STATION: Gesture-Based Authentication for Voice Interfaces

Sungbin Park, Xueqiang Wang, Kai Chen, *Member, IEEE*, and Yeonjoon Lee

***Abstract*—The popularity of smart home devices has led to an increase in security incidents happening in smart homes. A key measure to avoid such incidents is to authenticate users before they can interact with smart devices. However, current methods often require additional hardware. This article proposes STATION, a gesture-based authentication system, an effective gesture-based authentication method built on top of the voice interfaces already available in these smart home devices, without adding new hardware. STATION uses a gesture processing pipeline that identifies Doppler-existing frames and detects the direction of arrival of Reflection to authenticate users in low SNR environments and at longer distances. Furthermore, regarding the nature of gesture-based authentication, this system also supports detecting user liveness, preventing replay and synthesis attacks from remote attackers. The evaluation of STATION shows high accuracy with a false acceptance rate (FAR) of 0.08% and false rejection rate (FRR) of 3.10% for users within 1.5 m of the device.**

***Index Terms*—Acoustic sensing, device security, gesture-based authentication, low cost sensors and devices, security and privacy, sensor signal processing.**

I. INTRODUCTION

THE PROLIFERATION of smart home devices has also brought more security attacks targeting them. A key feature to prevent such attacks is always to authenticate users before they can actually use the devices. Due to the lack of an authentication-friendly input interface, existing authentication solutions for smart home devices are often built on additional hardware, such as a paired smartphone or a fingerprint sensor, which presents usability issues when the hardware is not nearby or runs out of battery, etc.

To address the potential issues, previous studies [1], [2], [3] have explored the feasibility of using voice interfaces (e.g., speakers and microphones) in *challenge-response*-based authentication, given that voice interfaces are commonly adopted by smart home devices. For example, W3C working group [1] and Carlini et al. [2] discussed the concept of audio CAPTCHAs where the challenge is composed of audio data that is difficult for machines to understand but can be easily recognized by humans. Diao et al. [3] proposed a voice fingerprint-based method based on wake-up commands. However, as evidenced by prior studies, the above methods are vulnerable to replay attacks [4], [5], and attacks that are based upon signal processing [6] and adversarial audio examples [7], etc.

In this article, we refer to smart home devices that have voice interfaces as VIF devices, such as smart speakers and smart cameras. We propose STATION, a new authentication method for the common VIF devices, through which VIF device users can enter their authentication credentials using hand gestures, similar to password- or PIN-based authentication methods. STATION works at a much further distance (∼1.5 m) than previous solutions and achieves false acceptance rate (FAR) of 0.08% and false rejection rate (FRR) of 3.10% at that distance. It also proves to be robust against remote attacks because of the challenges for attackers to mimic the features with a single compromised device, e.g., the movement and direction of the user's hand.

Specifically, STATION is based on the ability of microphones to detect human motion by measuring the *Doppler* effect from reflected signals. Essentially, we assume that four virtual buttons are placed in different directions around the VIF devices, which the device user can push and pull by hand to complete device authentication, just like she does on the physical buttons. To implement the virtual buttons in STATION, we first use the speaker on the smart device to emit an ultrasound signal. The device user then makes hand gestures with predefined push-then-pull (PTP) movement on the virtual buttons. After that, microphones on-device receive the reflected signal and analyze the *Doppler* effect. In this step, we extract the *Direction of Arrival* (DoA) *of Reflection* and *Hand Movement* from the received signal and authenticate the users by verifying the sequence of hand gestures.

Manuscript received 6 June 2023; revised 26 December 2023, 23 February 2024, and 8 March 2024; accepted 23 March 2024. Date of publication 28 March 2024; date of current version 7 June 2024. This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation [IITP, Artificial Intelligence Convergence Innovation Human Resources Development (Hanyang University ERICA)] funded by the Korea Government (MSIT) under Grant RS-2022-00155885, and in part by the National Research Foundation of Korea (NRF) funded by the Korea Government (MSIT) under Grant NRF-2022R1F1A1074999 and Grant NRF-2020R1G1A1102722. The work of Xueqiang Wang was supported in part by NSF under Grant OAC-2320974. *(Corresponding author: Yeonjoon Lee.)*

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

Sungbin Park is with the Department of Computer Science and Engineering, Major in Bio Artificial Intelligence, Hanyang University, Ansan 15588, South Korea (e-mail: pbt98@hanyang.ac.kr).

Xueqiang Wang is with the Department of Computer Science, University of Central Florida, Orlando, FL 32816 USA (e-mail: xueqiang.wang@ucf.edu).

Kai Chen is with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100012, China, and also with the School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: chenkai@iie.ac.cn).

Yeonjoon Lee is with the College of Computing, Hanyang University, Ansan 15588, South Korea (e-mail: yeonjoonlee@hanyang.ac.kr).

Digital Object Identifier 10.1109/JIOT.2024.3382721

To authenticate device users in a noisy indoor environment, STATION utilizes several new techniques. First, since the buttons are located in different directions of the device, we need to accurately identify the DoA of the reflected signals. This is done by applying a sound source localization (SSL) algorithm based upon GCC-PHAT-β, which can localize the source of sound in a low signal-to-noise ratio (SNR) environment. Second, in addition to PTP gestures, users may make redundant gestures and create noises when moving between different buttons. Therefore, we must robustly separate the PTP gestures from redundant gestures before we can determine exactly which buttons are pressed. To address the problem, we detect the signal frames with a strong Doppler effect (i.e., Doppler-existing frames) using a new density-based clustering method based on the observation that the standard deviation (STD) of Doppler-existing frames that are distinctive from other frames.

In particular, STATION enhances security and privacy in several ways. STATION recognizes the gesture with *DoA of Reflection* which is extracted from *time Difference of Arrival* (TDoA). This method effectively closes the door for remote adversaries as it is challenging for the adversaries to mimic the hand direction information. Even an adversary that can compromise a smart speaker in the victim's house and use the speaker to synthesize adversarial signals would not be able to mimic the hand movement information because of the presence of TDoA.

We implemented the prototype of STATION using low-cost hardware modules, including the Omnidirectional speaker and MATRIX Voice. Next, we evaluated the effectiveness and robustness of STATION by addressing three research questions. First, what are the overall FAR and FRR of STATION (RQ1)? Second, how robust is STATION to environmental noise and how does it fare with different characteristics of end users (RQ2)? Third, how does STATION compare to other authentication methods for smart home (or IoT) devices in terms of common criteria such as deployability, usability, accessibility, and security & privacy (RQ3)? Specifically, we answer the first two research questions (RQ1 and RQ2) by enrolling onsite participants to validate their gesture sequences in the prototype under various settings. We address the third research question (RQ3) empirically by comparing STATION to other authentication methods based on a set of well-established criteria from prior studies. The results indicate that STATION is highly effective, with a FAR of 0.08% and a FRR of 3.10% in recognizing gesture sequences within a range of 1.5 m. It also demonstrates robustness and is less impacted by environmental noise and human factors, such as height and handedness. Moreover, compared to other authentication methods, STATION potentially offers better deployability and usability, primarily because it does not require the installation and carrying of additional hardware. It is also more privacy-preserving since the hand gestures used are analogous to passwords rather than biometric data.

In summary, we make the following contributions.

1) We propose a new authentication method, STATION, for smart home devices with voice interfaces (i.e., VIF devices). Leveraging the motion detection capabilities of voice interfaces, STATION generates highly usable virtual buttons, enabling users to authenticate themselves using hand gestures. Also, STATION enhances the security of smart home devices against known attacks, including physical and remote attacks.
2) We implement a prototype of STATION on low-cost and commercial smart devices. The evaluation on this prototype shows that STATION can achieve a FAR of 0.08% and FRR of 3.10% with over 3,000 samples from 11 real participants, and is robust against environmental noise and noise caused by human factors.

The remainder of this article is organized as follows. Section II provides background knowledge of our research. Appendix presents our online surveys about how users interact with VIF devices. Section III introduces the system overview of STATION. Section IV illustrates the system design and implementation of STATION. Section V reports our evaluation of STATION. Section VI discusses the limitation and future work of this research, and Section VIII concludes this article.

II. BACKGROUND

Sound Navigation and Ranging (Sonar): Sonar is a technique that uses sound propagation to measure distance and navigate. There are two types of *Sonar*. One is *Active Sonar* which emits pulses of sound and listens and analyzes echoes of the sound to extract features. The other is *Passive Sonar*, which listens to the sound made by the target and analyzes it. In STATION, we take advantage of *Active Sonar*: a VIF device authenticates its users by emitting ultrasound and verifying user gestures by analyzing reflected signals from the users' hand (see Section III).

DoA: Acoustic source localization techniques determine the source *location* of an acoustic signal, which is essential for acoustic-based authentication. Since it is hard to measure the exact location of a moving hand, we measure another localization factor–the direction of the sound source, i.e., DoA. There are typically two ways to calculate DoA. The first way is the Multiple Signal Classification algorithm (MUSIC) [8]. MUSIC is based on the eigenvalue decomposition of the sensor covariance matrix observed at an array. MUSIC decomposes the spatial space into signal and noise subspaces with a covariance matrix. Then, it calculates DoA with orthogonality between noise and arrival vectors. The second way is the *SRP-PHAT* [9]. *SRP-PHAT* first estimates the TDoA of the signals. Then, it calculates the largest value of the sum of estimated cross-correlation score corresponding to TDoA as DoA among the predefined discrete points around the sensor array. In this study, we detect the direction of reflected acoustic signals using an approach based on SRP-PHAT, because it incurs less computation overhead than MUSIC-based techniques [10] and works well for wide-band signals.

TDoA: TDoA measures the difference between the time-of-arrival (ToA) of signals, which is commonly used in real-time locating systems. Given that it is challenging to accurately calculate TDoA because of the spatial ambiguity [11], different techniques have been proposed to *estimate* the TDoA. A commonly used technique is generalized cross-correlation

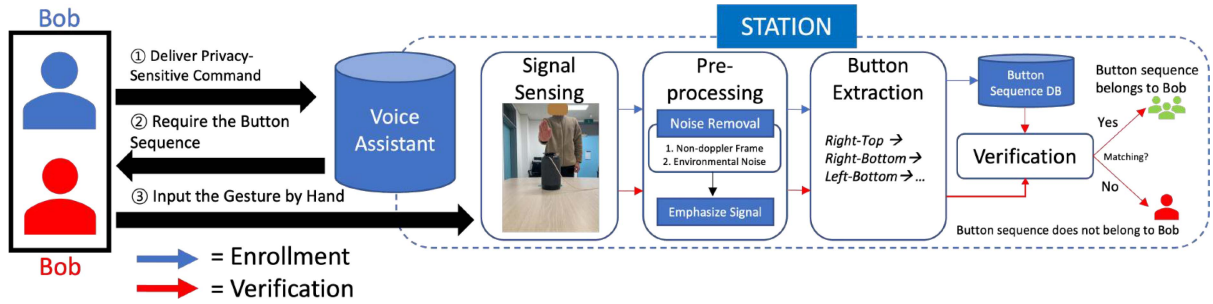


Fig. 1. Overview of STATION.

phase transform (GCC-PHAT) [12], which estimates the time delay between a signal and a reference signal by identifying the peak location of the cross-correlation between the signals. The disadvantage of GCC-PHAT is that it whitens the impact of the magnitude for each frequency bin and only focuses on the phase information of the signals, which limits its use to relatively low-noise environments [13]. To make the technique work in high-noise (i.e., low SNR) environments, GCC-PHAT-β [14] introduces a weighting factor to GCC-PHAT to account for the magnitude for each frequency bin.

III. STATION OVERVIEW

Stages of STATION*:* STATION is essentially a password-like authentication method for VIF devices. It allows users to input an authentication password by clicking a sequence of virtual buttons in the form of hand gestures. Therefore, similar to password-based authentication, STATION has two stages: 1) enrollment and 2) verification. Fig. 1 provides an overview of these stages.

To begin using STATION, users must enroll in the hand gestures during the out-of-box experience (OOBE) of their VIF device or on the first boot after system flashing. Once the user's gestures are received, the VIF device will process the gestures to extract the sequence of virtual buttons, which is then encrypted and stored in the device's internal storage. Additionally, STATION supports the ability to change the sequence. To do so, users simply need to enter their current sequence by gesture and input the new sequence in the device's settings, similar to changing a password or pattern on a mobile device [15].

In the verification stage, the user may authenticate himself to the VIF device by entering the same sequence as the enrolled one through his hand gestures. As shown in Fig. 1, the VIF device may ask the user to authenticate under various situations according to its security needs, e.g., before performing sensitive operations such as opening a smart lock or making online purchases, after device restart, or if the device hasn't authenticated the user for more than 48 h [15], [16]. For such cases, the VIF device may notify the user of the start of the authentication process by blinking LEDs, or playing a notification message via its speakers, e.g., "start authentication by making gestures."

Note that manufacturers can set the minimum length of button sequence with their security policies, similar to forcing the minimum length of the password [17]. Additionally, during enrollment, STATION records the total time it takes for users to press the virtual buttons. This information is used to set the timeout criterion. If users make a mistake while entering their authentication sequence during the verification stage, they will need to wait until the timeout and press the whole buttons again.

Challenges: There are several challenges that can affect the accuracy, usability, and robustness of STATION.

1) *Sensing Doppler Effect Reflected by Hand Gestures:* STATION distinguishes between different hand gestures by evaluating the Doppler effect of the reflected signals from the user's hand. However, accurate sensing of the Doppler effect from a distance is challenging due to signal attenuation, which reduces the magnitude of the shifted frequency. Also, the ultrasound emitted by the embedded speakers has inconsistent frequency magnitudes, leading to variations in the magnitude of reflected signals. In addition, gestures made by different users will also result in varying frequency shifts and magnitudes. Therefore, existing methods [18], [19] that rely on a static threshold to detect the Doppler effect are not effective, and new methods need to be designed to ensure accurate sensing of the Doppler effect (and gesture identification).
2) *Designing a Gesture That Balances Usability and Accuracy:* Slow and small hand movements cause minor signal reflections, which reduces the accuracy of gesture identification of STATION. Fast and big gestures result in more significant signal reflections, but users are difficult to make these kinds of gestures. Therefore, it is important to design hand gestures that can balance the accuracy and usability of STATION.
3) *Identifying Hand Gestures From Low SNR Signals:* The reflected signals that deliver hand gesture information are of low SNR, e.g., mixed with emitted ultrasound and noises from reverberation, etc. This will greatly affect the accuracy of hand gesture identification, because a high noise floor makes it difficult to differentiate the gesture signals from noises. Therefore, more robust methods are required to improve the handling of low SNR signals.

Gestures, and Gesture Processing Pipeline: Similar to password-based authentication that allows users to verify

Algorithm 1 STATION Procedure Algorithm

Input: PTP Gestures
Output: Authentication State

```
procedure STATION
Signal sensing start
    Emit acoustic signal
    Record raw signal reflected by PTP gestures
Signal sensing end
Pre-processing start
    Transforming the raw signal into frequency domain
    Distinguishing Doppler-existing frame
    Environment noise reduction
    Signal emphasis
Pre-processing end
Button extraction start
    Calculate DoA of Reflection
    Filter PTP gestures
    Button recognition
Button extraction end
  if Phase is Enrollment then
    Button Sequence DB ← Inputted sequence
    Authentication State = Enrolled
  else
    if Inputted sequence == Enrolled sequence then
      Authentication State = Pass
    else
      Authentication State = Reject
    end if
  end if
return Authentication State
end procedure
```

themselves by clicking a few keys (buttons) on screen, we design a set of unique hand gestures using the concept of *virtual buttons*. Essentially, we label different areas as virtual buttons according to their relative positions (e.g., direction) to the VIF devices, and hand gestures made in these areas are considered pushing the corresponding virtual buttons. We further define how to push virtual buttons using human-friendly [20] and easily-detectable hand gestures–*PTP* gestures (Section IV-A). For each enrollment and verification session, we ask the user to make a PTP gesture to indicate the base position of his hand (B-PTP), followed by gestures for authentication (A-PTP). Leveraging the virtual button concept and PTP gestures, the user can enter the password-like sequence (i.e., the button sequence) into the VIF device, and the sequence can be enrolled or verified in the VIF device like a password.

As shown in Fig. 1 and Algorithm 1, supporting the enrollment and verification of the above PTP gestures is a pipeline that captures, analyzes, and identifies the gestures from reflected signals of the user's hand. Specifically, the pipeline has three phases. The first phase is *signal-sensing phase*. After the VIF device notifies the user to start the authentication process, the embedded speaker on the VIF device will actively and continuously send out acoustic signals (until the authentication process is completed). The user will make a sequence of PTP gestures (i.e., gestures to be enrolled and for authentication) toward the VIF device. In this phase, the microphone array in the VIF device will record the signals reflected from the PTP gestures. An important decision made in this phase is about the acoustic signals sent by the speaker: we select a specific frequency range of inaudible sound to minimize the impact on the user and design signals with low-crest factor to make the signal resistant to environment noise (Section IV-B).

The next *signal preprocessing* phase takes as input the recorded signals and removes the noise irrelevant to hand gestures using a series of techniques. A key observation is that signals reflected from hand gestures usually have a frequency shift (i.e., Doppler effect) compared to the signal emitted by the embedded speaker. Therefore, we can first transform the recorded signals into the frequency domain [using short time Fourier transform (STFT)]. Then, for each frame, we can determine whether it has the Doppler effect by distinguishing the frequency shift patterns (Section IV-C). Those frames that do not have the Doppler effect are considered noise frames. Finally, we reduce the environmental noise by performing spectral subtraction on the frequency domain signal, and then we emphasize the signal. The outputs of the signal preprocessing phase are Doppler signals with low noise. A key challenge for previous studies [18], [19] that use static threshold for identifying Doppler shift is that, from a relatively further distance (e.g., 1.5 m), the magnitude of the Doppler signals are much smaller than the noise signals (including the emitted signals). To solve this problem, we design a new and more robust method for identifying Doppler shift by clustering the variation of frequency magnitudes for all signal frames (Section IV-C).

The last phase of the pipeline is the *button extraction* phase. In this phase, we first identify the signals that represent PTP gestures by identifying pairs of the higher frequency shift (i.e., the shift caused by pushing toward the VIF device) and the lower frequency shift (i.e., the shift caused by pulling from the VIF device). The first identified PTP gesture is taken as B-PTP, which indicates the base position (e.g., direction) of the user's hand. Next, the following A-PTP gestures are used to determine which virtual buttons were pressed by calculating their relative positions to the base position. The output of this phase is a series of hand gestures represented by virtual button presses.

The system in the VIF device is responsible for managing the behaviors in both the enrollment and verification stages, after extracting the button sequence. During the enrollment stage, the VIF device saves the button sequence in its internal storage. On the other hand, during the verification stage, the VIF device verifies the identity of the user by comparing the button sequence to the enrolled sequence. It is important to note that while designing STATION, we primarily focused on gesture processing pipelines rather than storing and verifying the sequence. This is because the sequence used in STATION is similar to a password, which can be addressed using the common standards and practices that exist well in the industry [15], [16], [21].

Scope and Threat Model: This study aims to introduce device authentication to common VIF devices available in user households. These devices include voice assistants (VAs) such as Alexa and Google Home, but also other devices that have embedded speakers and microphone arrays, such as a smart home robot.

1) *Threats Taken Down by* STATION*:* Many VIF devices do not support user authentication due to them being headless and thus difficult to apply authentication [22], or the assumption that VIF devices in the households are always used by trusted parties [23], [24]. However, lack of authentication can create security and privacy concerns given that burglars, which are in millions each year [25], and other short-term visitors can freely interact with the devices without notifying device owners, e.g., asking the devices for sensitive information and set up unwanted device schedules. STATION provides an effective solution for device owners to block such unauthorized accesses. Some VIF devices have started providing weak authentication, such as Voice ID [26], which are reported to be vulnerable to remote attacks that replay recorded or synthesized signals from other compromised VIF devices [27], [28]. We propose STATION to tackle this problem. On the one hand, STATION allows users to "press" a set of virtual buttons using carefully designed hand gestures and uses such button presses to verify the identity of the users. This practice is similar to password- and pattern-based authentication methods that require the user to enter his credentials. We rely on the confidentiality of the credentials (i.e., virtual button presses) rather than the unique gesture signatures for authentication. On the other hand, the inherent design of STATION also makes it capable of determining the liveness of users: it is difficult for remote attackers to spoof STATION, considering that it is challenging to imitate hand position/movement and direction information by signal synthesis without physical presence. STATION, with these advantages, becomes effective in protecting the device even if it is stolen by an adversary. Upon attempting to use the device, the adversary (i.e., thief) must guess the correct sequence of enrolled gestures, much like guessing a correct password for device unlock. For example, a gesture sequence with a length of n will have a "password space" of size 4^n. This large space poses a significant challenge for the adversary to guess, not to mention the other tactics we can deploy to thwart guesses, such as limiting the number or frequency of attempts like using exponential backoff [15].
2) *Other Assumptions:* We assume that STATION is operating in indoor environments, and adversaries do not have visibility to authentication gestures. We assume that adversaries have enough knowledge and technical capability to synthesize signals in order to target VIF devices with remote attacks. In addition, we assume that the devices are placed in houses following the common placement guideline [29], e.g., at least 8 In from the wall and not near the corner or beside noisy appliances (which can block the signals that are emitted by or flow into the smart home devices).

IV. STATION SYSTEM DESIGN

In this section, we elaborate on the design of STATION, including the selection of hand gestures, signals, and the pipeline to analyze reflected signals and identify hand gestures from them.

A. Gesture Design

An authentication solution should provide easy ways for device users to enter input and for devices to capture the input. This requirement can easily be met by traditional authentication solutions such as password and PIN but is challenging for hand gestures, e.g., users may make hand gestures that trigger very minor signal reflections, which complicates gesture capturing for VIF devices. Therefore, we need to design hand gestures that are not only easy for users to make but also can be effectively captured by VIF devices.

To meet the above needs, we choose the common *push-the-pull* (PTP) gestures. These gestures are human-friendly [20], and commonly used in daily life, such as opening/closing doors, using a gas pump, etc. Therefore, such gestures are easy for users to make and require less effort to learn. Also, the PTP gestures can introduce strong signal reflection, because, according to (1), the faster the users' hand moves toward VIF devices, the larger the frequency shift in the reflected signals (i.e., Doppler effect). Fig. 2(a)–(d) show a PTP gesture.

In this study, once the VIF devices indicate the start of authentication (by playing a command or LED), the users need to make a series of PTP gestures: a PTP gesture that is used to tell the base position of his hand (B-PTP), and followed by other gestures for authentication (A-PTP). The A-PTP gestures can be made in different areas according to their relative positions (e.g., direction) to the B-PTP gesture. We call such areas the *virtual buttons*. As shown in Fig. 2(e)–(h), there are four virtual buttons defined in STATION: Left-Top, Left-Bottom, Right-Top, and Right-Bottom, which are located in four distinctive directions from the B-PTP gesture.

B. Signal Design

During both gesture enrollment and verification, STATION uses the embedded speakers of VIF devices to emit signals. Then, the devices capture and analyze the reflected signals for identifying hand gestures (i.e., active sonar). The signals emitted are essential to the above steps and require careful consideration. We need to ensure that the signals (and their reflections) do not interfere with users' normal use of VIF devices and can be produced and recorded by common speakers and microphones. Therefore, we select a frequency range that represents inaudible sound to humans, i.e., from 17200 to 20200 Hz, similar to prior studies [30], [31].

We also apply a gap of 500 Hz between the peaks of the signals. This is motivated by the maximum Doppler shift one can make using his hand: according to the Doppler effect equation (1), a 500 Hz gap is able to handle gestures with a

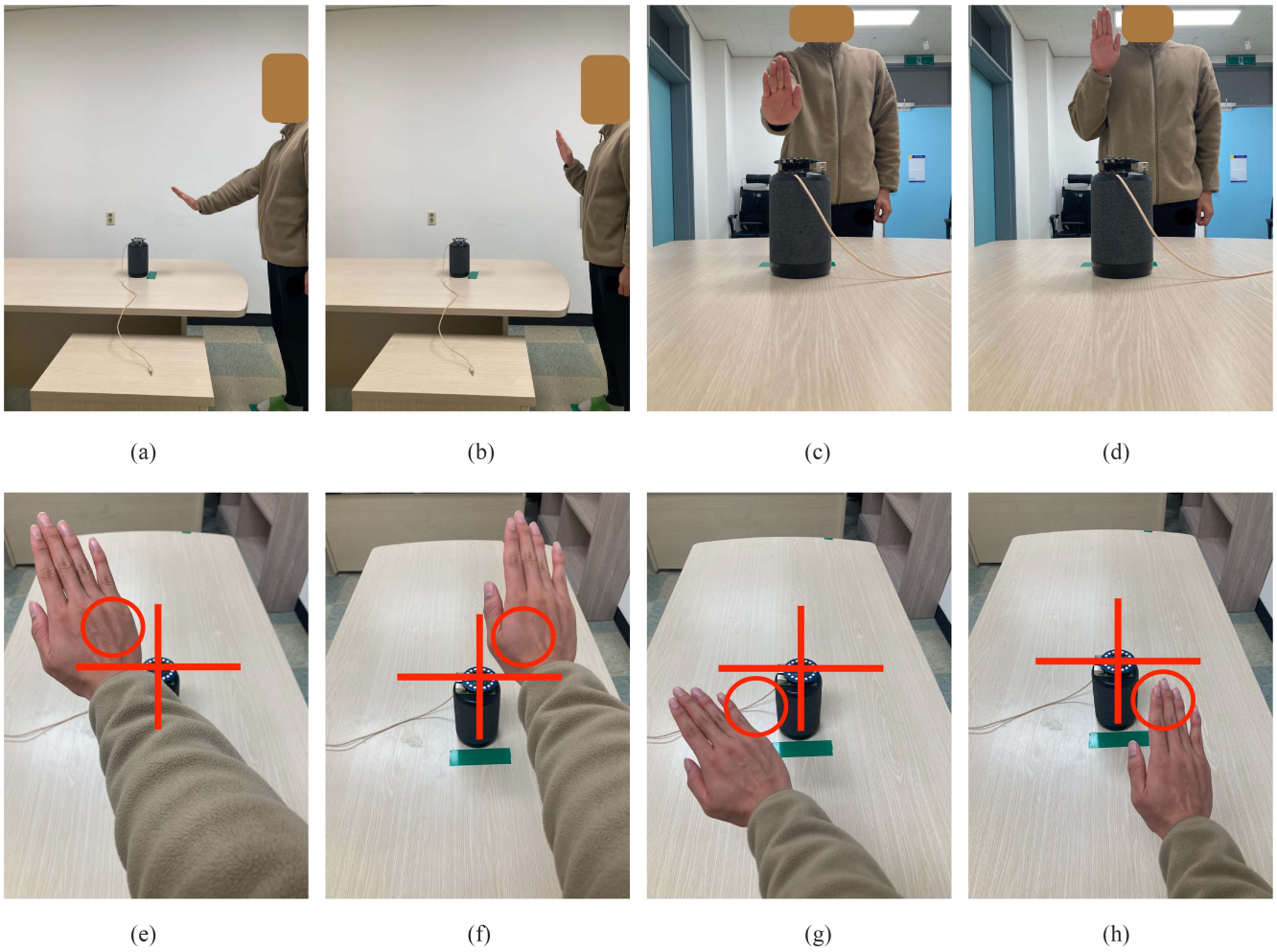


Fig. 2. PTP gestures. (a) Push movement (profile view). (b) Pull movement (profile view). (c) Push movement (front view). (d) Pull movement (front view). (e) Left-top. (f) Right-top. (g) Left-bottom. (h) Right-bottom.

maximum speed of over 4 m/s, which is the maximum speed of most users' gestures [32].

Further, we use wideband signals to handle the spatial ambiguity problem [33], and therefore improve the performance of gesture localization. Also, we design the signals to have low-crest factors. This is because a high-crest factor signal makes more distortion (e.g., phase distortion and harmonics distortion) than a low one, leading to a low SNR signal that is hard to distinguish the gesture with noise [34]

$$\Delta f = f_{\text{origin}} - f_{\text{Doppler}} = \frac{2v_{\text{object}}}{v_{\text{sound}} - v_{\text{object}}} f_{\text{origin}}. \quad (1)$$

C. Gesture Processing Pipeline

As described in Section III, the microphone array of the VIF devices will capture reflected signals from users' hand gestures (such as B-PTP and A-PTP gestures), and then the devices analyze the signals to determine which gestures they represent. These tasks require a gesture processing pipeline with multiple phases: *Preprocessing*, *Feature extraction*, and *Hand gesture recognition*. The *Preprocessing* phase transforms the signals into the frequency domain, sets the adaptive threshold value for sensing the Doppler signal with a B-PTP gesture, besides extracts the Doppler signal when the user pushes-then-pulls the virtual buttons. Then, the *Feature extraction* module extracts *DoA of Reflection* and *Hand Movement* information from the extracted Doppler signal. Finally, the *Hand gesture recognition* module removes the noise from *DoA of Reflection* information and identifies the gesture type while checking the validity with *Hand Movement* information, and authenticates the user by comparing the gesture sequence to the enrolled one.

Preprocessing: Detecting the hand gesture with the raw reflected signal is challenging because of the noise made by humans and other environmental factors, the signal emitted by speakers of the VIF devices, and the reflections from other stationary objects in the household, etc. Hence, we need to process the signal to minimize the noise and emphasize the Doppler signal reflected from the hand gestures. In this section, we describe the preprocessing phase with three steps: 1) transforming the raw signal into the frequency domain; 2) distinguishing the Doppler-existing frame; and 3) environment noise reduction and signal emphasis.

1) Transforming the Raw Signal Into Frequency Domain: In STATION, we leverage the frequency domain information to analyze the Doppler signals. Specifically, we transform the raw signals into the frequency domain information using the

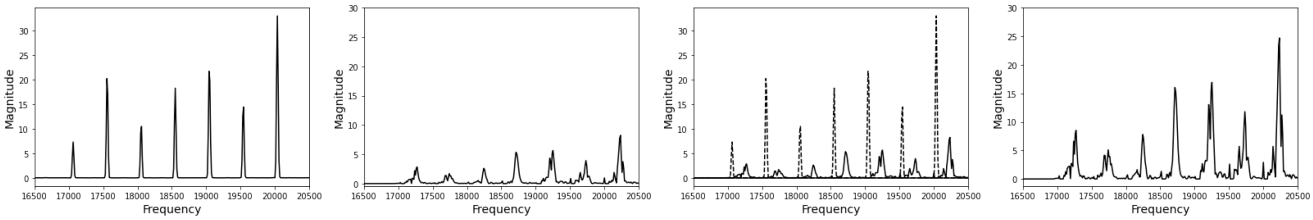


Fig. 3. Signal preprocessing. (a) Noise spectrum. (b) Doppler spectrum. (c) Doppler spectrum with noise. (d) Enhanced Doppler-bin.

STFT. Our sampling rate is set at 48 000 Hz, with a frame size of 4096 and a hop size of 1024 - parameters that are commonly used in other studies [30], [35] due to their suitability for analyzing Doppler shifts made by human gesture. To detect the frequency shifts more accurately, we utilize the Blackman window as our window function as it minimizes the side-lobe effect [36], thereby making the shifts more obvious. Note that the processing is done frame by frame and is generic to different types of VIF devices.

2) Distinguishing Doppler-Existing Frame: Sensing the Doppler effect precisely is challenging as the magnitude of the shifted frequency can vary with many factors, such as signal attenuation, signal fluctuation, and the difference in human body shape. Previous studies [18], [19] have addressed this challenge by using static thresholds. However, such an approach is less adaptive to the aforementioned factors and is less effective in dealing with variations in distance or body shape during gesture recognition.

To enhance the adaptiveness of STATION, we employ three techniques: normalizing the frequency magnitudes, incorporating the STD of frequency magnitudes around the pilot tone, and detecting the frames with Doppler shift with DBSCAN [37]. Specifically, for each frame, we gather the frequency magnitudes around the pilot tone and normalize them using the pilot tone. From the normalized values, we calculate the STD. We can then differentiate between frames with and without a Doppler shift based on the differences in their STD values since frames with a Doppler shift exhibit higher STD values compared to those without a Doppler shift. However, simply using the STD with a static threshold is insufficient for STATION to be adaptive. Thus, we use DBSCAN to cluster the frames based on the STD values. Since DBSCAN identifies low-density samples as noise, we can differentiate the frames with Doppler shift by identifying the frames with STD values labeled as noise. For optimal performance, we carefully select the minimum samples, taking into account the size of the data (i.e., number of frames), and set the epsilon value based on the k-distance average method [38].

3) Environment Noise Reduction and Signal Emphasis: For the identified D-frames, we need to remove noise further for the purpose of accurate gesture recognition. Specifically, we apply the spectral subtraction method to the signal–a common method for removing noise in the frequency domain. We take the non-Doppler spectrum as the environment noise spectrum [Fig. 3(a)], and then subtract the noise spectrum from the Doppler-existing spectrum [Fig. 3(c)]. If the result of the above subtraction is negative, we set it to be the value of the original spectrum multiplied by 0.005 (a treatment also used in [39]). The Doppler signal after subtracting noise is shown in Fig. 3(b). In the last step, we emphasize the Doppler spectrum throughout the frame to make the Doppler signal more visible to STATION [Fig. 3(d)].

Button Extraction: To determine the type of a gesture (i.e., which virtual button is pushed), we need to know the direction of the gesture toward the VIF device and whether the gesture is a PTP gesture or not. For this purpose, we need to extract two pieces of information from the emphasized Doppler spectrum: *DoA of Reflection* and *Hand Movement*. Below we elaborate on how STATION extracts the information.

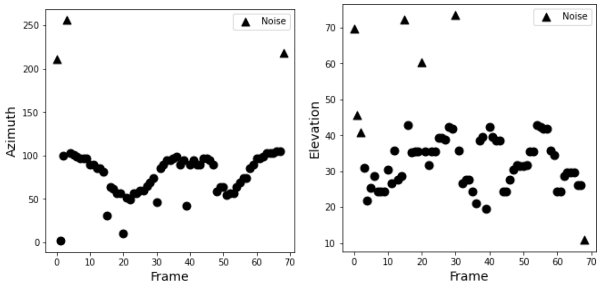
4) DoA of Reflection: In this study, we collect the direction of a user's hand toward his VIF device in the form of $\langle elevation, azimuth \rangle$ tuples, using sound localization algorithms.

We calculate the DoA from the sum of the cross-correlation score corresponding to fixed discrete points' TDoA for the microphone's position. Discrete points embedded in the virtual 3-D sphere around the VIF device represent the direction vector that has x, y, and z-axis components in radians from the sensing device. TDoA at the discrete point x to microphones $m1, m2$ is calculated with $\tau_{m1,m2}(x) = \lfloor f_s([\|x - x_{m1}\| - |x - x_{m2}\|]/c) \rceil$, where f_s is the sampling frequency of the system, c is the sound propagation speed, x_m is the position of mth microphone [9]

$$\hat{X}_s = \arg\max_{x \in g} P(x). \tag{2}$$

After calculating the TDoA of each discrete point, we calculate the cross-correlation score corresponding to TDoA from Doppler signals. A commonly adopted algorithm for calculating the score is the GCC-PHAT algorithm, which calculates the score from the spectral information between two microphones [40]. However, this algorithm is not applicable to our use case: it is designed to handle speech/sound in audible frequencies with high SNR, while we use inaudible sound and expect low SNR. To overcome the challenge, we adopt a new algorithm proposed in [14] – GCC-PHAT-β that introduces a weight factor in the PHAT filter to adjust the impact of spectral magnitude information. In the implementation of STATION, we set β to 0.5 since our experiments show that the β value achieves a more accurate cross-correlation score from the Doppler signals.

After getting the cross-correlation score, we run SRP-PHAT to calculate the DoA from the sum of the scores for each discrete point. As shown in (2), the algorithm calculates the



(a) (b)

Fig. 4. DoA of reflections. (a) Azimuth. (b) Elevation.

direction point corresponding to the max value of sums of the cross-correlation, where $P(x)$ is the result of the sum of the cross-correlation of each microphone pair for the discrete point X, g represents the set of all discrete points in the sphere, and $\hat{X}_s$ is the point of the sphere that corresponds to the maximum value.

The above results in the x, y, and z-axis represent the locations of the user's hand over the virtual 3-D space around the VIF device. It is relatively difficult to convert them to specific gesture types. Therefore, we further translate the above results into $\langle elevation, azimuth \rangle$ tuples. An issue for determining hand directions is that due to the noise and variance of human gestures, there are many outliers in the elevation of azimuth even for the same signal frame [as shown in Fig. 4]. Therefore, for each frame, we gather all the tuples of elevation and azimuth data and use the *median value* (which is more resistant to outliers than average) to represent the actual elevation and azimuth of the frame.

5) Hand Movement: Given our gesture design, the only valid gesture is the gesture with a PTP movement. Therefore, we need to identify the PTP movement from the Doppler signal. Specifically, the hand that approaches the VIF device causes the frequency to increase (i.e., red-shift), while moving away from the devices causes the frequency to decrease (i.e., blue-shift). Therefore, we can extract the PTP gestures by calculating the number of changes in the sign (i.e., increase, decrease) of the Doppler effect, and save the order of frames with push movement and frames with pull movement, respectively. For example, if the sign is changed from plus to minus, we record this as PTP pair.

6) Button Recognition: For each pair of push-and-pull frames, we extract the median data of *elevation* and *azimuth* points (with the same method as above). Afterward, the recognition of virtual buttons becomes trivial. In this study, we first identify the *elevation* and *azimuth* of the B-PTP gesture (the first PTP gesture in an authentication session), which is represented using $\langle E_b, A_b \rangle$. Then, for the following A-PTP, we extract the same direction information, noted as $\langle E_a, A_a \rangle$. We compare both tuples to decide which virtual button is pressed. For example, condition $\{E_a < E_b, A_a > A_b\}$ indicates the *Left-Top* button is pressed.

Fig. 5. Laboratory setup.

Fig. 6. Virtual buttons.

V. Evaluation

In this section, we evaluate the prototype implementation of STATION. Specifically, we aim to answer the following research questions.

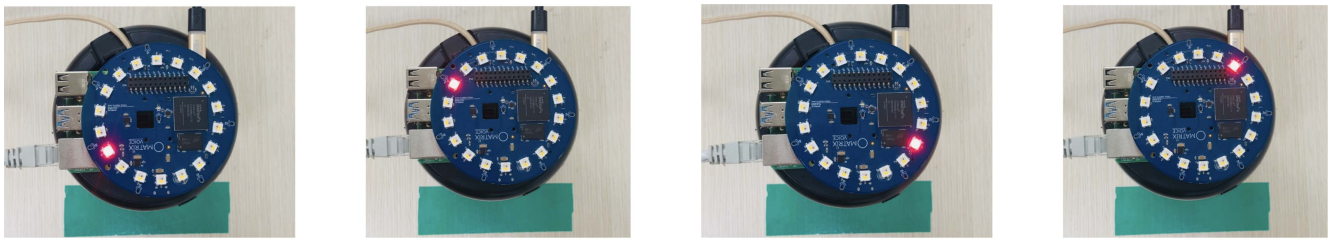
1) *RQ1:* What is the overall FAR and FRR of STATION?
2) *RQ2:* How robust is STATION to environmental noise and to different characteristics of end users?
3) *RQ3:* How does STATION compare to other authentication methods for smart home (or IoT) devices in terms of common criteria?

A. Evaluation Setup

Hardware: STATION requires two essential hardware modules to function, i.e., a speaker that emits inaudible sounds and a microphone array with at least four microphones to capture reflected signals for calculating DoA. To evaluate STATION, we implemented a prototype using an Omnidirectional speaker [41], a MATRIX Voice board (a radial array of seven microphones), and a Raspberry Pi 4 [42] for processing signals (see Fig. 6). In total, the modules cost us $185. It is worth noting that the cost is for setting up the evaluation environment and does not represent the additional cost that STATION introduces to VIF devices since STATION leverages the speaker and microphone array that are already present in the VIF devices, i.e., Amazon echo has a far-field 7-microphone array.

We made the signal file by adding sine waves whose frequencies correspond to the distance between peaks in our signal (Section IV, Fig. 3). In addition, we played the signal file using the Music application in the macOS Monterey after connecting the speaker to Macbook.

Room Setup: To mimic the room settings of regular users, we conduct all the experiments in a laboratory with a desk, a table, three bookshelves, and several chairs (see Fig. 5). We place the speaker on the table and ensure no objects are blocking the signals of the speaker within 1m.



(a) (b) (c) (d)

Fig. 7. LED guidance. (a) *Left-bottom* guidance. (b) *Left-top* guidance. (c) *Right-bottom* guidance. (d) *Right-top* guidance.

TABLE I
INFORMATION OF THE PARTICIPANTS

Participant	Sex	Height	handedness	Age
U1	Male	175cm	Right	27
U2	Male	176cm	Right	24
U3	Male	170cm	Right	24
U4	Male	175cm	Right	23
U5	Male	177cm	Right	21
U6	Male	175cm	Right	23
U7	Male	177cm	Right	24
U8	Male	168cm	Left	24
U9	Female	165cm	Right	16
U10	Female	165cm	Right	23
U11	Female	160cm	Right	22

Participants: To demonstrate the effectiveness of our approach, we recruited 11 participants (labeled *U1-U11*) to test STATION. We recorded the demographic information of the participants, e.g., height, handedness, and gender, for the purpose of evaluating the impact of such information on STATION. Overall, the participants are in a height range from 1.60 to 1.77 m. 10 participants (*U1-U7*, *U9-U11*) are right-handed, and one participant (*U8*) is left-handed. Also, 3 participants (*U9-U11*) are female, and the other participants are male. Lastly, the age of participants is from 16 to 27 (Table I).

To reduce unattended noises from each participant, we provided detailed guidance about how the participants can make hand gestures. Specifically, before conducting the experiment, we asked all the participants to view a tutorial video (similar to the training videos that are commonly seen on new devices and software) that shows how one may make gestures for authentication, e.g., how to PTP the virtual buttons. Then, we showed four sequences of gestures (labeled *GS1-GS4*) and displayed them to the participants via the LED on the MATRIX Voice board (see Fig. 7). The first sequence (*GS1*) consists of three gestures (i.e., Left-Bottom, Left-Bottom, Right-Bottom). The second sequence (*GS2*) consists of three gestures (i.e., Right-Top, Right-Top, Right-Bottom). The third sequence (*GS3*) consists of three gestures (i.e., Left-Bottom, Left-Bottom, Right-Top). The last sequence (*GS4*) consists of three gestures (i.e., Left-Top, Right-Top, Right-Top). A B-PTP gesture should be made first when the participant is asked to make a gesture sequence.

Our institute granted an IRB exemption for this evaluation since we did not collect or record any sensitive or personally identifiable information from the participants.

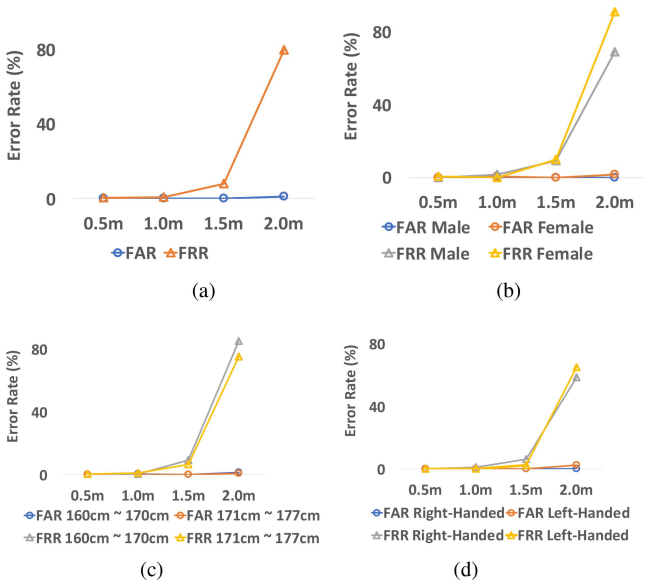


Fig. 8. FAR and FRR over various conditions. (a) Overall. (b) Over genders. (c) Over height ranges. (d) Over handedness groups.

B. RQ1: Overall FAR and FRR

We used two metrics to evaluate the accuracy of STATION: 1) FAR and 2) FRR. Specifically, FAR measures how likely can STATION accidentally take an incorrect gesture sequence as the enrolled sequence, and FRR measures how often a correct gesture sequence is classified as different to the enrolled sequence.

Specifically, we asked the participants to make hand gestures at different distances to the VIF device, i.e., 0.5, 1.0, 1.5, and 2.0 m. The participants make each enrolled gesture (*GS1-GS4*) 10 times, which leads to 440 authentication attempts for each distance (i.e., 4 gestures * 10 times * 11 participants). In selecting the number of authentication attempts and participants, we refer to other gesture-based authentication research [43] and gesture recognition research [44]. We count the number of successful authentication attempts and report the FRR of each distance as *1 - # success attempts / 440*. As shown in Fig. 8(a), STATION can process correct gesture sequences at pretty low rejection rates, with a 0.45, 0.91, and 7.93% FRR at the distance of 0.5, 1.0, and 1.5 m, respectively.

To measure FAR, we randomly sampled the other gesture sequences with three gestures other than the enrolled gesture. We asked the participants to make 10 incorrect gestures for

(a) (b) (c) (d)

Fig. 9. Setting to test impact of reverberation. (a) 10 cm apart from the wall. (b) Overall setting. (c) STATION with small objects. (d) Different furniture density room.

each enrolled gesture, which also leads to 440 authentication attempts for each distance. We count the number of successful authentication attempts and report the FAR as *# success attempts / 440* for each distance. The evaluation results indicate that STATION achieves a rather low FAR when provided with random and incorrect gesture sequences. For example, the FAR at a distance of 0.5 m is 0.00%, the FAR at a distance of 1.0 m is 0.23%, and the FAR at a distance of 1.5 m is 0.00%.

In the evaluation, we did not notice significant differences of FAR or FRR among different enrolled gestures and participants. Also, the FRR of STATION increases significantly to as high as 80.00%. This is due to the incapability to capture far-field signals using the microphone array available on the MATRIX Voice board. Interestingly, the FRR does not increase significantly at 1.5, which is the distance the users feel comfortable interacting with smart speakers (see our survey in Section A).

C. RQ2: Robustness of STATION

Robustness to Environmental Noises: To evaluate whether our approach is robust to environmental noises, we measured the FAR and FRR of STATION under several noisy settings, in the presence of reverberation and small objects, with different furniture densities and external noise levels. Specifically, we measured 80 authentication trials (i.e., 20 times for each gesture sequence with 10 correct sequences and 10 incorrect sequences) at 1∼1.5 m, the comfortable distance, which users can take tradeoff for authentication, to users in our survey (Section A).

1) *Impact of Reverberation From Walls:* To test the impact of reverberation, we placed the speaker 20, 10 cm from the wall [Fig. 9(a)] and evaluated the FRR of STATION by making gestures at 1 m from the speaker [Fig. 9(b)]. We noticed that the FRR increases significantly from 1.25% for the normal use case to 72.50% for the 10 cm reverberation setting. However, at the manufacturer's minimal recommendation distance [29], 20 cm reverberation setting, the FRR did not increase much, i.e., from 1.25% to 2.5%.
2) *Impact of Reverberation From Small Objects:* In addition to walls, reverberation can also be caused by small objects. Therefore, we placed small objects of different sizes and textures at 0.3 m from the speaker [Fig. 9(c)], and evaluated the FAR and FRR of STATION by making gestures at 1m from the speaker. The result suggests that the impact of small objects is minimal, with only a 3.75% of increase FRR and 0.31% of decrease FAR.
3) *Impact of Different Furniture Density:* To measure how different furniture density impacts STATION, we repeat the evaluation with a 1m distance with objects such as a sofa, TV, recliner, and air purifier in a room [Fig. 9(d)]. It's worth noting that we did not place large objects that are larger than the smart speaker in between the smart speaker and the user, as it is obvious that the system would not work without line-of-sight. Then, we conducted the experiment sitting on the sofa at 1m from the speaker. The FRR does not increase even if there were many furniture and appliances. The result suggests that there is no impact on furniture density.
4) *Impact of Noise:* To evaluate how noise impacts the system, we design the experiment referring to comparative noise level [45]. we repeat the evaluation of correct and incorrect gestures with a 1m distance with three noise levels; 50 dB without music (only environmental noise), and 70 dB with music (environmental noise+ music) within 0.3 m. The result indicates that STATION achieves satisfying results on different noise levels, with a 1.25 and 5.00% FRR for noise levels 50 and 70 dB.

Robustness to User Characteristics: We grouped participants by sex, height, and handedness. Then, we analyze the result by comparing the FAR and FRR between groups. We used the same data collected for evaluating robustness to noises here.

1) *Impact of Height:* To check the impact of height, we divide the users into two groups: a) from 160 to 170 cm and b) from 170 to 177 cm. There were five participants in the first group and six in the second group. The result is shown in Fig. 8(c). The differences in the average FRR of all gesture sequences are 0.08, 0.75, 2.89, and 10.08% at 0.5, 1.0, 1.5, and 2.0 m. The result shows that height does not have a significant impact on the FAR and FRR of gestures ∼1.5 m, at least in the range of 160 to 177 cm.
2) *Impact of Gender:* To check the impact of gender, we divide the users into the male group and the female group. Given that there are three participants in the female group, we chose three male participants randomly to make the number of participants in each group

similar. The differences in the average FRR of all gesture sequences are 0.83, 1.67, 0.83, and 21.66% at 0.5, 1.0, 1.5, and 2.0 m. So, there is no significant impact on the average FRR of all gesture sequences ∼1.5 m.

3) *Impact of Handedness:* To check the impact of handedness, we divide the users into the right-handed group and the left-handed group. Given that there is only one participant in the left-handed group, we choose two right-handed participants randomly to make the number of participants in each group similar. The differences in the average FRR of all gesture sequences are 0.00, 1.25, 3.75, and 6.25% at 0.5, 1.0, 1.5, and 2.0 m. So, there is no significant impact on average FRR of all gesture sequences ∼1.5 m. The result shows that handedness does not have a significant impact on the FRR of gestures.

D. RQ3: Comparing STATION With Other Methods

In this section, we first select several recently proposed authentication methods designed for VIF devices and compare STATION to them based on a series of criteria proposed by prior research studies. Then, we compare the entropy of STATION with a set of general-purpose authentication factors to establish the strength of STATION relative to these authentication factors.

Comparing STATION to Other Authentication Methods for VIF Devices: We conducted a comparison of STATION with other authentication methods, using a set of criteria ($C1$–$C20$) inspired by Stephenson et al. [46]. These criteria cover various dimensions, including usability, deployability, security & privacy, and accessibility. In this comparison, we also followed the criteria of satisfaction (e.g., *Full*, *Quasi*, and *No*) in Stephenson et al. [46], by empirically evaluating these authentication methods.

Specifically, we compared STATION to voice-based methods (e.g., VocalPrint [47] and 2MA [48]), gesture-based methods (e.g., Shen et al. [43], TwistIn [49], and HandLock [35]). We also compared STATION to PIN since STATION was designed to support virtual buttons similar to PIN. Our comparative study, as shown in Table III, highlights the main strengths of STATION stemming from the fact that it does not necessitate additional hardware and its flexibility to support re-enrollment in new hand gesture sequences. For instance, STATION can be readily implemented on VIF devices ($C2$) without the need for any additional hardware, and users do not have to carry authentication hardware around ($C9$). Moreover, STATION eliminates the necessity of storing users' biometric data and offers the flexibility to easily change authentication credentials ($C19$). Additionally, STATION does not frequently cause errors as demonstrated by its low FAR and FRR ($C10$, as evaluated in Section V-B).

Comparison of Entropy to General-Purpose Authentication Methods: Entropy analysis is a common method used to estimate the strength of an authentication method. In this study, we compared the entropy of STATION to that of both knowledge-based authentication methods, including password [50] and PIN [51], as well as biometric-based authentication methods, including methods using keystroke [52], iris [53], [54], face [55], finger vein [56], fingerprint [57], and pupillometry [58].[1] The entropy of knowledge-based authentication methods is determined using Shannon entropy, expressed as $H(x) = -\Sigma_i P(x_i) \log_2 P(x_i)$, where $P(x)$ represents the distribution of a variable x. We determine the entropy of STATION using this entropy as well, since STATION is essentially a variant of password-based authentication. Since users of STATION can choose any sequence of the four virtual buttons as a password, STATION provides an entropy of $2 * n$ bits, given that a sequence of length n is used. The Shannon entropy cannot be applied to biometric-based authentication methods, as it disregards intrauser variability and tends to overestimate biometric information [58]. Therefore, the entropy of biometric-based authentication methods is based on relative entropy, which quantifies the extent to which the distribution of a single user's biometric features differs from that of the population. This relative entropy is measured using Kullback-Leibler Divergence (KLD), where $K(x) = \Sigma_i P(x_i) \log_2 (P(x_i)/Q(x_i))$, with $P(x)$, $Q(x)$, and x representing the feature distribution of the target user, that of the reference set, and the feature space, respectively.

TABLE II
ENTROPY OF VARIOUS AUTHENTICATION METHODS

Category	Work	Authentication Method	Entropy (bits)
Biometric-based Authentication	Sae-Bae et al. [52]	Keystroke	3.48-4.62
	Krivokuca et al. [56]	Finger Vein	4.20-19.50
	Sutcu et al. [53]	Iris	8.90-10.00
	Takahashi et al. [57]	Fingerprint	18.60
	Adler et al. [55]	Face	37.00-55.60
	SoundLock [58]	Pupillometry	81.00
	Youmaran et al. [54]	Iris	278.00-288.00
Knowledge-based Authentication	Wang et al. [50]	Password	20.00-23.00
	Wang et al. [51]	PIN (4-digit, 6-digit)	8.41 (4-digit), 13.21 (6-digit)
	STATION	Button Sequence (entered by hand gestures)	12.00 (6 length sequence), 14.00 (7 length sequence)

Table II shows the entropy comparison of STATION with other authentication methods. Specifically, when $n = 6$, the entropy of STATION is lower than that of some biometric-based authentication methods, as demonstrated in [54], [55], [57], and [58]. However, the advantage of STATION is that it reuses the common voice interfaces within VIF devices, rather than relying on extra hardware such as iris or fingerprint sensors. Additionally, when $n = 6$, the entropy of STATION authentication falls between that of a 4-digit and a 6-digit PIN, which are commonly supported authentication methods in devices with a display [59], [60], [61], [62], [63]. Therefore, STATION has the potential to attain comparable security on VIF devices to these real-world authentication methods used by commercial headful devices.

E. System Performance

To assess the performance of STATION, we conducted two measurements. First, we measured the time taken for *preprocessing*, and second, we analyzed the time taken for *button*

[1] We could not compare STATION's entropy with the authentication methods in Table III because calculating entropy for those methods requires access to their code and models, which are not publicly available.

TABLE III
COMPARISON BETWEEN STATION AND OTHER METHODS

		Deployability				**Usability**							**Accessibility**					**Security & Privacy**			
		OS-Supported (C1)	*Platform-Agnostic* (C2)	*Mature* (C3)	*Low-Power-Consumption* (C4)	*Efficient-to-Use* (C5)	*Physically-Effortless* (C6)	*Memorywise-Effortless* (C7)	*Easy-to-Learn* (C8)	*Nothing-to-Carry* (C9)	*Infrequent-Errors* (C10)	*Acceptable-in-Public* (C11)	*Accessible-Visual* (C12)	*Accessible-Hearing* (C13)	*Accessible-Speech* (C14)	*Accessible-Mobility* (C15)	*Accessible-Cognitive* (C16)	*Resilient-to-Guessing* (C17)	*Resilient-to-Physical-Observation* (C18)	*Protects-User-Privacy* (C19)	*Multi-Factor* (C20)
Voice-based Authentication	VocalPrint [47]	✗	✗	✗	✗	✓	△	✓	✓	✗	△	✗	✓	✓	✗	✓	✓	△	✓	✗	✗
	2MA [48]	✗	✓	✗	✗	✓	△	✓	✓	✓	△	✗	✓	✓	✗	✓	✓	△	✓	✓	✗
Gesture-based Authentication	Shen et al. [43]	✗	✗	✗	✗	✗	✗	✓	✓	✗	△	✗	✓	✓	✓	✗	✓	△	✗	✗	✗
	TwistIn [49]	✗	✗	✗	✗	△	✗	✓	✓	✗	✓	△	✓	✓	✓	✗	✓	✓	✓	✓	✗
	HandLock [35]	✗	✓	✗	✗	△	✗	△	✓	✓	△	✗	✓	✓	✓	✗	✗	△	✗	✗	✗
	4-digit PIN	✓	✓	✓	✗	△	✗	✗	✓	✓	✓	✓	✗	✓	✓	✗	✗	✗	✗	✓	✗
	STATION	✗	✓	△	✗	△	✗	✗	✓	✓	✓	✗	✓	✓	✓	✗	✗	✗	✗	✓	✗

✓: *Full*-satisfaction, △: *Quasi*-satisfaction, ✗: *No*-satisfaction.

extraction. To evaluate the system, we gathered 20 samples and measured the time taken in each phase. The average time for *preprocessing* was 1.06 ms, while the average time for *button extraction* was 43.54 ms. Therefore, the average total latency was 44.60 ms. These results were calculated using a Raspberry Pi 4, which has specifications similar to those of common VIF devices.

VI. DISCUSSION

There are two typical types of security threats targeting VIF devices: 1) adversaries performing unauthorized physical accesses to VIF devices or 2) remote attacks conducted by adversaries via other compromised devices using techniques such as signal replay and synthesis. Previous studies [30], [64] have proposed methods for checking users' physical presence at VIF devices using liveness detection. While these methods help mitigate the second threat, they are unable to address the first. In this study, we utilize sequences of hand gestures (or virtual button clicks) as authentication passwords. As long as these passwords remain confidential and unknown to attackers, STATION helps mitigate both security threats.

STATION recognizes hand gestures by measuring the hand direction and movement features from the *Doppler* signal. This approach is limited in cases where the gestures are complicated, and the direction features are subtle. To complement this, we only allow hand gestures that are composed of a predefined set of pushing-then-pulling virtual buttons (Fig. 2). Additionally, while we have demonstrated the effectiveness of STATION across various environmental noises and user characteristics, we acknowledge that the evaluation may not cover all potential noises and may not be entirely representative of all user groups. To evaluate the capability of STATION in real-world and large-scale applications, we may need to perform an evaluation with a more diverse and larger set of participants.

VII. RELATED WORK

In this section, we discuss the authentication methods applicable to voice interfaces, i.e., voice- and gesture-based authentication. We further introduce related acoustic-based hand gesture recognition methods.

Voice-Based Authentication: Voice-based authentication has drawn attention from both academia [2], [3], [47], [48] and industry [65], [66]. For example, Carlini et al. [2] described the use of audio CAPTCHA system as a challenge-response authentication method. Diao et al. [3] proposed to authenticate users of Google Search app with voice fingerprinting. VocalPrint [47] supports user authentication by analyzing the vocal vibrations of voice commands. 2MA [48] develops a more secure authentication method with two microphones using DoA techniques. Compared to these studies that process voice data, STATION authenticates users by analyzing the *Doppler* signal reflected from the users' live hand gestures, which is more resistant to replay attacks, and hidden/inaudible voice attacks, etc.

Gesture-Based Authentication: Prior studies discussed gesture-based authentication systems for different devices, e.g., smartphones [67], [68], smart speakers [35], and smartwatches [43], [49], [69]. For example, Hong et al. [67] proposed to implement a motion gesture authentication system using the accelerometer of smartphones. TwistIn [49] presents to authenticate users by analyzing the motion data captured by the device (which the users authenticate to) and a smartwatch. Lewis et al. [69] developed a real-time authentication system with the accelerometer and gyroscope of smartwatches. Similarly, Shen et al. [43] implemented a hand-waving-based unlocking system using accelerometer data.

Most closely related to our research is HandLock [35], which also leverages hand gestures to authenticate device users with built-in microphones. However, HandLock applies its analysis to acoustic signals to extract gesture fingerprints. As a result, it is highly sensitive to SNR and only functions

effectively within a very short range (approximately 30 cm). In this study, we do not extract gesture fingerprints, but rather identify hand gestures representing a predefined set of virtual buttons (or password) using algorithms that are more robust to low SNR settings, such as GCC-PHAT-β and SRP-PHAT. This enables STATION to capture hand gestures over a greater distance (up to 1.5 m, as detailed in Section V). Additionally, HandLock utilizes arm acceleration from reflected signals but does not determine the signal's origin (i.e., hand direction). Therefore, compared to STATION, HandLock is unable to detect remote attacks that exploit signal replay and synthesis.

Acoustic-Based Gesture Recognition: The problem of recognizing hand gestures from acoustic signals has been discussed on different fronts. First, prior studies [44], [70], [71] explored methods to track fingers or gestures using the orthogonal frequency divison multiplexing (OFDM) signal. Second, some studies [72], [73], [74], [75], [76] proposed to use the frequency modulated continuous wave (FMCW)-based methods to recognize hand gestures. Additionally, other studies [18], [19], [32], [77] collected a variety of features corresponding to the *Doppler* effect for the purpose of recognizing user gestures. Especially, there were studies that achieved room-scale sensing [70], [76], they are 2-D-based approaches. Given that STATION utilizes the 3-D information to determine the pressed virtual button, these works are not fit our interest. Furthermore, while our study is also built upon the *Doppler* effect, we utilize several algorithms (e.g., sound localization, and density-based clustering on Doppler-existing frames) to achieve robust hand gesture detection even in the low SNR settings and at a further distance to the devices.

VIII. CONCLUSION

In this article, we present STATION, a gesture-based user authentication system, which recognizes the gesture by analyzing the Doppler signal reflected from the user's hand. We design and apply new techniques, such as density-based clustering and *DoA of Reflection*, to analyze the reflected signals and extract hand gestures from the signals. Using STATION, device users can prevent unauthorized accesses to their devices from both local and remote adversaries (e.g., signal replay and synthesis attacks). We implemented a prototype of STATION using low-cost voice interfaces and evaluated its effectiveness through onsite participants. The evaluation results show that STATION is resistant to changes of human factors, works well in the low SNR environment, and is highly accurate (i.e., FAR of 0.08% and FRR of 3.10% for ∼1.5-m distance between the users and devices).

APPENDIX

USER PERCEPTIONS TO STATION

We also conducted the surveys to understand VIF device users' concern about the security in using it and design STATION closer to real-world requirements (e.g., using distance). So, we asked the participants about the user habits, recognition of security, enhancement of security of the smart home devices, and convenience of STATION. We gathered 229 participants' responses from SurveyMonkey [78] with carefully designed questions (in the Appendix). The survey result provided us that 72.0% of online survey participants feel STATION is convenient and willing to take a tradeoff coming within 1.0∼1.5 m for authentication.

We conducted a user study with a group of online participants to understand VIF device users' security concerns and design STATION that meets real-world requirements. Specifically, the user study mainly focuses on three aspects: 1) The types of smart home devices used by users and the number of devices with voice interfaces; 2) How users interact with voice interfaces; 3) Whether users feel comfortable using gesture-based authentication solutions, and their perceptions of STATION. Results collected from the user study confirmed that users have smart devices with voice interfaces, feel it necessary to authenticate such devices for security, and STATION meets their expectations for usability and security.

Design of User Study: The study contains 13 questions grouped into four categories. In the first category, we asked users what types of smart devices they have (Q1), e.g., smart speakers and cameras, how many devices they have in total (Q2), and how many devices have voice interfaces (Q3), i.e., speaker and microphone.

The second category of questions focuses on collecting users' habits of interacting with the voice interfaces, such as how long they have owned and used the devices with voice interfaces (Q4), where they usually use the devices (Q5), and how far away (e.g., <0.5, 0.5–1.0 m, 1.0–1.5 m, 1.5–2.0 m, >2.0 m) they interact with the voice interfaces (Q6), etc.

We design the third category of questions to understand users' security expectations of smart devices and their perceptions of gesture-based authentication like STATION. Specifically, we first ask users that own smart devices whether they are concerned about device security, using a 0-10 score where ten means *strongly concerned* and 0 means *not concerned at all* (Q7). We then present STATION to users via text descriptions and screenshots. We ask whether the *virtual buttons* design of STATION is difficult to learn, using a score between 0-10 where 10 means *very difficult* and 0 means *not difficult at all* (Q8). We also ask users whether STATION provides a convenient method for device authentication (a 10 score means *extremely convenient*, and 0 means *very difficult to use*) (Q9). Afterward, we ask users whether they would like to come within 1.0 to 1.5 m of the smart devices for device authentication (Q10). In the end, we ask users, if STATION is deployed to their devices, how often they would like to use it.

Finally, we collect the demographic information of the participants, such as their age, gender, and highest education (Q11-Q13). Such information allows us to evaluate the overall perceptions of the general public. Note that we do not collect any personally identifiable information from the questions. A detailed list of questions is shown in Table IV.

To ensure the reliability of the survey, we carefully designed the questions and conducted a pilot study to enhance the clarity of the questions and assess the consistency of participants'

TABLE IV
ONLINE SURVEY QUESTIONS

Number	Question
Q1	Do you use any smart home device with a speaker and microphone (e.g., smart speaker, smart TV, smart refrigerator, smart security camera, smart baby monitor, etc.)?
Q2	How many smart home devices do you use?
Q3	How many smart speakers do you use?
Q4	How long have you used a smart speaker?
Q5	Where do you use a smart speaker usually?
Q6	How far do you use a smart speaker usually?
Q7	In your opinion, how secure do you perceive the smart home device to be against cyber attacks?
Q8†	Could you share your thoughts on the ease of learning STATION?
Q9†	What are your perceptions regarding the convenience of using STATION?
Q10†	How do you feel about the idea of being within 1.0-1.5 meters of the smart home device for the authentication process?
Q11	What is your age?
Q12	What is your gender?
Q13	What is the highest education level that you have completed?

† The questions are preceded by a textual description of STATION: "Please refer to the picture above to answer the question. We designed the user authentication method using a speaker and microphone that are built-in smart home devices. The key idea is pushing-pulling the virtual buttons (red circles in the picture) following their unique sequence (i.e., Left-Bottom, Left-Top, Right-Bottom); it is similar to inputting the passcode with a keypad. The system authenticates the user successfully if the user pushes-pulls the buttons in their unique sequence. In technical, when the speaker emits the inaudible ultrasound, the microphone records the signal reflected from the hand. Then, the system determines which virtual button was pushed-pulled by analyzing the recorded signal. Finally, if the inputted button sequence is the same as the user's unique sequence, the authentication is done successfully. Given the system design, our system does not gather any biometric information (e.g., fingerprint, iris) or privacy-sensitive information (e.g., face, voice) compared to other authentication methods."

responses. One of the major challenges we encountered in designing the questions was social desirability bias. To mitigate this bias, we employed indirect questioning [79]. Specifically, we used indirect questions when inquiring about preferences or feelings (e.g., Q7, Q8, Q9, Q10).

In the pilot study, we conducted a survey with a group of 10 participants recruited through social media postings. All participants were graduate students and faculty members aged between 22 and 37. Our objective was to assess the clarity of each question, so we asked participants to provide feedback on the clarity of each question through text responses. Additionally, we provided two survey versions, with the order of the questions shuffled, to evaluate response consistency. Based on our findings, we reflected on areas for improvement in the questions, such as the use of ambiguous pronouns and excessive reliance on professional terms. We did not observe any notable differences in responses between the two shuffled versions.

Then, we anonymously conducted the massive user study on Survey Monkey [78] using the finalized questions from December 1, 2022, to December 8, 2022. We target participants that are of age 18 or above, in the United States, and have at least one smart home device. We awarded each participant 4.87 USD for completing the study.

Results and Findings: We received a total of 229 valid responses, with 118 from female participants and the rest from male participants. Most participants are between 21 to 59 in age (74.7%) and have a high school or higher degree (96.1%). From the valid responses, we summarize the following findings.

1) Almost all participants have at least one smart home device with voice interfaces. According to Q1, smart speakers are among the most popular smart home devices that support voice interfaces (i.e., speakers and microphones), with 88.2% (202 out of 228) of participants owning at least one smart speaker. In addition, 39.9% of the participants own other smart home devices, such as smart cameras, smart locks, etc. On average, each participant has 1.62 smart speakers installed in his/her home.
2) Users typically interact with voice interfaces at close range (<2 m) in their private space. In practice, most participants issue commands to voice interfaces at close range, likely due to the short-range supported by the hardware: 71.8% of smart speaker users (0–50 cm, 5.45%, 50 cm–1 m, 21.29%, 1–1.5 m, 25.25%, 1.5–2 m, 19.80%) interact with their devices within 2 m. Additionally, almost all (88.6% in smart speakers, 83.7% other devices) participants use their devices in private spaces (such as bedrooms) rather than public spaces. This aligns with our assumption that STATION can be applied to most devices without being physically compromised, assuming that other members of the family are benign.
3) A significant portion of participants are concerned about the security of their smart home devices, and most participants feel that STATION (i.e., gesture-based authentication) is a convenient method to secure their devices. The answers to Q7 indicate that many participants are concerned about the security of their smart home devices, with an average score of 5.5. After showing them STATION, most of the participants feel that the gestures designed in STATION are convenient to users (with an average score of 6.8, Q9). Additionally, 76.0% of participants are willing to come within 1.0 to 1.5m of their devices in order to get the security feature – device authentication using STATION.

REFERENCES

[1] M. May, "Inaccessibility of CAPTCHA: Alternatives to visual Turing tests on the Web," W3C Working Group Note, W3C, Cambridge, MA, USA, 2005.

[2] N. Carlini et al., "Hidden voice commands," in *Proc. 25th USENIX Security Symp. (USENIX Security)*, 2016, pp. 513–530.

[3] W. Diao, X. Liu, Z. Zhou, and K. Zhang, "Your voice assistant is mine: How to abuse speakers to steal information and control your phone," in *Proc. 4th ACM Workshop Security Privacy Smartphones Mobile Devices*, 2014, pp. 63–74.

[4] S. Chen et al., "You can hear but you cannot steal: Defending against voice impersonation attacks on smartphones," in *Proc. IEEE 37th Int. Conf. Distrib. Comput. Syst. (ICDCS)*, 2017, pp. 183–195.

[5] M. E. Ahmed, I.-Y. Kwak, J. H. Huh, I. Kim, T. Oh, and H. Kim, "Void: A fast and light voice liveness detection system," in *Proc. 29th USENIX Security Symp. (USENIX Security)*, 2020, pp. 2685–2702.

[6] H. Abdullah, W. Garcia, C. Peeters, P. Traynor, K. R. Butler, and J. Wilson, "Practical hidden voice attacks against speech and speaker recognition systems," 2019, *arXiv:1904.05734*.

[7] M. Chen, L. Lu, Z. Ba, and K. Ren, "PhoneyTalker: An out-of-the-box toolkit for adversarial example attack on speaker recognition," in *Proc. IEEE Conf. Comput. Commun.*, 2022, pp. 1419–1428.

[8] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[9] J. H. DiBiase, *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays*. Providence, RI, USA: Brown Univ., 2000.

[10] F. Grondin and F. Michaud, "Lightweight and optimized sound source localization and tracking methods for open and closed microphone array configurations," *Robot. Auton. Syst.*, vol. 113, pp. 63–80, Mar. 2019.

[11] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques* (Transferred to Digital Print on Demand). Upper Saddle River, NJ, USA: PTR Prentice Hall, 2002.

[12] B. Kwon, Y. Park, and Y.-S. Park, "Analysis of the GCC-PHAT technique for multiple sources," in *Proc. ICCAS*, 2010, pp. 2070–2073.

[13] C. Zhang, D. Florêncio, and Z. Zhang, "Why does PHAT work well in lownoise, reverberative environments?" in *Proc. IEEE Int. Conf. Acoust., Speech Signal Proces.*, 2008, pp. 2565–2568.

[14] J. M. Villadangos, J. Ureña, J. J. García-Domínguez, A. Jiménez-Martín, Á. Hernández, and M. C. Pérez-Rubio, "Dynamic adjustment of weighted GCC-PHAT for position estimation in an ultrasonic local positioning system," *Sensors*, vol. 21, no. 21, p. 7051, 2021.

[15] "Android compatibility definition document." 2022. [Online]. Available: https://source.android.com/docs/compatibility/cdd

[16] "Apple platform security: Face ID, touch ID, passcodes, and passwords." 2022. [Online]. Available: https://support.apple.com/guide/security/face-id-touch-id-passcodes-and-passwords-sec9479035f1/web

[17] "Device compliance settings for android enterprise in Intune." 2024. [Online]. Available: https://learn.microsoft.com/en-us/mem/intune/protect/compliance-policy-create-android-for-work

[18] S. Gupta, D. Morris, S. Patel, and D. Tan, "SoundWave: Using the doppler effect to sense gestures," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2012, pp. 1911–1914.

[19] K.-Y. Chen, D. Ashbrook, M. Goel, S.-H. Lee, and S. Patel, "Airlink: Sharing files between multiple devices using in-air gestures," in *Proc. ACM Int.Joint Conf. Pervasive Ubiquitous Comput.*, 2014, pp. 565–569.

[20] D. A. Winter, *Biomechanics and Motor Control of Human Movement*. Hoboken, NJ, USA: Wiley, 2009.

[21] "Fido authenticator security requirements." 2021. [Online]. Available: https://fidoalliance.org/specs/fido-security-requirements/fido-authenticator-security-requirements-v1.5-fd-20211102.html

[22] "IoT security: Intel EPID simplifies authentication of IoT devices." 2016. [Online]. Available: https://www.networkworld.com/article/3121981/iot-security-intel-epid-simplifies-authentication-of-iot-devices.html

[23] O. Alrawi, C. Lever, M. Antonakakis, and F. Monrose, "SOK: Security evaluation of home-based IoT deployments," in *Proc. IEEE Symp. Security Privacy (SP)*, 2019, pp. 1362–1380.

[24] S. Demetriou et al., "HanGuard: SDN-driven protection of smart home WiFi devices from malicious mobile apps," in *Proc. 10th ACM Conf. Security Privacy Wireless Mobile Netw.*, 2017, pp. 122–133.

[25] T. Zebra. "Burglary statistics." 2023. [Online]. Available: https://www.thezebra.com/resources/research/burglary-statistics/

[26] "What is Alexa voice ID?" Amazon. 2023. [Online]. Available: https://www.amazon.com/gp/help/customer/display.html?nodeId=GYCXKY2AB2QWZT2X

[27] S. Esposito, D. Sgandurra, and G. Bella, "Alexa versus Alexa: Controlling smart speakers by self-issuing voice commands," in *Proc. ACM Asia Conf. Comput. Commun. Security*, 2022, pp. 1064–1078.

[28] X. Yuan et al., "All your Alexa are belong to us: A remote voice control attack against echo," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2018, pp. 1–6.

[29] "Where is the best place for a smart speaker in a room?" 2024. [Online]. Available: https://smaart.house/where-is-the-best-place-for-a-smart-speaker-in-a-room/

[30] Y. Lee et al., "Using sonar for liveness detection to protect smart speakers against remote attackers," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 4, no. 1, pp. 1–28, 2020.

[31] R. Nandakumar, A. Takakuwa, T. Kohno, and S. Gollakota, "CovertBand: Activity information leakage using music," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 3, pp. 1–24, 2017.

[32] W. Ruan, Q. Z. Sheng, L. Yang, T. Gu, P. Xu, and L. Shangguan, "AudioGest: Enabling fine-grained hand gesture detection by decoding echo signal," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2016, pp. 474–485.

[33] J. Dmochowski, J. Benesty, and S. Affès, "On spatial aliasing in microphone arrays," *IEEE Trans. Signal Process.*, vol. 57, no. 4, pp. 1383–1395, Apr. 2009.

[34] "Application Note #71: Multi-Tone: Testing, theory and practice." 2024. [Online]. Available: https://www.elektronikfokus.dk/wp-content/uploads/sites/5/AppNote71-MULTI-TONE-TESTING.pdf

[35] S. Zhang and A. Das, "HandLock: Enabling 2-FA for smart home voice assistants using inaudible acoustic signal," in *Proc. 24th Int. Symp. Res. Attacks, Intrusions Defenses*, 2021, pp. 251–265.

[36] S. Chakraborty, "Advantages of Blackman window over hamming window method for designing FIR filter," *Int. J. Comput. Sci. Eng. Technol.*, vol. 4, no. 8, pp. 1–9, 2013.

[37] A. Ram, S. Jalal, A. S. Jalal, and M. Kumar, "A density based algorithm for discovering density varied clusters in large spatial databases," *Int. J. Comput. Appl.*, vol. 3, no. 6, pp. 1–4, 2010.

[38] N. Rahmah and I. S. Sitanggang, "Determination of optimal Epsilon (EPS) value on DBSCAN algorithm to clustering data on peatland hotspots in sumatra," in *Proc. IOP Conf. Series Earth Environ. Sci.*, vol. 31, 2016, Art. no. 12012.

[39] S. V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, 3rd ed. Chichester, U.K.: Wiley, 2006.

[40] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[41] "GGMM D6 portable speaker for the echo dot." 2024. [Online]. Available: https://ideaing.com/product/ggmm-d6-portable-speaker

[42] "Tech specs: Elérhető," Raspberry pi. 2024. [Online]. Available: https://www.raspberrypi.org/products/raspberry-pi-4-model-b/specifications

[43] C. Shen, Z. Wang, C. Si, Y. Chen, and X. Su, "Waving gesture analysis for user authentication in the mobile environment," *IEEE Netw.*, vol. 34, no. 2, pp. 57–63, Mar./Apr. 2020.

[44] R. Nandakumar, V. Iyer, D. Tan, and S. Gollakota, "FingerIO: Using active sonar for fine-grained finger tracking," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2016, pp. 1515–1525.

[45] "Comparative examples of noise levels." IAC Acoustics. 2017. [Online]. Available: https://www.iacacoustics.com/blog-full/comparative-examples-of-noise-levels

[46] S. Stephenson, B. Pal, S. Fan, E. Fernandes, Y. Zhao, and R. Chatterjee, "SOK: Authentication in augmented and virtual reality," in *Proc. IEEE Symp. Security Privacy (SP)*, 2022, pp. 267–284.

[47] H. Li et al., "VocalPrint: Exploring a resilient and secure voice authentication via mmWave biometric interrogation," in *Proc. 18th Conf. Embedded Netw. Sens. Syst.*, 2020, pp. 312–325.

[48] L. Blue, H. Abdullah, L. Vargas, and P. Traynor, "2MA: Verifying voice commands via two microphone authentication," in *Proc. Asia Conf. Comput. Commun. Security*, 2018, pp. 89–100.

[49] H.-M. C. Leung, C.-W. Fu, and P.-A. Heng, "TwisTin: Tangible authentication of smart devices via motion co-analysis with a smartwatch," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 2, pp. 1–24, 2018.

[50] D. Wang, H. Cheng, P. Wang, X. Huang, and G. Jian, "Zipf's law in passwords," *IEEE Trans. Inf. Forensics Security*, vol. 12, pp. 2776–2791, 2017.

[51] D. Wang, Q. Gu, X. Huang, and P. Wang, "Understanding human-chosen pins: Characteristics, distribution and security," in *Proc. ACM Asia Conf. Comput. Commun. Security*, 2017, pp. 372–385.

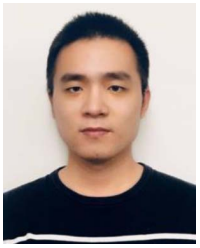
[52] N. Sae-Bae and N. Memon, "Distinguishability of keystroke dynamic template," *PLoS ONE*, vol. 17, no. 1, 2022, Art. no. e0261291.

[53] Y. Sutcu, E. Tabassi, H. T. Sencar, and N. Memon, "What is biometric information and how to measure it?" in *Proc. IEEE Int. Conf. Technol. Homeland Security (HST)*, 2013, pp. 67–72.

[54] R. Youmaran and A. Adler, "Measuring biometric sample quality in terms of biometric feature information in iris images," *J. Electr. Comput. Eng.*, vol. 2012, p. 22, Jul. 2012.
[55] A. Adler, R. Youmaran, and S. Loyka, "Towards a measure of biometric feature information," *Pattern Anal. Appl.*, vol. 12, no. 3, pp. 261–270, 2009.
[56] V. Krivokuca, M. Gomez-Barrero, S. Marcel, C. Rathgeb, and C. Busch, "Towards measuring the amount of discriminatory information in finger vein biometric characteristics using a relative entropy estimator," in *Handbook of Vascular Biometrics*. Cham, Switzerland: Springer, 2020, p. 507.
[57] K. Takahashi and T. Murakami, "A measure of information gained through biometric systems," *Image Vis. Comput.*, vol. 32, no. 12, pp. 1194–1203, 2014.
[58] H. Zhu, M. Xiao, D. Sherman, and M. Li, "SoundLock: A novel user authentication scheme for VR devices using auditory-pupillary response," in *Proc. NDSS*, 2023, pp. –18.
[59] "How to find your pin code." 2022. [Online]. Available: https://www.samsung.com/sg/support/tv-audio-video/how-to-find-your-pin-code/
[60] "How to set a screen lock on my device?" 2023. [Online]. Available: https://www.sony-asia.com/electronics/support/articles/SX671401
[61] "Reset your meta quest pin." 2023. [Online]. Available: https://www.meta.com/en-us/help/quest/articles/accounts/account-settings-and-management/reset-oculus-pin/
[62] "Use a passcode with your iPhone, iPad, or iPod touch." 2024. [Online]. Available: https://support.apple.com/en-us/HT204060
[63] "Set screen lock on an android device." 2024. [Online]. Available: https://support.google.com/android/answer/9079129
[64] Y. Meng et al., "Your microphone array retains your identity: A robust voice liveness detection system for smart speakers," in *Proc. 31st USENIX Security Symp. (USENIX Security)*, 2022, pp. 1077–1094.
[65] "Teach Google assistant to recognize your voice with voice match." 2024. [Online]. Available: http://surl.li/hrczg
[66] "Aware voice authentication." 2024. [Online]. Available: https://www.aware.com/voice-authentication/
[67] F. Hong, M. Wei, S. You, Y. Feng, and Z. Guo, "Waving authentication: Your smartphone authenticate you on motion gesture," in *Proc. 33rd Annu. ACM Conf. Extended Abstracts Human Factors Comput. Syst.*, 2015, pp. 263–266.
[68] Y. Song and Z. Cai, "Integrating handcrafted features with deep representations for smartphone authentication," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 6, no. 1, pp. 1–27, 2022.
[69] A. Lewis, Y. Li, and M. Xie, "Real time motion-based authentication for smartwatch," in *Proc. IEEE Conf. Commun. Netw. Security (CNS)*, 2016, pp. 380–381.
[70] W. Mao, M. Wang, W. Sun, L. Qiu, S. Pradhan, and Y.-C. Chen, "RNN-based room scale hand motion tracking," in *Proc. 25th Annu. Int. Conf. Mobile Comput. Netw.*, Oct. 2019, pp. 1–16.
[71] N. Zhu, H. Chen, and Z. Yang, "Fine-grained multi-user device-free gesture tracking on today's smart speakers," in *Proc. IEEE 18th Int. Conf. Mobile Ad Hoc Smart Syst. (MASS)*, Oct. 2021, pp. 99–107.
[72] H. Chen, F. Li, and Y. Wang, "EchoTrack: Acoustic device-free hand tracking on smart phones," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, May 2017, pp. 1–9.
[73] A. Wang and S. Gollakota, "MilliSonic: Pushing the limits of acoustic motion tracking," in *Proc. CHI Conf. Human Factors Comput. Syst.*, May 2019, pp. 1–11.
[74] D. Li, J. Liu, S. I. Lee, and J. Xiong, "FM-track: Pushing the limits of contactless multi-target tracking using acoustic signals," in *Proc. 18th Conf. Embedded Netw. Sens. Syst.*, Nov. 2020, pp. 150–163.
[75] D. Li, J. Liu, S. I. Lee, and J. Xiong, "LaSense: Pushing the limits of fine-grained activity sensing using acoustic signals," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 6, no. 1, pp. 1–27, 2022.
[76] D. Li, J. Liu, S. I. Lee, and J. Xiong, "Room-scale hand gesture recognition using smart speakers," in *Proc. SenSys*, 2022, pp. 462–475.
[77] E. A. Ibrahim, M. Geilen, J. Huisken, M. Li, and J. P. de Gyvez, "Low complexity multi-directional in-air ultrasonic gesture recognition using a TCN," in *Proc. Design, Autom. Test Europe Conf. Exhibition (DATE)*, 2020, pp. 1259–1264.
[78] "Surveymonkey." 2023. [Online]. Available: https://www.surveymonkey.com
[79] D. T. Campbell, "The indirect assessment of social attitudes," *Psychol. Bull.*, vol. 47, no. 1, p. 15, 1950.

Sungbin Park received the B.S. degree from Hanyang University, Ansan, Republic of Korea, in 2022. He is currently pursuing the Ph.D. degree with Hanyang University, Ansan, Republic of Korea.

His research interests include IoT security, usable security, and cybercrime.



Xueqiang Wang received the Ph.D. degree from Indiana University Bloomington, Bloomington, IN, USA, in 2021.

He was a Security Engineer with Amazon Lab126, Sunnyvale, CA, USA, and joined the University of Central Florida, Orlando, FL, USA as an Assistant Professor in October 2022. His research interests include software supply chain security, mobile/IoT security, and privacy analysis.

Kai Chen (Member, IEEE) received the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China, in 2010.

He joined Chinese Academy of Sciences, Beijing, in January 2010, where he became an Associate Professor in September 2012 and became a Full Professor in October 2015. His research interests include software analysis and testing, smartphones, and privacy.

Yeonjoon Lee received the B.S. degree from Hanyang University, Seoul, Republic of Korea, in 2012, and the Ph.D. degree in security informatics from Indiana University Bloomington, Bloomington, IN, USA, in 2019.

He is currently an Assistant Professor with the College of Computing, Hanyang University, Ansan, Republic of Korea. His research interests include mobile security, usable security, cybercrime, and AI security.