The Insecurity of Home Digital Voice Assistants – Vulnerabilities, Attacks and Countermeasures

Xinyu Lei*, Guan-Hua Tu*, Alex X. Liu*, Chi-Yu Li†, Tian Xie*
* Michigan State University, East Lansing, MI, USA
Email: leixinyu@msu.edu, ghtu@msu.edu, alexliu@cse.msu.edu, xietian1@msu.edu
† National Chiao Tung University, Hsinchu City, Taiwan
Email: chiyuli@cs.nctu.edu.tw

***Abstract*—Home Digital Voice Assistants (HDVAs) are getting popular in recent years. Users can control smart devices and get living assistance through those HDVAs (e.g., Amazon Alexa, Google Home) using voice. In this work, we study the insecurity of HDVA services by using Amazon Alexa and Google Home as case studies. We disclose three security vulnerabilities which root in their insecure access control. We then exploit them to devise two proof-of-concept attacks, home burglary and fake order, where the adversary can remotely command the victim's HDVA device to open a door or place an order from Amazon.com or Google Express. The insecure access control is that HDVA devices not only rely on a single-factor authentication but also take voice commands even if no people are around them. We thus argue that HDVAs should have another authentication factor, a physical presence based access control; that is, they can accept voice commands only when any person is detected nearby. To this end, we devise a Virtual Security Button (`VSButton`), which leverages the WiFi technology to detect indoor human motions. Once any indoor human motion is detected, the HDVA device is enabled to accept voice commands. Our evaluation results show that it can effectively differentiate indoor motions from the cases of no motion and outdoor motions in both laboratory and real world settings.**

I. Introduction

In recent years, more and more home digital voice assistant (HDVA) devices are deployed at home. Its number is forecasted to grow thirteen-fold from 2015 (1.1 million) to 2020 (15.1 million), a compound annual growth rate of 54.74% [1]. Thanks to the continuous efforts of the leading HDVA device manufacturers (e.g., Amazon and Google) and the third party voice service developers (e.g., CapitalOne, Dominos, Honeywell), users can do a great number of things using voice commands. They include playing music, ordering pizzas, shopping online, scheduling an appointment, checking weather, making a payment, controlling smart devices (e.g., garage doors, plug, thermostats), to name a few. To provide users with usage convenience, most of HDVA devices (e.g., Amazon Echo, Google Home) adopt an always-listening mechanism which takes voice commands all the time. Specifically, users are not required to press or hold a physical button on HDVA devices before speaking commands. This is the major difference between the HDVAs and phone assistants. The phone assistants are carried by users and only take voice commands after the phones are unlocked. Such great convenience may expose users to security threats due to the openness nature of voice channels. Therefore, we believe that the HDVA security should be specially considered. At this point, a natural question is: *Do these commercial off-the-shelf (COTS) HDVAs employ necessary security mechanisms to authenticate users and protect users from acoustic attacks?*

Unfortunately, our study on Amazon Alexa and Google Home yields a negative answer. We identify three security vulnerabilities from them and devise two proof-of-concept attacks. The victims may suffer from home security breach and fake order attacks. All the parties including the HDVA service provider (i.e., Amazon), HDVA devices, and the third party voice service developers, should take the blame. The Alexa and Google Home services employ only a single-factor authentication method based on a password-like voice word (e.g., "Alexa", "Hi, Google"). For any person/machine who speaks the correct authentication word ahead of a voice command, the command can be accepted by the HDVA devices. The HDVAs accept voice commands no matter whether any persons are around. It works for all the sounds whose sound pressure levels (SPL) are higher than 60dB. In addition, no access control is deployed at Alexa-enabled smart devices, since device vendors consider that all the voice commands from the Alexa service are benign. They are thus exposed to security threats once false voice commands can be delivered through the Alexa service.

At first glance, the remedy seems to be straightforward. HDVA devices shall authenticate users by their voice biometrics before taking voice commands. However, on the second thought, it may not be an easy task due to two reasons. First, users' voices may vary with their ages, illness, or tiredness. Second, human voice is vulnerable to replay attacks. Some of the prior works are proposed to deploy wearable devices for user authentication. For example, a study [2] develops a proprietary wearable device which collects the skin vibration signals of users. The collected vibration signals are then continuously matched with the voice signals received by HDVA devices. However, users may be reluctant to wear this device at home all the time. Another solution is to force users to press a physical button to explicitly activate Alexa devices before using them. Therefore, the main challenge for the remedy is *how to effectively secure HDVAs without scarifying user convenience or introducing extra deployment cost?*

After a careful study on the discovered security vulnerabilities, we observe that acoustic attacks are mainly launched while victims are not at home; otherwise, these acoustic

978-1-5386-4586-4/18/$31.00 ©2018 IEEE

attacks will be heard by the victims. If HDVA devices stop taking voice commands when there are no surrounding users, the adversaries' fraudulent voice commands would not be accepted. To this end, we propose to deploy a Virtual Security Button (`VSButton`) on the HDVA devices. The `VSButton` system leverages COTS WiFi infrastructure deployed at home to detect tiny indoor human motions (e.g., waving a hand). Once indoor motions are detected, `VSButton` enables the microphone of HDVA devices to take voice commands for a period of time (e.g., 1 minute). Our experimental results show that `VSButton` can accurately differentiate indoor motions from the cases of no motions and outdoor motions.

Due to the similarity of Amazon Alexa and Google Home, we mainly present the results of the former, which is more popular. All the findings can be applied to both of them unless it is explicitly specified. In summary, this paper makes three major contributions.

1) We unveil three vulnerabilities of HDVAs towards acoustic attacks. HDVA voice service providers merely employ a weak single-factor authentication for users; HDVA voice services do not have physical presence based access control; the HDVA third party voice service developers do not enforce security policies on their connected devices. We devise two proof-of-concept attacks (i.e., home burglary and fake order) based on the identified vulnerabilities.

2) We design and develop a novel security mechanism (Virtual Security Button) which improves the security of HDVA voice services without sacrificing user convenience. Our evaluation results show that small indoor human motions (e.g., wave a hand) can activate the HDVAs, whereas big outdoor motions (e.g, jump) do not. Our proposed remedy can be further applied to all home digital voice assistants with proper software upgrades.

3) Compared with previous WiFi-based gesture/motion recognition works ([3]–[6]), `VSButton` is designed to differentiate indoor and outdoor motions, and be resistant to environment changes. Two novel manners are introduced to achieve them. First, it augments the effects of indoor motions by carefully selecting principal components (see Section IV). Second, it dynamically adapts CSI baselines, which are used to indicate the conditions of no indoor motions, to different environments over time. Therefore, it does not require manual re-calibration for environment changes, but only the initial calibration. Both of two manners are absent in the prior studies.

II. Background: Amazon Alexa

In this section, we introduce Amazon Alexa devices and their common voice service model.

A. Alexa Devices

There are three kinds of Alexa devices: Amazon Echo, Amazon Tap, and Echo Dot. To support voice commands, they connect to a cloud-based Amazon voice service, *Alexa*. Amazon Echo always stays in a listening mode, so it does not take any voice commands until a voice word "Alexa" wakes it up. Every time it wakes up, it serves only one voice command and then returns to the listening mode. It appears as a 9.25-inch-tall cylinder speaker with a 7-piece microphone array. Amazon Tap is a smaller (6.2-inch-tall), portable device version with battery supply, but has similar functions. Echo Dot is the latest generation, which is a 1.6-inch-tall cylinder with one tiny speaker. Both the Echo Dot and the Amazon Echo, which require plug-in power supplies, are usually deployed at a fixed location (e.g., inside a room). In this work, we focus on the Echo Dot to examine the Alexa voice service.

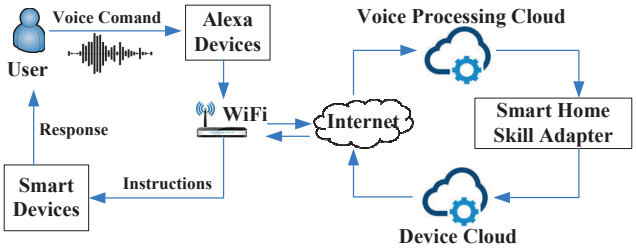


Fig. 1. Alexa voice service model.

B. Alexa Voice Service Model

The Alexa voice service supports the recognition of voice commands to Alexa devices. Figure 1 illustrates how the voice service works with Alexa devices to control smart home devices (e.g., smart bulb, thermostat, etc.). To control a smart device, a user can speak a voice command to an Alexa device after waking it up with voice "Alexa". The Alexa then sends the sounds of that voice command to a remote voice processing cloud via its connected WiFi network. Once the cloud recognizes the sounds as a valid command, it is forwarded to a server, called *smart home skill adapter*, which is maintained by Amazon to enable the cooperation with third-party service providers. Afterwards, that command is sent to another cloud which can control the corresponding smart device remotely. Note that in addition to the control of smart devices, some functions (e.g., checking the weather, placing orders on Amazon.com, etc.) provided by Alexa devices can also be accessed by voice commands.

III. Breaking Access Control in Alexa

Alexa voice service enables Alexa devices to receive voice commands. However, it has insecure access control so that malicious voice commands may be accepted. More threateningly, they may command smart home devices to assist in crimes. We here identify two vulnerabilities related to the Alexa's access control and one vulnerability existing in the Alexa-enabled smart devices. We then propose two proof-of-concept attacks. In the following, we start with our threat model.

Threat Model. The adversary has no access to the voice processing cloud, the smart home skill adapter, the device cloud, and smart devices. Victims are the owners of Alexa devices. The adversary does not require any physical access to the victims' Alexa devices or to be physically present nearby them. They may need to compromise indoor acoustic devices (e.g., Google Chromecast [7], answer machines [8],

or Bluetooth speakers [9]) to play arbitrary sounds at victims' home, but do not require to record, overhear, or replay victims' spoken voice commands.

A. (V1) Weak Single-factor Authentication

Alexa employs a single-factor authentication method based on a password-like voice word, to authenticate users who intend to access the voice service. The voice command recognition mechanism of Alexa services does not consider if the speakers are authorized users (e.g., the Alexa device owner or the owner's family members) but the semantics of the received voice commands. Therefore, any users who know the voice commands can access the Alexa services on behalf of the victims.

Validation. We validate that a valid voice command, which follows a correct authentication word, can be always accepted by an Alexa device, no matter where the voice comes. We consider voice from both human sounds and text to speech (TTS) service with natural sounding voices. We find that Alexa accepts voice commands from humans with different races, ages, and genders. And it also accepts voice commands from a variety of machines including computers (e.g., laptop and desktop), music/audio players (e.g., MP3 player, Bluetooth speaker, and home theater system), and mobile devices (e.g., smartphones and tablets).

B. (V2) No Physical Presence based Access Control

Alexa voice service is designed for the scenario that a user nearby his/her Alexa device speaks voice commands to request services from the device. Since it does not have physical presence based access control, the device can still accept voice commands even if no people are around it. It works for all the sounds reaching it at the sound pressure level (SPL) 60dB or higher. Therefore, the sounds from an adversary outside the owner's space or a speaker device may successfully deliver malicious voice commands to the Alexa device.

Validation. We validate this vulnerability by testing whether the sounds from a speaker device can successfully queue up requests in an Alexa device where nobody is nearby. We use a Belkin Bluetooth speaker to play commands around Alexa. Our finding is that the Alexa device accepts our voice commands from the speaker as long as the sound arriving at the Alexa device is loud enough.

C. (V3) Insecure Access Control on Alexa-enabled Devices

Alexa owners can control Alexa-enabled smart devices by speaking their names (e.g., "My Door") and commands (e.g., "Open") via Alexa devices. Most vendors allow the devices' default names to be replaced, but it is usually not mandatory. Since the device cloud accepts all the voice commands sent from the Alexa, the security threats caused by Alexa devices may propagate to them. This insecure access control can aggravate the damage caused by the Alexa's vulnerabilities.

Validation. To validate this vulnerability, we examine whether Alexa-enabled smart devices can be controlled via an Alexa device by their default voice names and commands. We find numerous feasible smart devices including Garageio Garage Door, Tp-link smart plug, Wemo smart switch, just to name a few.

D. Proof-of-concept Attacks

We now devise two practical attack cases based on the above vulnerabilities: home burglary and fake order. To launch these attacks, the adversary needs to spread his/her malware to discover suitable victims whose Alexa devices can be abused to receive fraudulent voice commands without the adversary's nearby presence. Note that our attack cases are used for the victims who are identified to be vulnerable by the malware, but not to target specific victims. Even though the percentage of possible victims to all Alexa owners is low, not only can those victims get unexpected damage, but also are the other owners still under the risk of the crimes. Before presenting the attacks, we seek to answer two major questions.

- How to abuse an Alexa device without being present nearby?
- How to enable victims discovery in the malware?

In the following, we present some example cases to show the attack feasibility, but the available avenues are not limited to them.

1) Abusing an Alexa device without being present nearby: One possible way to deliver sounds to an Alexa device without being present nearby is to exploit an indoor acoustic device which is placed in the same room. It can also be achieved by broadcasting from the radio or a fire fighting speaker in a house or a building. We here focus on the acoustic device.

Bluetooth speakers. The adversary outside the victim's house can connect his/her smartphone to the victim's Bluetooth speaker and then play an MP3 audio file of voice commands. To show the viability, we use a Nexus smartphone and a Belkin Bluetooth speaker, as well as generate the audio using the TTS system, in our experiments. It is observed that the Alexa takes the voice commands from the audio and acts accordingly.

Smart TVs. Smart TVs, which are getting popular in recent years, support to play streaming videos from Youtube.com. A user is allowed to cast a specific Youtube video from his/her mobile device (e.g., smartphone, tablet, etc.) to a smart TV through the home WiFi. The adversary can steal the victim's home WiFi password through the malware without root permission [10], [11], and then cast a video with voice commands to the victim's smart TV when next to his/her house. We confirm the viability using an LG smart TV. This attack can also happen occasionally, e.g., a TV show discussed that Echo automatically ordered dolls houses [12].

2) Discovering victims using the malware: The adversary can discover potential victims, which have deployed Alexa devices or/and Alexa-enabled smart devices at home, using the smartphone malware without root permission. Since both the Alexa devices and the smart devices need to connect to a WiFi network, the malware can scan the home WiFi network with which its smartphone host associates to detect whether they exist or not.

3) Two Attack Cases: We now describe two attacks based on the Alexa's vulnerabilities.

Home Burglary. An adversary can burglar the house of an Alexa device owner by requesting the Alexa to open a door via an Alexa-enabled smart lock. For example, an adversary can open a victim's Garageio garage door by using the default voice command, "Alexa, tell Garageio to open my door" [13].

Fake Order. The Alexa service may be abused to place a fake order of a device owner by the adversary, and then the owner may suffer from financial loss. For example, the adversary can deliver sounds to the Alexa device to place an order of any prime-eligible items on Amazon.com and then stealthily pick them up in front of the victim's house. Note that we have verified that for this kind of orders, Amazon carriers usually leave the ordered items to the front door without the need of a signature.

IV. Virtual Security Button (VSButton): Physical Presence based Access Control

We propose an access control mechanism which is based on physical presence to Alexa devices or other devices/services that require the detection of physical presence. It not only addresses V2, but also makes Alexa's authentication to be two-factor instead of current single factor (i.e., V1). We name this mechanism as virtual security button (VSButton), because whether physical presence is detected is like whether a virtual button is pushed. The access to a device/service with VSButton is not allowed when the virtual button is not in a push state (i.e., physical presence is not detected). As a result, this mechanism enables Alexa devices to prevent fraudulent voice commands which are delivered when no persons are nearby them.

The mechanism detects physical presence based on the WiFi technology. It results in negligible overhead on the Alexa device/service, because it reuses the user's existing home WiFi network and needs negligible change on how the user requests Alexa services. It does the detection by monitoring the channel state information (CSI) of the channel used by the home WiFi network. The CSI changes represent that some human motions happen nearby the Alexa device. Once any human motion is detected, the Alexa device is activated to accept voice commands. Therefore, the user just needs to make a motion (e.g., waving a hand for 0.2 meters) to activate the device before speaking his/her voice commands.

We believe that detecting human motions based on WiFi signals is a practical yet low-cost solution approach due to two reasons. First, home WiFi networks are commonly deployed, so no extra deployment cost is needed. Second, only a software upgrade is required for the Alexa devices, since all of them have been equipped with WiFi. Before presenting our VSButton design, we introduce the CSI primer and the CSI-based human motion detection, which is based on the multi-path/multi-reflection effects.

A. CSI Primer

It is commonly used to characterize the channel state properties of WiFi signals. Current WiFi standards like IEEE 802.11n/ac rely on the Orthogonal Frequency Division Multiplexing (OFDM) technique, which divides a channel into multiple subcarriers, and use the Multiple-Input Multiple-Output (MIMO) technology to boost speed. Each CSI value represents a subcarrier's channel quality (i.e., channel frequency response) for each input-output channel. The mathematical definition of the CSI value is presented below. Let $\mathbf{x}_i$ be the N_T dimensional transmitted signal and $\mathbf{y}_i$ be the N_R dimensional received signal in subcarrier number i. For each subcarrier i, the CSI information $\mathbf{H}_i$ can be obtained based on the following equation: $\mathbf{y}_i = \mathbf{H}_i\mathbf{x}_i + \mathbf{n}_i$, where $\mathbf{n}_i$ represents noise vector. Commodity WiFi devices can record thousands of CSI values per second from numerous OFDM subcarriers, so even for subtle CSI variations caused by a motion, the CSI values can still provide descriptive information to us.

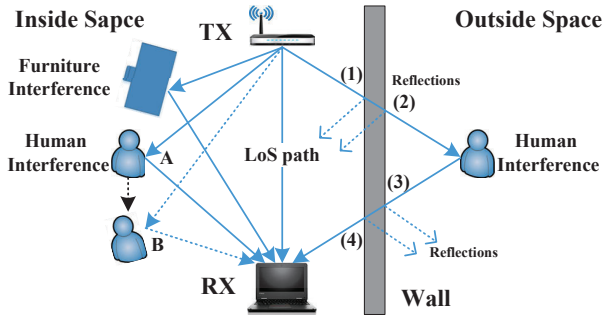


Fig. 2. An illustration of multi-path and multi-reflection effects.

B. CSI-based Human Motion Detection

We detect whether there are any human motions and identify whether they happen inside a house/room by leveraging the multi-path and multi-reflection effects on CSI, respectively.

Multi-path Effect for Human Motions Detection. The multi-path effect refers to the signal propagation phenomenon that a wireless signal reaches a receiving antenna along two or more paths. As shown in Figure 2, the receiver (RX) receives multiple copies from a common signal along multiple paths, the line-of-sight (LoS) path, one reflection from the human at location A, and one reflection from the furniture. Different lengths of the paths along which the WiFi signals are sent result in phase changes of the signals, thereby leading to various CSI values. Therefore, human motions can change the paths of WiFi signals and further make CSI values to change. For example, when the human moves from location A to location B, the new signal reflection path is being substituted for the original one. This move can thus distort the CSI values observed at the receiver.

Multi-reflection Effect for Identifying Where the Motions are. The multi-reflection effect means that a wireless signal may be reflected by multiple objects and it can cause the receiver to receive multiple copies of signals from different reflections. When we consider human motions inside and outside a room/house, they can be differentiated in terms of variation degrees of CSI values due to multiple different reflections. As shown in Figure 2, when a WiFi signal is reflected by the human body outside the wall, the signal

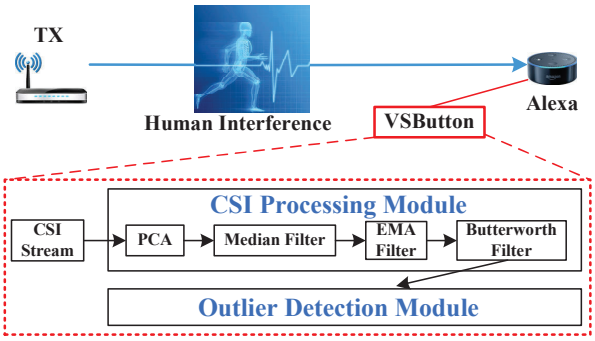


Fig. 3. VSButton design.

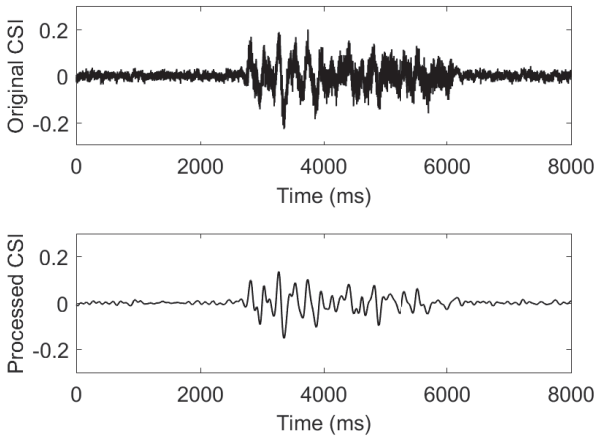


Fig. 4. Comparison between original/processed CSI over time.

arriving at the receiver along this path should have experienced four reflections: (1), (2), (3), and (4). They happen since the signal traverses different media, from air to wall and from wall to air. Such multi-reflection effect causes the signal received by the receiver to suffer from a serious attenuation. Our results show that it makes outside motions to result in only a small variation of CSI values, compared with a significant variation caused by inside human motions. The variation degrees of CSI values can thus be leveraged to identify the human motions occurring inside and outside the wall.

C. `VSButton` Design

In this section, we introduce our `VSButton` design as shown in Figure 3. It resides at the Alexa device and monitors human motions by its collected CSI values of the data packets received from the WiFi AP. Based on the CSI variations of the wireless channel between the device and the AP, `VSButton` can determine whether any human motion happens inside or not. Note that, in order to keep collecting CSI values over time, the Alexa device can send ICMP packets to the AP at a constant rate (e.g., 200 ICMP messages/second in our experiments) and then keep receiving the packets of ICMP reply messages from the AP.

The detection of inside human motions in the `VSButton` mainly consists of two phases, *CSI Processing Phase* and *Outlier Detection Phase*. When receiving CSI values, the former eliminates noises from them so that the CSI variation patterns caused by human motions can be augmented. Based on the output of the first phase, the latter relies on a real-time outlier detection method to detect the CSI patterns of the human motions inside a room/house. In the following, we present the details of each phase and then give an example of identifying indoor human motions.

1) CSI Processing Phase: This phase consists of three modules: principal component analysis (PCA) [14], median and exponential moving average (EMA) filter [15], [16], and Butterworth low-pass filter [17]. We first use the PCA module to reduce the dimensions of CSI values by removing those uncorrelated to motions detection. The last two modules are used to eliminate bursty noise and spikes, and filter out high-frequency noise in CSI stream values, respectively.

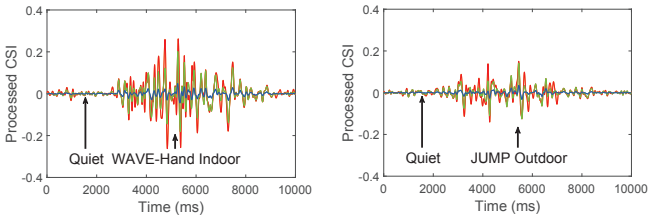
Figure 4 gives an example of the comparison between the original CSI over time and the CSI which has been processed by those three modules. It is shown that most of the noises in the original CSI are removed so that the processed one can give us more accurate information for motions detection. We elaborate the details of each module below.

PCA module. We employ the PCA to remove uncorrelated noisy information from the collected CSI by recognizing the subcarriers which have strong correlations with motions detection. The PCA is usually used to choose the most representative principal components from all CSI time series. It can accelerate the subsequent signal processing because the collected CSI may contain too much noisy information. In our experiments, it is observed that the first four components almost show the most significant changes in CSI streams but the first one is sensitive to signal noise of outdoor motions. As a result, we keep only the second, the third, and the fourth components for further analysis.

Median and EMA Filter Module. We next use a combination of a median filter and an EMA filter to eliminate bursty noise and spikes in CSI stream values. They happen because commodity WiFi interface cards may have a slightly unstable transmission power level and also be affected by dynamic channel conditions (e.g., air flow, humidity, etc.). Median filter can smooth out short-term fluctuations and highlight long-term trends. Note that the number of neighboring entries (i.e., moving window size) is a configurable parameter. We also adopt the EMA filter to smooth CSI values. It applies weighting factors which decrease exponentially to each older CSI value. The window size of EMA is a configurable parameter.

Butterworth Filter Module. We finally apply the Butterworth low-pass filter to filtering out high frequency CSI, since human motions cannot be generated too fast. Specifically, it is observed that the CSI variations caused by human motions mainly happen in the low frequency domain (i.e., less than 100 Hz). Given a proper cut-off frequency, 100 Hz, high frequency noise can be removed by the filter.

2) Outlier Detection Phase: We detect human motions using a real-time hyper-ellipsoidal outlier detection method over non-stationary CSI data streams [18]. It improves accuracy over the typical moving average method, which detects an



(a) Indoor motion CSI variation (b) Outdoor motion CSI variation

Fig. 5. Comparison of indoor and outdoor CSI variations.

anomaly based on a threshold of the distance between current value and the average of old values, from two aspects. First, it employs a new distance metric, i.e., *Mahalanobis distance*, which is more accurate. Second, it exploits exponential moving average (EMA) to update the previous sample mean.

This detection method can be mathematically described as follows. Let $X_k = \{x_1, \cdots, x_k\}$ be the first k samples of CSI readings. Each sample is a $d \times 1$ vector, where d is the number of chosen components. The sample mean m_k and sample covariance S_k are given by

$$m_k = \frac{1}{k}\sum_{i=1}^{k} x_i, \quad S_k = \frac{1}{k-1}\sum_{i=1}^{k}(x_i - m_k)(x_i - m_k)^T. \quad (1)$$

The Mahalanobis distance of a sample reading r from the X_k is defined as $D(r, X_k) = \sqrt{(r - m_k)^T S_k^{-1}(r - m_k)}$. By using Mahalanobis distance, we consider the reading r as an anomaly if $D(r, X_k) > t$, where t is a threshold parameter and needs to be carefully selected according to the experiments. All of the points bounded by $D(r, X_k) \leq t$ are considered as normal readings.

When we apply it to the non-stationary CSI streams, the sample mean is updated by $m_{k+1} = \alpha m_k + (1 - \alpha)x_{k+1}$, where $\alpha \in (0, 1)$ denotes a forgetting factor. The closer the receiving of a CSI value reading is, the larger weight it has to determine the next sample mean.

For the motions detection, we avoid false detections by specifying a threshold for the number of consecutive anomalous CSI values (10 is used in our experiments). It is because the noises may occasionally lead to some anomaly detections. However, human motions can make anomalous CSI value readings to last for a long period of time (e.g., consecutive 10 readings).

3) An Example of Identifying Indoor Motions: We show that the CSI values of indoor and outdoor motions can be clearly differentiated so that the indoor motions can be properly detected. Figures 5(a) and 5(b) plot the processed CSI values respectively for a small indoor motion (i.e., waving a hand) in one laboratory room and a large outdoor motion (i.e., jumping). It shows that even the small indoor motion can lead to the CSI variations which are much larger than those of the strong outdoor motion. Therefore, with proper parameter configurations, `VSButton` is able to correctly identify indoor motions and then activate an Alexa device to accept voice commands.

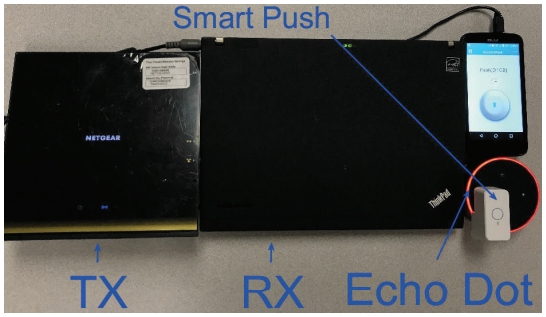


Fig. 6. `VSButton` prototype.

V. Implementation and Evaluation

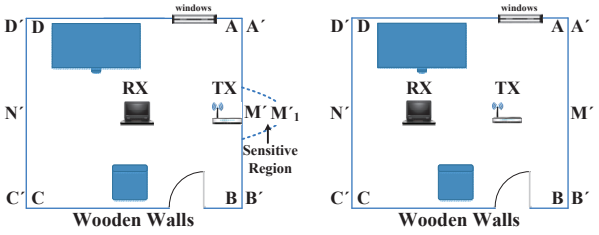
In this section, we present the implementation of our `VSButton` prototype and the evaluation results of both laboratory and real-world settings.

A. Prototype Implementation

Our `VSButton` prototype is based on commercial off-the-shelf (COTS) devices, as shown in Figure 6. The TX device is a Netgear R63000v2 WiFi router, which can be considered as a home WiFi AP. It is employed as the transmitter for the packets used by the Alexa device to collect CSI over time. The AP is set to the 802.11n mode [19], because the Alexa devices have not supported 802.11ac yet [20]. Since the Alexa devices are not open to development, we implement the motions detection module on a laptop, which is tagged to be RX. The Alexa device, Echo Dot, can then get motions detection result from the laptop. The RX device is a Lenovo X200 laptop equipped with an Intel Link 5300 WiFi adapter, which is able to collect CSI values using the tool developed by the work [21]. To emulate the Alexa device's access control, the module controls MicroBot Push [22] to turn on/off the Alexa device's microphone through a smartphone, once any status change of access control is detected by the module based on motions detection. Note that in our current `VSButton` prototype, there are a wireless router, a laptop, a smartphone, and a MicroBot. Seemingly, the deployment cost is not small. However, most of people have deployed a wireless router at home for Internet access and the others' functions can be integrated to the Alexa devices based on only software upgrades.

B. Evaluation

We next introduce our experimental settings and evaluate the performance of our `VSButton` prototype in three space settings: *square room*, *rectangle room*, and *two-bedroom apartment*. The performance refers to whether three cases, no motion, indoor motion, and outdoor motion, can be correctly identified. We recruit six volunteers to participate in the experiments. They are required to do three motions including waving a hand (WAVE-HAND), sitting down and standing up (SIT-DOWN-STAND-UP), and jumping (JUMP, 0.5m), inside and outside a room. They represent three degrees of human motions, weak, medium, and strong, respectively. Note that we examine the Mahalanobis distance for each measurement and see whether indoor motions can be clearly detected or not.



(a) Configuration 1 (b) Configuration 2

Fig. 7. Square room: two deployment configurations.

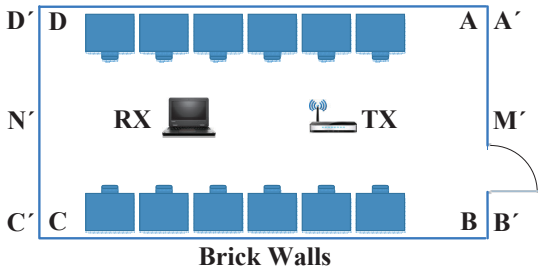


Fig. 8. Rectangle Room Configuration.

1) Experimental Settings: In all experiments, the RX (the laptop with an Echo dot) sends 200 ICMP `Echo Request` messages per second to the TX (the WiFi router) so that it can keep collecting CSI over time from the ICMP `Echo Reply` messages sent by the TX. A CSI stream with the sampling rate of 200 values per second can thus be used for motions detection. Each ICMP message size is only 84 bytes, so network bandwidth is low.

The window sizes of the median and EMA filters are set to 9 and 15, respectively. Our experimental results show that these two numbers are large enough for the filters to remove noise. The cut-off frequency [17] for the Butterworth filter is set to $\omega_c = \frac{2\pi \times 100}{200} = 1\pi$ rad/s[1], because human motions lead to only low-frequency CSI variations, which are typically less than $f = 100$ Hz. In the outlier detection module, we set the forgetting factor α to be 0.98. It means that we give a larger weight to the recent CSI value readings.

2) A Square Lab Room: We deploy the `VSButton` in a wooden square room and evaluate it with two deployment configurations as shown in Figure 7. In the first configuration, the laptop with an Echo dot (RX) is placed at the center of the room and the WiFi router (TX) is located at the edge. In the second configuration, the RX and the TX are placed between locations N' and M' to divide the distance into three equidistant portions. In the experiments, the six participants do the aforementioned three motions (i.e., WAVE-HAND, SIT-DOWN-STAND-UP, and JUMP) at four indoor locations (A, B, C, and D) and six outdoor locations (A', B', C', D', M', and N').

Configuration 1. The Mahalanobis distances we measured are summarized in Table I. Note that each number of the indoor results is the minimum value of all the numbers measured among the participants, whereas that of the outdoor results is the maximum. This way can easily show whether indoor and outdoor motions can be clearly differentiated based on the Mahalanobis distances or not. We observe that all the indoor motions can be differentiated from no-motion cases and outdoor motions at all the locations except Location M'. Some indoor motions, such as the WAVE-HAND motion at Location D, have smaller distances than the motions, JUMP and SIT-DOWN-STAND-UP, at Location M'. The main reason is that that location is very close to where the WiFi router is deployed. As a result, the router shall not be deployed at the location close to the wall next to outdoor space.

[1]rad/s is the unit of rotational speed (angular velocity)

Configuration 2. The results of this configuration are summarized in Table II. We observe that the distance of each indoor motion is higher than the maximum distance (i.e., 0.241 from JUMP at Location M')of all the outdoor motions. It represents that `VSButton` can activate the Alexa service only due to indoor motions with this configuration.

3) A Rectangle Lab Room: We now evaluate the prototype in a rectangle room with brick walls. The RX and the TX are placed between Locations N' and M' to divide the distance into three equidistant portions, as shown in Figure 8. Table III summarizes the measurement results. Note that the distance at each indoor location is the minimum value of all the numbers measured among the participants, whereas that at each outdoor location is the maximum.

There are two findings. First, the result is similar to that of the square room with the same configuration (i.e., Configuration 2). The minimum Mahalanobis distance (i.e., 0.147 from WAVE-HAND at Location A) among all indoor motions is higher than the maximum distance (i.e., 0.042 from JUMP at Location M') of all outdoor motions. Their difference is as large as 0.105. Second, that difference is higher than that (i.e., 0.071) observed in the square room with the same configuration. The main reason is that WiFi signals are unable to penetrate the brick walls as easy as wooden walls.

There are two lessons learned. First, the wall materials can influence the performance of `VSButton`. The harder the wall materials are, the better performance the `VSButton` can get. Our experimental results show that `VSButton` can be applied to two kinds of wall materials, wood and brick. Second, users should not deploy `VSButton` and WiFi router at the location which is close to outdoor space. Note that though `VSButton` has some limitations/preferences of the locations of WiFi routers deployed (e.g., the center of a room is recommended), users may also benefit from this, e.g., the coverage of WiFi signal will be larger than that which WiFi router is deployed close to outdoor space.

Parameters Calibration. Before deploying the `VSButton` in a real-world scenario, we need to perform parameter calibration to determine a proper threshold t for the outlier motion detection module. It is because the threshold can change with different environments. The calibration process includes two major steps. First, the Alexa owner chooses an indoor location to deploy his/her Alexa device and then determines which area is allowed to enable the device with human motions. At that location, s(he) does the smallest indoor motion (e.g., waving a hand) and collects its minimum Mahalanobis distance value.

Square Room	**Indoor Locations**				**Outdoor Locations**					
locations	A	B	C	D	A'	B'	C'	D'	M'	N'
WAVE-HAND	0.218	0.213	0.195	**0.191**	0.104	0.101	0.079	0.083	0.156	0.121
SIT-DOWN-STAND-UP	0.277	0.271	0.258	0.253	0.118	0.113	0.088	0.092	**0.238**	0.139
JUMP	0.392	0.391	0.371	0.366	0.132	0.128	0.099	0.103	**0.373**	0.165
DO NOTHING	0.026	0.021	0.027	0.024	0.023	0.027	0.028	0.023	0.020	0.023

TABLE I. MAHALANOBIS DISTANCE MEASURED IN A SQUARE ROOM WITH CONFIGURATION 1.

Square Room	**Indoor locations**				**Outdoor locations**					
locations	A	B	C	D	A'	B'	C'	D'	M'	N'
WAVE-HAND	**0.312**	0.315	0.401	0.409	0.041	0.043	0.049	0.051	0.092	0.063
SIT-DOWN-STAND-UP	0.345	0.349	0.423	0.430	0.060	0.062	0.069	0.071	0.121	0.089
JUMP	0.401	0.407	0.451	0.459	0.069	0.071	0.084	0.086	**0.241**	0.099
DO NOTHING	0.025	0.021	0.022	0.024	0.028	0.026	0.021	0.022	0.023	0.025

TABLE II. MAHALANOBIS DISTANCE MEASURED IN A SQUARE ROOM WITH CONFIGURATION 2.

Rectangle Room	**Indoor locations**				**Outdoor locations**					
locations	A	B	C	D	A'	B'	C'	D'	M'	N'
WAVE-HAND	**0.147**	0.150	0.180	0.183	0.020	0.022	0.025	0.027	0.035	0.030
SIT-DOWN-STAND-UP	0.181	0.184	0.216	0.217	0.024	0.026	0.028	0.029	0.039	0.033
JUMP	0.254	0.255	0.287	0.288	0.029	0.029	0.032	0.033	**0.042**	0.035
DO NOTHING	0.022	0.021	0.022	0.027	0.028	0.026	0.021	0.022	0.020	0.025

TABLE III. MAHALANOBIS DISTANCE MEASURED IN A RECTANGLE ROOM.

Second, the owner finds all the outdoor locations which are not allowed to enable the Alexa device. S(he) does the strongest outdoor motion (e.g., jumping) and collects its maximum Mahalanobis distance value. We then set the threshold t to be the half of the difference between the above two distance values. Note that the whole calibration process can be done within 5 minutes and only one-time calibration is needed for the initial deployment of the Alexa devices. Note that our solution doesn't require manual re-calibration, but only initial calibration. It is because all the parameters are optimized and fixed in our design, but only the threshold t and the CSI baselines need to be adapted to environment changes. Specifically, the threshold t is affected only by wall materials and room layout, so it requires only initial calibration if the wall is not altered. The CSI baselines, which are used to indicate the conditions of no indoor motions, are automatically, dynamically adapted to different environments (e.g., moving a table nearby).

VI. DISCUSSIONS

We next discuss some limitations of `VSButton`.

Motions not from Humans. In our current prototype, we do not consider the motions which are not from humans. For example, the `VSButton` may activate the HDVA service due to the jump of a pet when the owner is not around. In our future work, we plan to develop a pet-immune `VSButton` system.

WiFi Hijacking Attack. The attacker may compromise the WiFi router and control the transmission power to cause a large WiFi signal variance which activates `VSButton`. However, this attack may not be very practice since it requires a strong attack assumption. The adversary has to obtain the administrator username and password of the victim's WiFi router, and further control the router's transmission power.

Tradeoff between Security and Convenience. `VSButton`'s resistance to malicious outdoor motions is a tradeoff between the security and user convenience. To accommodate a variety of user demands and use environment, it is configurable by our design. For example, by increasing the threshold *t*, `VSButton` has a strong resistance to outdoor motions whereas users have a shorter communication range with their HDVA devices.

Physical Invasive Attack. Our threat model assumes that the adversaries are from the outside space and cannot physically break into the room. This is because if the adversary can invade the victim's house, s(he) is able to do many things much eviller than attacking the HDVA.

VII. RELATED WORK

WiFi-based Indoor Human Activity/Motion Sensing. There have been several related works [3]–[6] which study how to detect indoor human motions or activities by WiFi technologies. Kosba et al. [3] propose to use RSS (received signal strength) to detect human motions by RASID. Due to the limitation of RSS (i.e., providing coarser granularity of wireless channel information than CSI), RASID is mainly designed to detect relative larger motions, e.g., walking, instead of small motions, e.g., waving a hand. Pu et al. [5] leverages Doppler shift to recognize nine whole-home gestures, requiring specialized receiver that extracts carrier wave features that are not reported in current WiFi systems. However, `VSButton` is developed on top of off-the-shelf equipments. Wang et al. [6] leverages CSI to detect fine grained human activities (e.g., walking or cooking). Those CSI-based motion recognition techniques are sensitive to the changes of the environment (e.g., moving the furniture) and the locations of users. If the usage environment changes, they require a re-calibration which is inconvenient for HDVA users.

Mangled & Inaudible Voice Attack. Mangled voice attack is first proposed in [23] and further developed in [24]. They show

that the software voice assistant (Google Voice) on phones can receive voice commands that are unrecognizable to human but interpretable by the voice assistants. The attack works if adversaries (speakers) to be not more than 3.5 meters far away from the phones and the victims do not notice the hearable mingle voice commands. Different from them, we study home digital voice assistants instead of software voice assistants on smartphones. Their use scenarios and security issues are different. For example, users usually have their smartphones with them, whereas HDVA users leave the home DAV devices at home. The inaudible voice commands attacks towards Amazon Echo is provided in [25]. The attacks require two strong perquisites, which may be barely satisfied in practice. First, the inaudible voice commands need to be generated from a customized ultrasound microphone. Second, the customized microphone requires to be deployed next to the Alexa device within 2 meters.

Voice Authentication for Digital Voice Assistants. The Biometric-based voice authentication [26]–[29] has been proposed to protect HDVA devices against acoustic attacks. However, they still have some challenges to address. For example, human voice changes with age, illness and tiredness. This approach may require a re-training process periodically. In addition to biometric-based voice authentication, a study [2] proposed a solution that needs a DVA user to wear a voice-enabled device; the approach in [29] requires the user to carry a smartphone to help user authentication. Different from them, `VSButton` tackles the above challenges while preserving user convenience. It enables the HDVA to take voice commands only when users can prove that they are around where the HDVA is. Therefore, adversaries are unable to launch remote acoustic attacks.

VIII. Conclusions

In this work, we identify several security vulnerabilities of HDVAs by considering Amazon Alexa and Google Home as case studies. Surprisingly, the Alexa and Google Home services rely on only a weak single-factor authentication, which can be easily broken. To secure the HDVA services, we seek to propose an additional factor authentication, physical presence. An HDVA device can accept voice commands only when any person is physically present nearby. We thus design a solution called virtual security button (`VSButton`) to do the physical presence detection. We prototype and evaluate it on an Alexa device. Our experimental results show that `VSButton` can do accurate detection in both laboratory and real-world home settings. We hope our initial efforts can stimulate further research on the HDVA security.

References

[1] D. Watkins, "Strategy analytics: Amazon, google to ship nearly 3 million digital voice assistant devices in 2017," https://www.strategyanalytics.com/strategy-analytics/news/strategy-analytics-press-releases/strategy-analytics-press-release/2016/10/05/strategy-analytics-amazon-google-to-ship-nearly-3-million-digital-voice-assistant-devices-in-2017#.WQtiXeXyuUk, 2016.

[2] H. Feng, K. Fawaz, and K. G. Shin, "Continuous authentication for voice assistants," in *MobiCom*, 2017.

[3] A. E. Kosba, A. Saeed, and M. Youssef, "Rasid: A robust wlan device-free passive motion detection system," in *PerCom*, 2012.

[4] Y. Zeng, P. H. Pathak, C. Xu, and P. Mohapatra, "Your ap knows how you move: Fine-grained device motion recognition through wifi," in *HotWireless*, 2014.

[5] Q. Pu, S. Gupta, S. Gollakota, and S. Patel, "Whole-home gesture recognition using wireless signals," in *Mobicom*, 2013.

[6] Y. Wang, J. Liu, Y. Chen, M. Gruteser, J. Yang, and H. Liu, "E-eyes: device-free location-oriented activity identification using fine-grained wifi signatures," in *MobiCom*, 2014.

[7] D. Petro, "Rickrolling your neighbors with google chromecast," in *Black Hat*, 2014.

[8] "Hacking answering machines," https://www.youtube.com/watch?v=bq6aV0Cxhl0, 2012.

[9] "How to prevent unauthorized access to bluetooth speakers?" https://superuser.com/questions/548592/how-to-prevent-unauthorized-access-to-bluetooth-speakers, 2013.

[10] "How to recover wifi password in android without root?" http://www.viralhax.com/recover-wifi-password-android/, 2017.

[11] "Default username and password of wireless routers," http://www.routerpasswords.com/, 2017.

[12] "Amazon echo rogue payment warning after tv show causes 'alexa' to order dolls houses," http://www.telegraph.co.uk/news/2017/01/08/amazon-echo-rogue-payment-warning-tv-show-causes-alexa-order/, 2017.

[13] "Work with garageio," https://garageio.com/workswith/echo, 2017.

[14] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002.

[15] G. R. Arce, *Nonlinear signal processing: a statistical approach*. John Wiley & Sons, 2005.

[16] C. C. Holt, "Forecasting seasonals and trends by exponentially weighted moving averages," *International journal of forecasting*, vol. 20, no. 1, pp. 5–10, 2004.

[17] S. Butterworth, "On the theory of filter amplifiers," *Wireless Engineer*, vol. 7, no. 6, pp. 536–541, 1930.

[18] M. Moshtaghi, C. Leckie, S. Karunasekera, J. C. Bezdek, S. Rajasegarar, and M. Palaniswami, "Incremental elliptical boundary estimation for anomaly detection in wireless sensor networks," in *ICDM*, 2011.

[19] "Ieee std. 802.11n-2009: Enhancements for higher throughput," http://www.ieee802.org, 2009.

[20] "Connect echo dot to wi-fi," https://www.amazon.com/gp/help/customer/display.html?nodeId=202011800, 2017.

[21] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "Tool release: Gathering 802.11 n traces with channel state information," in *SIGCOMM*, 2011.

[22] "Microbot push," https://prota.info/, 2017.

[23] T. Vaidya, Y. Zhang, M. Sherr, and C. Shields, "Cocaine noodles: exploiting the gap between human and machine speech recognition," in *WOOT*, 2015.

[24] N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou, "Hidden voice commands," in *USENIX Security*, 2016.

[25] L. Song and P. Mittal, "Inaudible voice commands," https://arxiv.org/pdf/1708.07238.pdf, 2017.

[26] A. Das, O. K. Manyam, M. Tapaswi, and V. Taranalli, "Multilingual spoken-password based user authentication in emerging economies using cellular phone networks," in *SLT*, 2008.

[27] M. Kunz, K. Kasper, H. Reininger, M. Möbius, and J. Ohms, "Continuous speaker verification in realtime." in *BIOSIG*, 2011.

[28] M. Baloul, E. Cherrier, and C. Rosenberger, "Challenge-based speaker recognition for mobile authentication," in *BIOSIG*, 2012.

[29] L. Zhang, S. Tan, and J. Yang, "Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication," in *ACM CCS*, 2017.