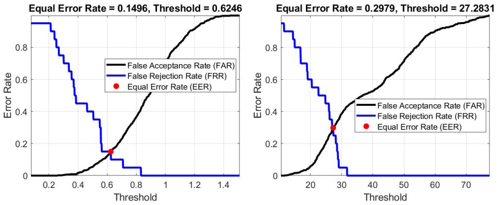


Fig. 30. The accuracy of signature matching using Signver (left) and Signet (right).

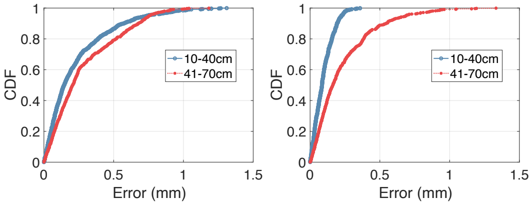


Fig. 31. 1D results for 40/42kHz (left) and 60/62kHz (right) for different distances.

7.6 Impacts of External Conditions

7.6.1 Distance to the microphone. Figure 31 shows the CDF plot of 10-40cm and 41-70cm for 40/42kHz and 60/62kHz primary channel, respectively. The plot shows that the system performs better in the 10-40cm range, with median errors of 0.14mm and 0.07mm for 40/42kHz and 60/62kHz frequencies. It also confirms that higher frequency performs better following the theory. When the distance increases to 41-70cm, the median errors are 0.23mm and 0.24mm. The increase of error is because the SNR of the acoustic signal reduces with distance. Moreover, when the distance increases, 60/62kHz frequency does not perform better than 40/42kHz. The reason is the higher frequency signal attenuates faster than lower frequency signals. Overall, the median error is still within 0.25mm in 41-70cm distance.

7.6.2 Plane shift. We evaluate the performance when writing at different planes, including flat plane on top, flat plane besides, slant plane besides, and vertical plane on top. As shown in Figure 32(a), aside from writing on the flat plane beside the array, the median errors are all within 2mm. The median error of writing on the flat plane beside the array is 3.3mm. The reason is when writing on the same plane with the microphone array, the pattern

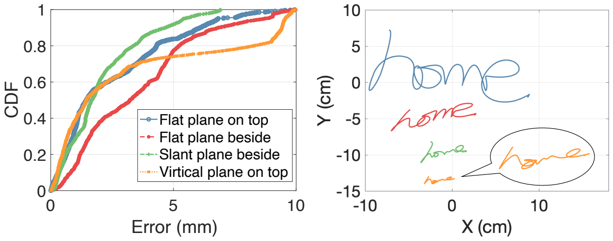


Fig. 32. Performance under different plane shifts (left) and size of written word (right).

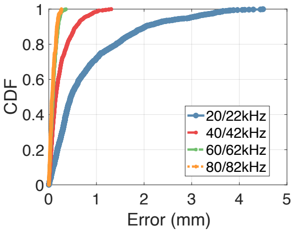


Fig. 33. The CDF of ranging error.

of distance changing is similar for all the microphones. Thus, same distance estimation error will cause a bigger location error.

7.6.3 Size of the word. To test the effectiveness of writing recovery with different sizes, we write the word "home" in widths from 3 to 20cm. As shown in Figure 32(b), we find that even in a width of 3cm (0.75cm per character), we still can recover the word clearly and can be recognized by Google Keep. This result means *Scribe* is applicable for capturing regular sizes of writing for humans.

7.7 Ablation Study

We perform an ablation study on our end-to-end system to evaluate the performance and effectiveness of different components. Here we present an analysis on the operating frequency for the pure-tone-based ranging, the virtual pen-lift detection and the elimination method, the surface flattening and 2-dimensional projection technique, the multipath elimination technique in the context of the cross-frequency sonar, and the system's robustness to Doppler shift. We also measure the individual contribution of different modules on the overall latency of the system.

7.7.1 Performance for different operating frequencies. To evaluate the impact of frequency, we collect data 5 times for each frequency for the primary channel. During this experiment, we maintain a 7kHz separation between the primary and the secondary frequencies. Figure 33(a) shows the cumulative distribution function (CDF) plot of the ranging error with frequencies 20kHz, 40kHz, 60kHz, and 80kHz. The signals are transmitted by transmitters resonate at 20kHz [2], 40kHz [21], 60kHz [3], and 80kHz [4]. Note that there is no non-linearity of the microphone for 20kHz signal, so we directly estimate the distance from the 20kHz frequency. The frequency was hopped between the base frequency and 2kHz higher. This plot shows that the accuracy increases with the higher frequencies, with median errors of 0.48mm, 0.16mm, 0.076mm, and 0.073mm. The accuracy difference between 60kHz and 80kHz is small.

7.7.2 Performance of the Pen-lift elimination module. To evaluate the performance of pen lift removal, we first identify the types of pen lifts: (1) pen lift before starting writing, (2) pen lifts between short lines, (3) pen lift between drawing circles, (4) pen lift when changing the line of writing, (5) pen lift when finish writing, (6) pen lift between characters, (7) pen lift within characters. We design four trajectories to evaluate all types of pen lifts. All the trajectories include types 1&5. The first trajectory includes two short dashes in the first line, and two circles in the second line, for the evaluation of types 2, 3, and 4. To evaluate type 5, we have two words "yes" and "go". Another word "fit" is designed to evaluate type 6. We ask 5 users (4 male and 1 female) to write these trajectories in the air. The plane is 30cm on the top of the microphone array. The accuracy is shown in Figure 34(a). Overall accuracy is higher than 93%. We also show the CDF of the ratio between the number of removed points over the number of points in pen lifts in Figure 34(b). The median value is 97%, only 3% to the 100%.

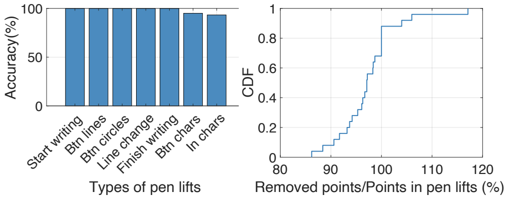


Fig. 34. The accuracy of removing different types of pen lifts.

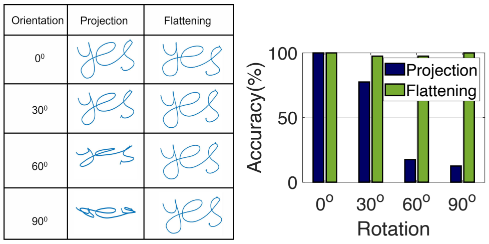


Fig. 35. Evaluation of surface flattening for different orientations of the writing plane.

7.7.3 *Performance of surface flattening.*

When people are writing in the air, they may naturally deviate from the plane of writing. We take samples with various inclinations of the writing plane. After flattening the accuracy for recognition improves. We evaluate this feature by writing various words in different orientations in space and compare the text recognition results using Google Keep before and after flattening with the one written on paper by the same person. We find that when there is an overlap between two characters or characters ex without flattening, they cannot be recognized successfully. For accuracy comparison, we compute the percentage of correctly recognized characters. Figure 35 shows the results of projection and our flattening technique when written at different orientations, and the accuracy plot shows that text recognition accuracy degrades with projection but remains almost the same with flattening in all orientations.

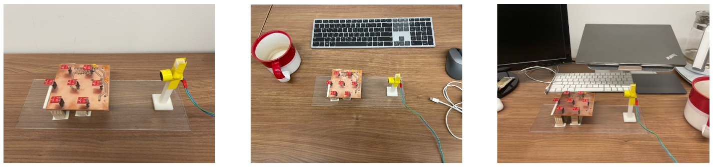


Fig. 36. Scenes of different multipath (a) wall, (b) clutters (3) both clutter and wall.

7.7.4 *Robustness to multipath.*

We consider the common multipath when putting Alexa at different locations in a room, including a wall in behind, in clutter, and both in clutter and a wall in behind. The scenes are shown in Figure 36 and results in Figure 37(a). When there is multipath, the median errors are still within 1.4mm, which means *Scribe* is robust to multipath. To test the effectiveness of the cross-frequency approach for multipath elimination, we do the same experiment with a wall behind, but only using one pair of frequencies. As shown in Figure 37(a), the median error increases to 10mm, meaning our multipath elimination approach is effective.

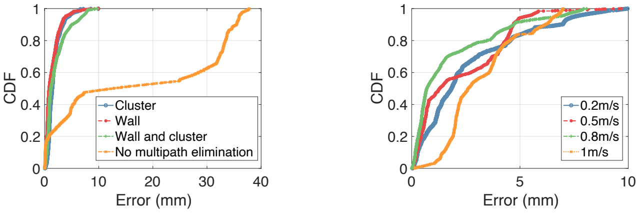


Fig. 37. (a) Comparison of the performance under multipath with and without CFCW for multipath elimination. (b) Performance of *Scribe* with different frequency resolutions under Doppler shift.

7.7.5 Robustness to Doppler shift due to movements. To test the effect of moving speed, we draw three shapes at speeds of $0.2m/s$, $0.5m/s$, $0.8m/s$, and $1m/s$. The frequency band of the primary speaker is 40-42kHz. The maximum frequency shift caused by the Doppler effect is 126Hz. As shown in Figure 37(b), the median errors are all within 1.9mm when the speed is within 0.8m/s. The error increases to 2.6mm when the speed is $1m/s$. This result shows our system is reliable with a $1m/s$ moving speed, which is an extreme case of hand movement in the air.

7.7.6 Latency. Latency plays a crucial role in user experience. We measure the latency of each building block using MATLAB. The latency of start point detection is 2.1 seconds. Note that we suppose we already detected the start point before the user touches the stylus and starts the writing, thus the time length for start point detection does not count as a delay. The time length for detecting a location in 3D is 0.4ms, causing a delay of 0.4ms. The pen lift removal and plane flattening is processed word by word. The latency of pen lift removal is 0.5s and plane flattening is 0.9s. These two modules cause a delay of 1.4 seconds. The latency of the whole system is 1.4s, which is tolerable for applications such as two-factor authentication and 'liveness' detection. For applications that only require raw stylus tracking, the latency is only 0.4ms.

8 DISCUSSION

Here we discuss the potential design implications of the presented work, limitations and future directions.

8.0.1 Frequency vs accuracy vs distance. Although in theory, the frequency values of the ultrasound signals are the higher the better to achieve higher accuracy, the frequency value still has a limit in practice based on the application we want to support. A higher frequency can provide higher accuracy. At the same time, the higher the frequency, the faster the attenuation in the air. Thus, a ranging/localization system should carefully select the frequency values based on both accuracy and distance requirements. For a system that has higher accuracy but lower distance demands, it is applicable to increase the frequency.

8.0.2 Tracking multiple concurrent trajectories. In this paper, we consider the application scenarios in which we only need to track one user. But it is possible to track several users simultaneously, if we carefully select the frequencies of the signals. The first requirement is the down-converted signals should fall into different frequency bins. That is the frequencies of the primary speakers in hand should have a clear gap between each other. Moreover, we need to make sure the ultrasound signals transmitted in the air do not interfere with each other, i.e., the frequencies of the multipath signals do not overlap with the frequencies for gesture recovering. We keep this topic for future work.

8.0.3 e-Signature interface. A signature is basically a stylized form of handwriting, and we are not the first to recognize the application of in-air writing. With the advent of smart devices, several techniques have been proposed to enable a virtual writing interface in the air [48, 53, 70, 87]. While writing recognizable words is possible with these techniques, they are far from being a natural interface for capturing signatures. Unlike written or drawn-on paper or tablet screens, existing in-air writing interfaces cannot capture the individualized style of writing or replicate the precise strokes of a drawing, therefore cannot be used to capture the human signatures. The voice assistants come with a well-designed acoustic frontend including a planar microphone array ideal for localization and spatial analysis of sound. This makes voice assistants a natural interface for users signing their signatures beside them. More than the opportunity to precisely capture the signatures, the IP addresses, voice characteristics, and even encryption can be applied to voice assistants to further secure the e-signature collection and verification process.

9 RELATED WORK

The literature is rich in spatial analysis and localization. We sample below three topics closely related to this work.

9.0.1 Acoustic motion tracking.

Acoustic signals are actively being explored by the research community for precise localization and tracking [10, 11, 14, 16, 31, 32, 35, 76, 86]. Existing studies detect minute body movement to monitor sleep apnea events [51] and breathing for adults [81] and infants [72]. AAMouse [86] applies Doppler shifts of acoustic signals to track hand movements in real-time. CAT [48] further enhances the tracking accuracy by analyzing both FMCW and Doppler shift of acoustic signals. FMCW maps time difference to frequency shift, without the need for precise synchronization. SoundTrak [87] tracks accumulated phase shift for continuous tracking of a speaker in the air. MilliSonic [70] also applies the FMCW signal on hand tracking. To achieve sub-millimeter level accuracy, it tracks the phase shift of FMCW. On the other hand, LLAP [77], FingerIO [53], and DeepRange[49] achieve device-free mm-level finger motion tracking. In our work, we explore the possibility of finer localization that does not interfere with voice interface on low sampling rate voice assistants. The CFCW sonar technique used in *Scribe* enables household acoustic devices to achieve tens of micrometer-level motion tracking accuracy.

9.0.2 RF and inertial sensor-based motion tracking.

We have seen rich literature on developing radio frequency techniques for localization and tracking, such as commercial battery-free tags [36–38, 73, 75, 80, 83] and custom RF backscatter [18, 20, 40, 45, 52, 68, 74]. However, they still cannot outperform acoustic motion tracking due to the tradeoff between latency, frame rate, and motion detection. Systems which can provide centimeter-level results either require a highly constrained environment [73, 83] or have frame rates less than 1 Hz [41, 46] since these systems need to step over hundreds of megahertz of bandwidth and take several seconds to compute a single location estimate. More recently, novel systems have been proposed to significantly reduce the latency of localization systems to tens of milliseconds. However these systems either require using multiple antennas on the backscatter tags [52] or using multiple RF transmitters by scanning over the different spectrum [44]. Inertial sensor-based motion capture is based on miniature intertial sensors, sensor fusion algorithm, and biomechanical models [17, 63, 64]. Inertial sensors can only measure the motions, but not the absolute locations.

9.0.3 Channel non-linearity.

We are not the first to exploit non-linearity to down-convert the frequency of the recorded signal. The notion of exploiting non-linearity was originally studied in 1957 by Westervelt's seminal theory [78, 79]. The closer use of *Scribe* to our System is Backdoor [59] which shows that ultrasound signals can be designed to become recordable by unmodified microphones. More studies use non-linearity of microphone for side-channel attack [28, 43, 57, 89], indoor localization [42], and communication [12, 13]. Different from existing studies applying second order of non-linearity, another work [19] uses the third order term to further boost the vibration sensing granularity. While third order term can further boost the phase change to sense minor movement, it drastically hinds the range of sensing due to the low signal strength.

10 CONCLUSION

This paper develops an acoustic motion tracking interface for voice assistants. *Scribe* proposes a high-resolution CFCW-based distance and tracking estimation algorithm leveraging the nonlinearity of the microphone. Moreover, *Scribe* is the first motion tracking work that can co-exist with voice interface on low sampling rate voice assistants. Evaluations of the prototype show a 1-D ranging error of 73 micro-meter and below 1.4 millimeters of median error in 3D trajectory tracking. This paper presents multiple benchmarks, pilot user studies and evaluation for some specific applications. The design of *Scribe* enables an interface on voice assistants for capturing handwritten notes, drawings, and signatures and can extend to a non-verbal mode of human-machine interaction.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their helpful comments. We thank Nakul Garg for helping with the prototype and other iCoSMoS members for their active participation. The work was partially supported by NSF CAREER Award 2238433 and Meta Research Awards. We also thank the various companies sponsoring the iCoSMoS laboratory at UMD.

REFERENCES

[1] 2021. Roughly 1 in 4 U.S. adults now owns a smart speaker, according to New Report. https://martech.org/roughly-1-in-4-u-s-adults-now-owns-a-smart-speaker-according-to-new-report/

[2] 2022. 20kHz speaker. https://www.digikey.com/en/products/detail/pui-audio,-inc./ASX05408-HD-R/7227653utm_adgroup=Speakers&utm_source=google&utm_medium=cpc&utm_campaign=Shopping_Product_Audio.

[3] 2022. 60kHz ultrasound speaker. https://www.steminc.com/PZT/en/ultrasonic-air-transducer-60-khz.

[4] 2022. 80kHz ultrasound speaker. https://www.steminc.com/PZT/en/ultrasonic-air-transducer-80-khz.

[5] 2022. Google Speech-To-Text API. https://cloud.google.com/speech-to-text.

[6] 2022. PMM-3738-VM1010-R MEMS Microphone. https://www.puiaudio.com/products/pmm-3738-vm1010-r.

[7] 2022. Voice-enabled devices steer the growth of the voice assistant application market as per the Business Research Company's voice assistant application global market report 2022. https://www.globenewswire.com/news-release/2022/02/02/2377858/0/en/Voice-Enabled-Devices-Steer-The-Growth-Of-The-Voice-Assistant-Application-Market-As-Per-The-Business-Research-Company-s-Voice-Assistant-Application-Global-Market-Report-2022.html

[8] Diego A Acosta, Oscar E Ruiz, Santiago Arroyave, Roberto Ebratt, Carlos Cadavid, and Juan J Londono. 2016. Geodesic-based manifold learning for parameterization of triangular meshes. *International Journal on Interactive Design and Manufacturing (IJIDeM)* 10, 4 (2016), 417–430.

[9] Michelle Annett, Fraser Anderson, Walter F. Bischof, and Anoop Gupta. 2014. The Pen is Mightier: Understanding Stylus Behaviour While Inking on Tablets. In *Proceedings of Graphics Interface 2014* (Montreal, Quebec, Canada) *(GI '14)*. Canadian Information Processing Society, CAN, 193–200.

[10] Yang Bai, Nakul Garg, and Nirupam Roy. 2022. Spidr: Ultra-low-power acoustic spatial sensing for micro-robot navigation. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*. 99–113.

[11] Yang Bai, Nakul Garg, and Nirupam Roy. 2022. Ultra-low-power acoustic imaging. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*. 523–524.

[12] Yang Bai, Jian Liu, Yingying Chen, Li Lu, and Jiadi Yu. 2019. Poster: Inaudible High-throughput Communication Through Acoustic Signals. In *The 25th Annual International Conference on Mobile Computing and Networking*. 1–3.

[13] Yang Bai, Jian Liu, Li Lu, Yilin Yang, Yingying Chen, and Jiadi Yu. 2020. BatComm: enabling inaudible acoustic communication with high-throughput for mobile devices. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 205–217.

[14] Yang Bai, Li Lu, Jerry Cheng, Jian Liu, Yingying Chen, and Jiadi Yu. 2020. Acoustic-based sensing and applications: A survey. *Computer Networks* 181 (2020), 107447.

[15] Rafael Moniz Caixeta and João Felipe Coimbra Leite Costa. 2021. A robust unfolding approach for 3-D domains. *Computers & Geosciences* 155 (2021), 104844.

[16] Gaoshuai Cao, Kuang Yuan, Jie Xiong, Panlong Yang, Yubo Yan, Hao Zhou, and Xiang-Yang Li. 2020. Earphonetrack: involving earphones into the ecosystem of acoustic motion tracking. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 95–108.

[17] Yifeng Cao, Ashutosh Dhekne, and Mostafa Ammar. 2021. ITrackU: tracking a pen-like instrument via UWB-IMU fusion. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*. 453–466.

[18] Liqiong Chang, Jie Xiong, Ju Wang, Xiaojiang Chen, Yu Wang, Zhanyong Tang, and Dingyi Fang. 2018. RF-copybook: A millimeter level calligraphy copybook based on commodity RFID. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–19.

[19] Xiangru Chen, Dong Li, Yiran Chen, and Jie Xiong. 2022. Boosting the sensing granularity of acoustic signals by exploiting hardware non-linearity. In *Proceedings of the 21st ACM Workshop on Hot Topics in Networks*. 53–59.

[20] Li-Xuan Chuo, Zhihong Luo, Dennis Sylvester, David Blaauw, and Hun-Seok Kim. 2017. Rf-echo: A non-line-of-sight indoor localization system using a low-power active rf reflector asic tag. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*. 222–234.

[21] Murata Manufacturing Co. 2020. Ultrasound speaker. https://www.murata.com/-/media/webrenewal/products/sensor/ultrasonic/open/datasheet_maopn.ashx.

[22] Andy Cockburn, David Ahlström, and Carl Gutwin. 2012. Understanding performance in touch selections: Tap, drag and radial pointing drag with finger, stylus and mouse. *International Journal of Human-Computer Studies* 70, 3 (2012), 218–233.

[23] Kurt M DeGoede, James A Ashton-Miller, Jimmy M Liao, and Neil B Alexander. 2001. How quickly can healthy adults move their hands to intercept an approaching object? Age and gender effects. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 56, 9 (2001), M584–M588.

[24] Analog Devices. 2013. Admp401: Omnidirectional microphone with bottom port and analog output.

[25] Sounak Dey, Anjan Dutta, J Ignacio Toledo, Suman K Ghosh, Josep Lladós, and Umapada Pal. 2017. Signet: Convolutional siamese network for writer independent offline signature verification. *arXiv preprint arXiv:1707.02131* (2017).

[26] Victor Dibia. 2022. Signver - A deep learning library for signature verification. https://devpost.com/software/signver-a-deep-learning-library-for-signature-verification.

[27] Evgueni Doukhnitch, Muhammed Salamah, and Emre Ozen. 2008. An efficient approach for trilateration in 3D positioning. *Computer communications* 31, 17 (2008), 4124–4129.

[28] Habiba Farrukh, Tinghan Yang, Hanwen Xu, Yuxuan Yin, He Wang, and Z Berkay Celik. 2021. S3: Side-Channel Attack on Stylus Pencil through Sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–25.

[29] Michael S Floater and Kai Hormann. 2005. Surface parameterization: a tutorial and survey. *Advances in multiresolution for geometric modelling* (2005), 157–186.

[30] Daniel Garcia-Romero and Carol Y Espy-Wilson. 2011. Analysis of i-vector length normalization in speaker recognition systems. In *Twelfth annual conference of the international speech communication association.*

[31] Nakul Garg, Yang Bai, and Nirupam Roy. 2021. Microstructure-guided spatial sensing for low-power iot. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services.* 503–504.

[32] Nakul Garg, Yang Bai, and Nirupam Roy. 2021. Owlet: Enabling spatial information in ubiquitous acoustic devices. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services.* 255–268.

[33] Sidhant Gupta, Daniel Morris, Shwetak Patel, and Desney Tan. 2012. Soundwave: using the doppler effect to sense gestures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* 1911–1914.

[34] Andreas Holzinger, Martin Holler, Martin Schedlbauer, and Berndt Urlesberger. 2008. An investigation of finger versus stylus input in medical scenarios. In *ITI 2008-30th International Conference on Information Technology Interfaces.* IEEE, 433–438.

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

[36] Chengkun Jiang, Yuan He, Songzhen Yang, Junchen Guo, and Yunhao Liu. 2019. 3D-OmniTrack: 3D tracking with COTS RFID systems. In *2019 18th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN).* IEEE, 25–36.

[37] Haojian Jin, Jingxian Wang, Zhijian Yang, Swarun Kumar, and Jason Hong. 2018. Rf-wear: Towards wearable everyday skeleton tracking using passive rfids. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers.* 369–372.

[38] Haojian Jin, Jingxian Wang, Zhijian Yang, Swarun Kumar, and Jason Hong. 2018. Wish: Towards a wireless shape-aware world using passive rfids. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services.* 428–441.

[39] Keysight. 2020. https://www.keysight.com/us/en/products/waveform-and-function-generators.html.

[40] Manikanta Kotaru, Pengyu Zhang, and Sachin Katti. 2017. Localizing low-power backscatter tags using commodity wifi. In *Proceedings of the 13th International Conference on emerging Networking EXperiments and Technologies.* 251–262.

[41] Xin Li, Yimin Zhang, and Moeness G Amin. 2009. Multifrequency-based range estimation of RFID tags. In *2009 IEEE International Conference on RFID.* IEEE, 147–154.

[42] Qiongzheng Lin, Zhenlin An, and Lei Yang. 2019. Rebooting ultrasonic positioning systems for ultrasound-incapable smart devices. In *The 25th Annual International Conference on Mobile Computing and Networking.* 1–16.

[43] Yihao Liu, Kai Huang, Xingzhe Song, Boyuan Yang, and Wei Gao. 2020. MagHacker: eavesdropping on stylus pen writing via magnetic sensing from commodity mobile devices. In *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services.* 148–160.

[44] Zhihong Luo, Qiping Zhang, Yunfei Ma, Manish Singh, and Fadel Adib. 2019. 3D Backscatter Localization for {Fine-Grained} Robotics. In *16th USENIX Symposium on Networked Systems Design and Implementation (NSDI 19).* 765–782.

[45] Yunfei Ma, Xiaonan Hui, and Edwin C Kan. 2016. 3D real-time indoor localization via broadband nonlinear backscatter in passive devices with centimeter precision. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking.* 216–229.

[46] Yunfei Ma, Nicholas Selby, and Fadel Adib. 2017. Minding the billions: Ultra-wideband localization for deployed rfid tags. In *Proceedings of the 23rd annual international conference on mobile computing and networking.* 248–260.

[47] I Scott MacKenzie, Abigail Sellen, and William AS Buxton. 1991. A comparison of input devices in element pointing and dragging tasks. In *Proceedings of the SIGCHI conference on Human factors in computing systems.* 161–166.

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

[49] 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.

[50] Seyedali Mirjalili. 2019. Genetic algorithm. In *Evolutionary algorithms and neural networks*. Springer, 43–55.

[51] Rajalakshmi Nandakumar, Shyamnath Gollakota, and Nathaniel Watson. 2015. Contactless sleep apnea detection on smartphones. In *Proceedings of the 13th annual international conference on mobile systems, applications, and services*. 45–57.

[52] Rajalakshmi Nandakumar, Vikram Iyer, and Shyamnath Gollakota. 2018. 3d localization for sub-centimeter sized devices. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*. 108–119.

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

[54] Sarah Perez. 2019. Report: Voice assistants in use to triple to 8 billion by 2023. https://techcrunch.com/2019/02/12/report-voice-assistants-in-use-to-triple-to-8-billion-by-2023/

[55] Domenico Prattichizzo, Leonardo Meli, and Monica Malvezzi. 2015. Digital handwriting with a finger or a stylus: a biomechanical comparison. *IEEE transactions on haptics* 8, 4 (2015), 356–370.

[56] Signal Processing and Speech Communication Laboratory. 2019. Pitch Tracking Database. https://www.spsc.tugraz.at/databases-and-tools/ptdb-tug-pitch-tracking-database-from-graz-university-of-technology.html.

[57] Soundarya Ramesh, Rui Xiao, Anindya Maiti, Jong Taek Lee, Harini Ramprasad, Ananda Kumar, Murtuza Jadliwala, and Jun Han. 2021. Acoustics to the Rescue: Physical Key Inference Attack Revisited. In *30th USENIX Security Symposium (USENIX Security 21)*. 3255–3272.

[58] Nirupam Roy, Haitham Hassanieh, and Romit Roy Choudhury. 2017. BackDoor: Making Microphones Hear Inaudible Sounds. In *ACM MobiSys*.

[59] Nirupam Roy, Haitham Hassanieh, and Romit Roy Choudhury. 2017. Backdoor: Making microphones hear inaudible sounds. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. 2–14.

[60] Wenjie Ruan, Quan Z Sheng, Lei Yang, Tao Gu, Peipei Xu, and Longfei Shangguan. 2016. AudioGest: enabling fine-grained hand gesture detection by decoding echo signal. In *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing*. 474–485.

[61] Alla Sheffer, Emil Praun, Kenneth Rose, et al. 2007. Mesh parameterization methods and their applications. *Foundations and Trends® in Computer Graphics and Vision* 2, 2 (2007), 105–171.

[62] Ann M Swartz, Scott J Strath, DAVID R BASSETT, WILLIAM L O'BRIEN, George A King, and Barbara E Ainsworth. 2000. Estimation of energy expenditure using CSA accelerometers at hip and wrist sites. *Medicine & Science in Sports & Exercise* 32, 9 (2000), S450–S456.

[63] Synertial. 2021. Synertial Motion Capture. https://www.synertial.com/.

[64] Xsens Technologies. 2021. Xsens Motion Capture. https://www.xsens.com/.

[65] Joshua B Tenenbaum, Vin de Silva, and John C Langford. 2000. A global geometric framework for nonlinear dimensionality reduction. *science* 290, 5500 (2000), 2319–2323.

[66] Huawei Tu, Xiangshi Ren, and Shumin Zhai. 2012. A comparative evaluation of finger and pen stroke gestures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1287–1296.

[67] Vicon Motion Systems Ltd UK. 2022. Vicon Motion Systems. https://www.vicon.com/.

[68] Deepak Vasisht, Guo Zhang, Omid Abari, Hsiao-Ming Lu, Jacob Flanz, and Dina Katabi. 2018. In-body backscatter communication and localization. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*. 132–146.

[69] Daniel Vogel and Patrick Baudisch. 2007. Shift: A Technique for Operating Pen-Based Interfaces Using Touch. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 657–666. https://doi.org/10.1145/1240624.1240727

[70] Anran Wang and Shyamnath Gollakota. 2019. Millisonic: Pushing the limits of acoustic motion tracking. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.

[71] Anran Wang, Dan Nguyen, Arun R Sridhar, and Shyamnath Gollakota. 2021. Using smart speakers to contactlessly monitor heart rhythms. *Communications biology* 4, 1 (2021), 1–12.

[72] 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.

[73] Jue Wang and Dina Katabi. 2013. Dude, where's my card? RFID positioning that works with multipath and non-line of sight. In *Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM*. 51–62.

[74] Jingxian Wang, Chengfeng Pan, Haojian Jin, Vaibhav Singh, Yash Jain, Jason I Hong, Carmel Majidi, and Swarun Kumar. 2019. Rfid tattoo: A wireless platform for speech recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (2019), 1–24.

[75] Jingxian Wang, Junbo Zhang, Ke Li, Chengfeng Pan, Carmel Majidi, and Swarun Kumar. 2021. Locating everyday objects using nfc textiles. In *Proceedings of the 20th International Conference on Information Processing in Sensor Networks (co-located with CPS-IoT Week 2021)*. 15–30.

[76] Mei Wang, Wei Sun, and Lili Qiu. 2021. {MAVL}: Multiresolution Analysis of Voice Localization. In *18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21)*. 845–858.

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

[78] Peter J Westervelt. 1951. The theory of steady forces caused by sound waves. *The Journal of the Acoustical Society of America* 23, 3 (1951), 312–315.

[79] Peter J Westervelt. 1957. Scattering of sound by sound. *The Journal of the Acoustical Society of America* 29, 2 (1957), 199–203.

[80] Fu Xiao, Zhongqin Wang, Ning Ye, Ruchuan Wang, and Xiang-Yang Li. 2017. One more tag enables fine-grained RFID localization and tracking. *IEEE/ACM Transactions on Networking* 26, 1 (2017), 161–174.

[81] Xiangyu Xu, Jiadi Yu, Yingying Chen, Yanmin Zhu, Linghe Kong, and Minglu Li. 2019. Breathlistener: Fine-grained breathing monitoring in driving environments utilizing acoustic signals. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. 54–66.

[82] Tomer Yanay and Erez Shmueli. 2020. Air-writing recognition using smart-bands. *Pervasive and Mobile Computing* 66 (2020), 101183.

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

[84] Kiwon Yeom, Jounghuem Kwon, JooHyun Maeng, and Bum-Jae You. 2015. [POSTER] Haptic Ring Interface Enabling Air-Writing in Virtual Reality Environment. In *2015 IEEE International Symposium on Mixed and Augmented Reality*. IEEE, 124–127.

[85] Mahwish Yousaf, Tanzeel U Rehman, and Li Jing. 2020. An extended Isomap approach for nonlinear dimension reduction. *SN Computer Science* 1, 3 (2020), 1–10.

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

[87] Cheng Zhang, Qiuyue Xue, Anandghan Waghmare, Sumeet Jain, Yiming Pu, Sinan Hersek, Kent Lyons, Kenneth A Cunefare, Omer T Inan, and Gregory D Abowd. 2017. Soundtrak: Continuous 3d tracking of a finger using active acoustics. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 1–25.

[88] 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.

[89] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. 2017. Dolphinattack: Inaudible voice commands. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*. 103–117.

[90] Kun Zhou, John Synder, Baining Guo, and Heung-Yeung Shum. 2004. Iso-charts: stretch-driven mesh parameterization using spectral analysis. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*. 45–54.