Traversing the Quagmire that is Privacy in your Smart Home

Chuhan Gao*
University of Wisconsin-Madison

Varun Chandrasekaran*
University of Wisconsin-Madison

Kassem Fawaz
University of Wisconsin-Madison

Suman Banerjee
University of Wisconsin-Madison

ABSTRACT

Voice has become an increasingly popular User Interaction (UI) channel, with voice-activated devices becoming regular fixtures in our homes. The popularity of voice-based assistants (VAs), however, have brought along significant privacy and security threats to their users. Recent revelations have indicated that some VAs record user's private conversations continuously and innocuously. With the VAs being connected to the Internet, they can leak the recorded content without the user's authorization. Moreover, these devices often do not pack authentication mechanisms to check if the voice commands are issued by authorized users. To address both shortcomings, we propose a framework to impose a security and privacy perimeter around the user's VA. Our proposed framework continuously jams the VA to prevent it from innocuously recording the user's speech, unless the user issues a voice command. To prevent unauthorized voice commands, our framework provides a scheme similar to two-factor authentication to only grant access when the authorized user is in its vicinity. Our proposed framework achieves both objectives through a combination of several techniques to (a) continuously jam one (or many) VA's microphones in a manner inaudible to the user, and (b) provide only authenticated users *easy* access to VAs.

CCS CONCEPTS

• **Security and privacy** → *Usability in security and privacy*;

KEYWORDS

Privacy, voice assistant, smart home, authentication, ultrasound jamming

ACM Reference Format:
Chuhan Gao, Varun Chandrasekaran, Kassem Fawaz, and Suman Banerjee. 2018. Traversing the Quagmire that is Privacy in your Smart Home. In *IoT S&P'18: ACM SIGCOMM 2018 Workshop on IoT Security and Privacy , August 20, 2018, Budapest, Hungary.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3229565.3229573

*Both authors contributed equally to the paper

IoT S&P'18, August 20, 2018, Budapest, Hungary

ACM ISBN 978-1-4503-5905-4/18/08…$15.00
https://doi.org/10.1145/3229565.3229573

1 INTRODUCTION

Advances in the various facets of machine learning and natural language processing have ensured that *smart devices* are increasingly permeated into our lives and our homes. These devices aid users in tasks ranging from automated management control using smart plugs to providing various forms of home automation based on a user's vocal input. Among these devices, voice assistants (VAs) are the most commonly used. As these devices often lack the computational power to process the user's speech, they constantly communicate with a cloud counterpart to provide the desired processed output. While these devices undoubtedly make users' lives easier, they do pose a significant privacy risk. They constantly record user's data and potentially share it with a cloud provider for subsequent processing. For example, Samsung indicated that its smart TVs not only record private conversations in their vicinity, but also pass them on to third parties [1]. Recording this form of information has severe consequences to the privacy of individuals[3, 27].

Apart from the privacy violations, VAs also introduce new security threats to the user's home. The majority of such smart devices have no mechanism for user authentication other than simple voice recognition. If such smart devices are integrated in a smart home environment, an unauthorized user can control sensitive home automation system; for example, an adversary can issue a command to open the garage door while standing outside the home.

To make matters worse, a home environment may have several such VAs, each of which pose potential privacy and security threats. For example, a home may be equipped with an Echo speaker [4], a Google Home assistant [7] and a smart TV [12], each independently recording private user conversations. With each of these devices controlling different aspects of the user's home and information, unauthorized access exacerbate the potential security problems. In this paper, we propose a framework to preserve user privacy, and provide greater security without sacrificing usability of these VAs. Our framework achieves the following goals: *(i) Preventing innocuous recording of user data on various voice-enabled devices, and (ii) preventing unauthorized users from accessing these smart devices.* Achieving both these goals is challenging as these VAs are typically closed black-boxes; these devices usually can't be modified or altered in any way.

To prevent data collection, an intuitive solution is to throttle the uplink network flow from the VA *i.e.* restrict the network traffic from the device to the Internet. In addition to degrading the functioning of the VA, this solution does not prevent the VA from caching the data and slowly streaming it to the cloud later. We propose a more proactive approach to prevent innocuous recording through ultrasound jamming. The authors of [33, 34, 36] note that

non-linearity in microphones used in VAs can be used to record inaudible (to the human ear) signals.

Such signals can also be used to jam the VA's microphone *i.e.* prevent it from recording ongoing user conversation. By generating and propagating customized signals from a trusted device, over multiple frequencies, one can jam multiple devices at the same time. By solving the innocuous recording problem, we introduce another: *how do we provide legitimate access to a jammed VA device in a usable manner?* To address this problem, we equip the jamming device (obfuscator) with a microphone to identify voice commands and lift the jamming accordingly. Note that achieving speech recognition at the obfuscator is challenging as the naive omnidirectional jamming signal will cause self interference at the obfuscator. Performing signal correction at the obfuscator, can approximate a noisy input signal. We, however, propose an alternate approach - utilizing *ultrasound* microphones at the obfuscator (see §5) which are immune to the ultrasound jamming.

Existing solutions for authenticating users to VAs primarily revolve around biometric authentication, such as voice-based authentication. However, it is trivial for any unauthorized party to simply record and replay a user's voice passphrase. Additionally, these forms of authentication are non-resilient *i.e.* there is no easy form of recovery if the adversary records the user's voice to replay or synthesize a command to compromise the VA. Other solutions for voice authentication are through the use of trusted hardware, such as tokens [13], or wearables [24] which hinder usability, as the user is now expected to purchase additional hardware. Our framework mitigates these problems by utilizing the user's smartphone to participate in a simple challenge-response protocol to authenticate the legitimate user (see §6).

To summarize, our framework (i) prevents innocuous user recording by a wide array of VAs, and (ii) prevents unauthorized users from injecting commands to VAs. For the first objective, we deploy an audio obfuscator that plays specially designed ultrasound signal to jam the VAs' microphones, which prevents the VA from recording in the human speech frequency range. The obfuscator is equipped with an ultrasound microphone, which allows it to hear the human speech without being affected by the obfuscation signal. To achieve the second objective, we propose a scheme similar to traditional two-factor authentication through a challenge-response mechanism (on a different, secure channel). This mechanism is subsequently hosted on the smartphone carried by legitimate/authorized users, in the form of a background application.

While these techniques are not entirely original, piecing them together to create a usable, privacy preserving framework is the contribution of this work. We begin by discussing the smart home ecosystem and our threat model in §2. In §3, we provide an overview of the proposed framework. In the remainder of the paper, we discuss the technical details to each of the solutions proposed. We conclude with a discussion of future research directions that stem from our solution in §7, and a brief survey of related works in §8.

2 SMART HOME ECOSYSTEM THREATS

Before we elaborate on the solutions to the requirements discussed earlier, we describe the smart home ecosystem. We consider a home environment populated with various forms of voice-based assistants (VAs) devices such as smart speakers and televisions. While there are other forms of IoT devices such as smart cameras, we restrict our discussion to those devices which respond to audio input only. These VAs are black-boxes, meaning their internal hardware/software configurations cannot be modified. They might innocuously record user activities (audio signals in this specific case) for subsequent profiling of the users. Additionally, there could be unauthorized actors[1] in such an ecosystem attempting to unauthorized use of these smart devices.

In this ecosystem, the first adversary is the VA itself. The VA might innocuously record the user's private speech using its multi-directional microphones. The second adversary in this ecosystem is the presence of an unauthorized user who can command the VA unbeknownst to the user. We specify the properties of the second adversary below:

VA Access. We assume that the second adversary may target any VA of her choice, and has direct access to the device. She can physically touch the device, but can not alter the device settings, or install malware. We also assume that she is fully aware of the characteristics of the target devices. However, we assume that the adversary is not present during configuration of the device, i.e. at the time of first use of the device.

No Owner Interaction. We assume that the target device(s) may be in the owner's vicinity, but may not be in use. In addition, the device may be unattended, which can happen when the owner is temporarily away.

Inaudible. Since the goal of an adversary is to inject voice commands without being detected, she will use the sounds inaudible to human, *i.e.*, ultrasound signals.

Attacking Equipment. We assume that the adversary can acquire both the speakers designed for transmitting ultrasound and commodity devices for playing audible sounds. For instance, she may secretly leave a remote controllable speaker around the victim's desk or home. Alternatively, she may be carrying a portable speaker while walking by the victim.

3 SYSTEM OVERVIEW

Our proposed framework employs two physical components: an audio obfuscator that prevents the VA from performing innocuous recording, and the user's smartphone that allows the user to interact with the obfuscator. The obfuscator is a trusted hardware device deployed in user's home environment with the purpose of jamming the audio inputs of any VAs in its vicinity. Additionally, the obfuscator is not connected to the Internet and communicates only with the user's smartphone through a secure out-of-band (OOB) channel, such as Bluetooth. The obfuscator is equipped with an ultrasound speaker and a microphone. Our framework only requires the smartphone to install an app that allows it to interact with the obfuscator through the OOB channel. In order to create a private and secure voice-controlled environment, our proposed framework aims to accomplish the following three objectives:

- Preventing the VA(s) from hearing any private dialogue or sound other than the intended voice commands.

[1] Henceforth referred to as the adversary

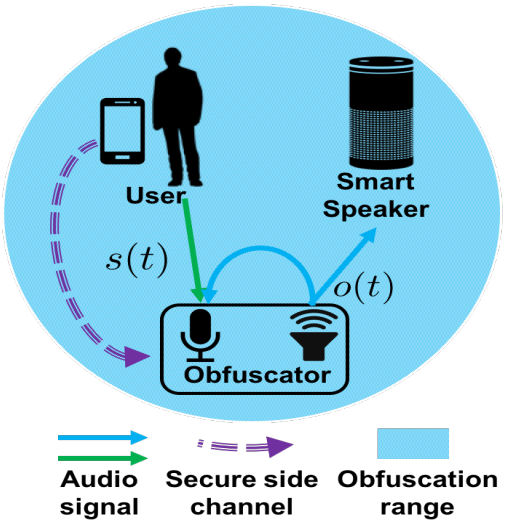

Figure 1: The obfuscator jams the smart speaker's microphone with inaudible obfuscation sound, while listening to user's hotword to lift the jamming. User authentication is achieved by a secure side channel.

- Preserving intended voice access to the VA(s), by allowing it to respond to intended voice commands following certain hotwords, such as "Alexa", or "OK Google". During the intended voice command, the obfuscator "lifts" the jamming signal.
- Realizing user authentication to prevent unauthorized users from issuing any voice command(s) to the VA(s).

To achieve these objectives, we leverage three different techniques. There are a number of ways to physically prevent the VA from recording any sound in the environment, such as powering off the device or turning off the microphone via on-device switches. However, such solutions force the user to physically access the device every time she wishes to issue a voice command, degrading its usability. To prevent innocuous recording with minimal user intervention, we adopt an audio obfuscation technique that injects audio jamming signals to microphones by playing human-inaudible ultrasound [33, 34, 36]. Both BackDoor [33] and DolphinAttack [36] utilize the non-linearity that exists in all microphones to make the microphone record sound in the human-audible range by playing an inaudible ultrasound signal. Such non-linearity exists outside the microphones' recording frequency range, which is usually below 22kHz. Within the audible frequency range, the microphone hardware primarily exhibits linear frequency response and records normally. Passing through the non-linear components and power amplifier in a microphone, a pair of ultrasound tones (*e.g.* 40kHz and 50kHZ) generate a "shadow" in low frequency (*e.g.* 10kHz), which is recorded by the microphone. By manipulating the frequency and amplitude of the ultrasound tones, one could control the frequency and amplitude of the shadow tone, which can be used to jam unauthorized recording of a private conversation, or even achieve human-inaudible audio communication. By leveraging this technique, we are able to use the human-inaudible sounds played by the obfuscator to jam the microphone of the smart speaker.

When the obfuscator is actively jamming the VA, there needs to be a way to allow the legitimate user issue voice commands to the VA. Ideally, the obfuscation should be constantly active, and deactivated during the user's intended voice commands automatically. To identify the user's intent to issue voice commands, we utilize the obfuscator's microphone. We leverage the fact that most smart speakers require the user to speak a hotword, such as "Alexa" or "OK Google" to indicate the start of a command. The obfuscator listens to such hotwords to determine when to lift the obfuscation signal. To prevent the obfuscator's microphone from being jammed as well, we choose to use ultrasound microphones with a recording frequency range covering the obfuscation signal. Since the obfuscator's microphone still operates linearly at this frequency, the obfuscation signal does not get aliased to lower frequencies as in the smart speaker. Instead, the audible frequency part of the received signal won't be affected by the obfuscation. The obfuscator, lacking any connection to the Internet, will not leak any private speech it hears. The user's speech can be processed locally as the obfuscator only needs to recognize the hotwords, further validating the decision to keep it disconnected from the Internet.

To achieve user authentication, commercially deployed VAs often perform voice recognition to distinguish speech patterns of different people. Such a solution can be ported to the obfuscator and have the VA only respond to authorized users. However, the adversary could potentially record the user's voice and play it back to control the VA. To address this shortcoming of biometric-based solutions, we achieve user authentication by utilizing the user's smartphone, which we also consider to be a trusted device. The user's smartphone contains a private key, that can be authenticated by the obfuscator via a secure OOB channel, such as Bluetooth Low Energy (BLE). The obfuscator is able to authenticate the user only when the user's smartphone is within its BLE's connection range.

4 OBFUSCATING THE VOICE ASSISTANT

As we have described earlier, the VA can record a user's private conversations and leak them to the cloud without the user's authorization or even awareness. One approach to prevent this leakage is to throttle the uplink traffic from the VA when the user is not commanding the device. The throttling can be lifted when the VA receives an authorized user's voice command, during which it is supposed to upload the speech for further processing. This approach, however, does not prevent the VA from caching the private speech, and uploading them at a reduced/throttled rate. A more effective direction would be to prevent the VA from recording any private speech. Although most VAs on the market offer an on device switch to turn the microphones on and off, it is highly impractical for the user to access the on-device switch every time she wishes to issue a voice command.

To prevent VAs' microphones from recording speech, with minimum user intervention, we use audio obfuscation techniques to jam the microphones. This is essentially a form of Denial-of-service (DoS) attack on the microphone to prevent innocuous recording. Typically, the obfuscation signal needs to be of the same frequency as the private speech that our system protects (which is the human

speech frequency range i.e 200Hz-16kHz). In such a case, the obfuscation signal would be audible to the human ear, rendering the solution unusable. We adopt the ultrasound obfuscation technique to *silently* jam the VA's microphone. We refer readers to BackDoor [33] and DolphinAttack [36] for the technical details of the technique, but provide a brief description of its working principle.

For a typical voice recording device, the sound received by the microphone passes through an amplifier, a low-pass filter, and then the ADC before being converted to a digital signal. These components generally have linear frequency responses in the audible frequency range (below 24kHz), but exhibit strong non-linearity above this frequency. As a result, the recored sound can be expressed in terms of the input S as,

$$S_{out} = A_1 S + A_2 S^2 + A_3 S^3 + \cdots \tag{1}$$

where A_i is the complex gain of the i^{th} order of input S. In practice, the power of the third and higher order terms are negligible. As a simple example, consider S with two inaudible tones at frequency ω_1 and ω_2. $S = S_1 + S_2 = \sin(\omega_1 t) + \sin(\omega_2 t)$. In the recorded sound, the multiplication of S_1 and S_2 resulted by the second order term creates frequency components $\omega_1 + \omega_2$ and $\omega_1 - \omega_2$. More formally, the second order term of the recored sound can be expressed as:

$$\begin{aligned} A_2(S_1 + S_2)^2 =& 1 - \frac{1}{2}\cos(2\omega_1 t) - \frac{1}{2}\cos(2\omega_2 t) \\ &+ \cos((\omega_1 - \omega_2)t) - \cos((\omega_1 + \omega_2)t) \end{aligned} \tag{2}$$

By carefully selecting ω_1 and ω_2 (*e.g.* $2\pi \cdot 40$kHz and $2\pi \cdot 50$kHz), the frequency component $\cos((\omega_1 - \omega_2)t)$ (10kHz) falls inside the human audible range, and be recorded by the microphone. By superimposing multiple ultrasound tones at different frequencies, the obfuscation covers the entire human speech frequency range at the VA's microphone. Due to the fundamental differences in the design of the human ear and microphone system hardware, the sound does not go through such non-linear transformation in the human ear. Thus the obfuscation signal is only audible to microphone.

In summary, the dedicated obfuscator equipped with ultrasound speaker plays carefully crafted obfuscation signals in a human-inaudible frequency. Superimposed by the obfuscation signal, the private user speech cannot be decoded by the smart speaker. Using this technique, we prevent the smart speaker from gathering the private information, thus eliminating the privacy threat.

This technique faces several challenges. First, the obfuscation signal also jams the intended voice commands. Second, a strong adversary may be able to utilize advanced signal processing techniques to separate the obfuscation signal from normal human speech. This is conceptually possible with the audio beamforming technique [16, 26], enabled by multiple microphones (Amazon Echo has a 7-mic array [8]). Finally, it is possible that there are multiple VAs deployed in the same environment, to which the user may wish to enforce different privacy configurations. For example, the user may wish to obfuscate inputs to her Echo speaker, but allow access to the smart TV. Since all the devices' microphones exhibit similar non-linearity, it would be challenging to apply fine-grained obfuscation, *i.e.*, only obfuscating certain devices, while leaving the other devices in the environment to function normally. The first problem is crucial in ensuring system usability, and we address this in §5. Addressing the remaining two problems is challenging, but would enable more advanced levels of privacy; we briefly discuss potential solutions in §7.

5 PRESERVING ACCESS

In this section, we explain how our framework permits legitimate voice commands to VAs under jamming. As the obfuscator jams both the private user speech and intended voice commands, our system needs to have the ability to distinguish these two types of sounds, and automatically lift jamming when appropriate. We achieve this by equipping the obfuscator with its own microphone. It constantly listens for keywords during the jamming, and lifts the obfuscation signal only when hotwords are detected.

One obvious problem is that the obfuscator's microphone might exhibit similar non-linearity as the VA. This means that it could be affected by the ultrasound obfuscation signal as well. To solve this problem, we leverage the working principle of BackDoor and DolphinAttack, and equip the obfuscator with an *ultrasound microphone*. Both BackDoor and DolphinAttack have shown that the microphone hardware has linear frequency response in the intended working frequency range (typically below 22kHz) [33, 36]. As most microphones on mobile devices and VAs only intend to record human generated sounds, the cut-off frequency of the employed low-pass filter is typically between 20kHz-22kHz. Strong non-linearity only exists beyond the cut-off frequency of low-pass filter. The input sound signal S only experience linear transformation $A \cdot S$ when S's frequency is below the filter cut-off frequency. The absence of higher order term S^2 prevents the sound from being converted to other frequencies in this case.

Suppose the obfuscation signal's frequency ranges from f_l to f_h, where f_l and f_h are above human audible frequency, we equip the obfuscator with ultrasound microphones that can record in the range of (f_l, f_h). To record at a higher frequency, ultrasound microphone's low-pass filter has a much higher cut-off frequency. For a microphone that can record up to f_h, the obfuscation signal's frequency range is still inside the linear region of the ultrasound microphone's frequency response. Therefore, the obfuscation signal received by ultrasound microphone does not create the shadow tones in low frequency, as in the case of the VA's microphone.

With the ability to hear human speech under ultrasound obfuscation signal, our obfuscator identifies a user's voice commands, and lift the jamming signal accordingly. To simplify the audio signal processing on the obfuscator, we have the obfuscator only detect the hotwords issued prior to the actual voice commands. Different from general speech recognition, detecting a small set of hotwords can be very efficient via comparing the recorded sound against hotwords' known speech pattern with cross-correlation or Dynamic time warping (DTW) [31].

6 LEGITIMATE USER AUTHENTICATION

So far, we have addressed the problem of innocuous recording of the user's private speech. Recall that the obfuscator jams the VA continuously and lifts the jamming when a user issues a voice command. However, adversaries present in the smart home ecosystem could do the same *i.e.* issue a voice command to lift the jamming signal and perform unauthorized actions. Thus, it is essential to

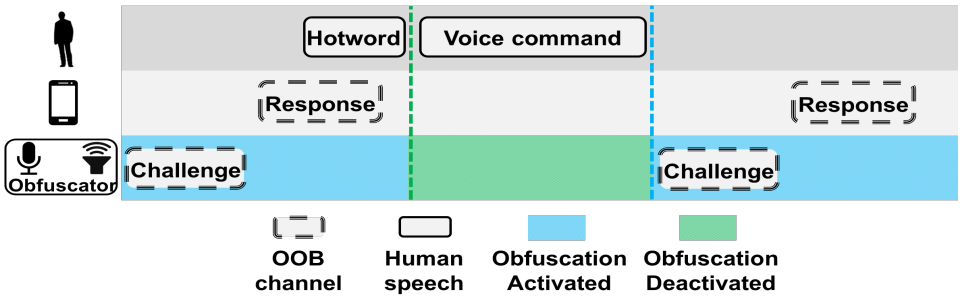

Figure 2: User's smartphone and audio obfuscator constantly exchange challenge response messages over secure OOB channel for authentication. Obfuscator lifts jamming signal after detecting hotwords in user's speech.

ensure that users are authenticated before they are allowed to use the VA.

Conventionally, authentication between the user and a device happens using the knowledge of either a shared secret (such as a password –something you know) or through some form of trusted hardware token (–something you have). However, using either of the aforementioned schemes adds additional usability burden in the VA ecosystem; the legitimate user should not have to do anything she normally would not do while using the VA. One solution is to use a unique feature of a user as a signature. For example, facial recognition can be used to identify the legitimate user of the system. This would require retrofitting the obfuscator with a camera and sufficient compute capabilities; the cost to purchase (and use) such a device will be prohibitively high. Alternatively, if one were to use an existing, external IoT camera to perform the classification, then (i) the external device would have to be trusted *i.e.* not record the users activities outside its designated period of usage, and (ii) the user would have to be in the line-of-sight of the device while issuing commands to the VA. Both of these requirements are too strict. A more fundamental problem with biometric solutions is their lack of resilience, *i.e.* once compromised, these solutions are no longer useful. For example, if a user's fingerprint is recorded by an adversary, then the user can no longer be uniquely authenticated.

Thus, we observe that (i) it is essential that the authentication mechanism be computationally inexpensive, so as to be easily performed on the obfuscator, and (ii) the authentication mechanism is resilient. To that end, we employ a traditional challenge-response mechanism for legitimate user authentication. We assume that each legitimate user creates a public-private key pair and stores it on a hardware device[2]. The user then proceeds to communicate the public key to the obfuscator (in an initial and secure registration phase). The challenge-response mechanism by itself is very simple. The obfuscator issues a random nonce (*challenge*) to each user, and the user returns the cryptographically signed nonce (*response*) back to the device. The nonce is signed with the user's private key, and can be verified for correctness using the public key listed on the device. Assumptions made in §2 prohibit the adversary from injecting it's public-key at the time of configuration (we discuss more stringent checks in §7). Thus, the obfuscator is now able to distinguish between authorized users and others.

[2]We address this limitation later in the section.

If a hotword is used in conjunction with successful authentication, the obfuscator proceeds to stop its jamming, and communicates the hotword to the VA (refer Figure 1) via ultrasound injection. Subsequent voice commands are relayed directly to the VA. To minimize user involvement, the challenge-response mechanism continuously occurs between the user's device and the obfuscator. However, one can envision a scenario where the user is far away from the obfuscator, and the challenge-response mechanism is successful. This provides any unauthorized user in the vicinity a window to inject a voice command. To prevent such scenarios, we propose that the range of transmission of the challenge from the obfuscator be restricted to several feet. This ensures that unauthorized users are proximate to the obfuscator (and consequently the legitimate user), and any authenticated activity she carries out will be detected, and can be dealt with reactively.

As we alluded to earlier in the section, the private key generated by the user is stored in a hardware device. Recall that one of the requirements of our framework is to ensure that the user need not purchase any additional hardware (apart from the obfuscator). To that end, we believe that a legitimate user's smartphone is the best option to both house the secret key, and perform the challenge-response protocol passively as a background task. Smartphones are pervasively used, with 32.3% of the world's population using a smartphone in 2017. This number is projected to increase to 40% by 2021 [2]. Additionally, they are computationally equipped to perform the operations required for computing digital signatures.

To restrict the transmission range, we recommend the usage of the Bluetooth low energy (BLE) standard on both the smartphone and the obfuscator. BLE has various transmission modes that restrict the transmission range. BLE also overcomes passive eavesdropping by encrypting the data being transferred using AES-CCM cryptography [15]. However, how the keys are (initially) exchanged has a great effect on the security of the connection. Various solutions such as Just Works, OOB pairing, Passkey, and Numeric Comparison can be used to alleviate this problem [6].

Fig. 2 shows above mentioned interactions between user's smartphone and the audio obfuscator that enables continuous user authentication.

7 DISCUSSION

Despite the components described in §4, §5, and §6, realizing such a system still faces several challenges.

Pending challenges in the obfuscation domain: If the VA is aware of the obfuscator's jamming, it could potentially utilize multiple microphones to perform beamforming to suppress the jamming signal. It is worth noting that many commercial VAs are already being equipped with a number of microphones to perform beamforming for the purpose of improving human speech recording quality in noisy environments [8]. Beamforming enables a VA to form a highly directional receiving beam to only receive the sound coming from a specific direction, while filtering out the rest of the sound. In today's VAs, sounds from different microphones are combined with different weights to form a beam that maximize the signal strength of the user's speech, thus suppressing the noise. A similar principle can be applied to suppress obfuscation signal. When the user and obfuscator are in different positions, their corresponding sounds have different arriving direction at the smart speaker, which makes them separable by beamforming. Although the smart speaker may not be fully aware of the directions of these two sound sources, there exists mature techniques such as blind source separation to isolate different signal sources effectively [19]. One possible solution to this problem is to equip the obfuscator with multiple speakers to perform transmitting beamforming. This allows the obfuscator to create directional transmitting beams towards different directions, which could potentially be superimposed with the user voice's direction at the smart speaker. When these two sounds are coming from the same direction, it becomes extremely difficult for the VA to separate them using beamforming. A significantly challenge is the obfuscator being unaware of the position of the user, and is unable to predict the transmitting direction best suited to overlap with the user voice's direction.

Ideally, we would hope to enable fine-grained privacy configuration and control, where the user can decide to jam certain VAs while leaving the others functional in the same environment. This is challenging as all sound recording devices exhibit similar properties, and all would be affected by the obfuscation signal. To obfuscate only certain devices, we could take advantage of slight differences across different audio hardware. Although all microphones are fundamentally affected by the above non-linearity we described, the differences in their hardware components make them more sensitive to jamming signals at different ultrasound frequencies. By carefully examining such hardware dependencies (which is a one-time process that can be conducted in lab environment for different VAs), it is possible to design ultrasound obfuscation signal that jams certain devices more effectively than others. Consequently, it becomes possible to utilize such obfuscation signals to only affect certain intended devices, while others could still be able to record.

Pending challenges in the authentication domain: Our discussion in §6 assumes that the unauthorized user is in close proximity to the legitimate user while attempting to inject commands to the obfuscator. This is the only time period when the unauthorized user has access to the unjammed VA. However, recent work by Roy et al. [34] provides a mechanism for such an unauthorized user to inject voice commands from a long range. In such a scenario, one possible outcome that might occur is that the legitimate user's command and the unauthorized user's command are both multiplexed and the device does not understand what it must do. In this case, the legitimate user is notified. Alternatively, if the unauthorized user's command is the only one processed, then based on the execution outcome, the user knows that someone is propagating signals from range and is notified. In both these aforementioned scenarios, the user is notified of unauthorized activities based on feedback from the VA. Observe that the inaudible command propagated by the unauthorized user is recognizable by both the VA and the obfuscator. If the VA does not provide vocal/visible feedback, one can envision a solution where the obfuscator processes and forwards the inaudible command to the user, through the same BLE channel used for the challenge response mechanism. While all these solutions are reactive, discovering the actions of such an adversary in a proactive fashion is future work.

8 RELATED WORK

Smart home device privacy threats. Smart devices have been shown to be able to collect a wide range of user information, including sleeping patterns [5], exercise routines [11], child behavior [9], and medical information [10]. Their always-on sensors pose severe privacy concerns [17], as well as security threats [28, 32]. Moorthy *et al.* investigate the privacy concerns of using VAs in public areas [23]. Diao *et al.* demonstrate that it is possible to hijack an Android smartphone's speaker to play malicious voice commands and control the VA [22]. Recent research shows that it's possible to inject mangled voice commands such that they are unrecognizable to humans, but can be decoded by VAs [20, 35].

Audio obfuscation techniques. Audio obfuscators and sound maskers have been used to protect private conversations from being eavesdropped by unauthorized voice recording devices. Most previous studies obfuscate in human-audible frequency range [14, 25, 30]. Recently, BackDoor [33] and DolphinAttack [36] utilize the non-linearity in microphone hardware to inject human-audible frequency sound by playing specially crafted ultrasound. As follow up, LipRead [34] extends the injection range, and also proposes a defense. The defense identifies whether the recorded sound is from actual human speaking or played by an ultrasound speaker, to reject fake voice commands. However, it does *not* prevent the microphone from recording the inaudible sound at audible frequency, and thus does not prevent the obfuscation.

Continuous user authentication. For VAs, there have been a number of proposals on using voice as a biometric for authentication [18, 21, 29]. Voice recognition is vulnerable to simple replay attacks if the adversary records user's voice in advance. VAuth takes a different approach and authenticates the voice commands with wearable devices [24]. It senses the vibrations during the speech and matches the motion information with the speech for authentication. The trade-off is that VAuth requires user to wear devices equipped with motion sensors.

9 ACKNOWLEDGEMENTS

This work is supported in part by the Wisconsin Alumni Research Foundation and the US National Science Foundation through awards CNS-1345293, CNS-14055667, CNS-1525586, CNS-1555426, CNS-1629833, CNS-1647152 and CNS-1719336.

REFERENCES

[1] 2015. Samsung's warning: Our Smart TVs record your living room chatter. https://www.cnet.com/news/samsungs-warning-our-smart-tvs-record-your-living-room-chatter/. (2015).

[2] 2017. Smartphone user penetration as percentage of total global population from 2014 to 2021. https://www.statista.com/statistics/203734/global-smartphone-penetration-per-capita-since-2005/. (2017).

[3] 2017. Stuffed toys leak millions of voice recordings from kids and parents. https://money.cnn.com/2017/02/27/technology/cloudpets-data-leak-voices-photos/index.html. (2017).

[4] 2018. Amazon Echo. https://www.amazon.com/Amazon-Echo-And-Alexa-Devices/b?ie=UTF8&node=9818047011. (2018).

[5] 2018. BEDDIT 3 SLEEP MONITOR. https://www.beddit.com/. (2018).

[6] 2018. BLE Features. http://www.cypress.com/file/224826/download. (2018).

[7] 2018. Google Home. https://store.google.com/us/product/google$_{h}$ome. (2018).

[8] 2018. Google Home Max vs. HomePod and Google Home Mini vs. Amazon Echo Dot: battle of the smart speakers. https://www.theverge.com/2017/10/5/16425142/google-home-mini-vs-amazon-echo-dot-max-apple-homepod. (2018).

[9] 2018. Internet of Things Teddy Bear Leaked 2 Million Parent and Kids Message Recordings. https://motherboard.vice.com/en$_{u}$s/article/internet-of-things-teddy-bear-leaked-2-million-parent-and-kids-message-recordings. (2018).

[10] 2018. Nest and the Internet of Broken Promises (and how to fix it). https://medium.com/minimum-reliableproduct/nest-and-the-internet-of-broken-promises-and-howto-fix-it-a26a1fc3349c#.ugh23el9o. (2018).

[11] 2018. Running on data: Activity trackers and the Internet of Things. https://www2.deloitte.com/insights/us/en/deloitte-review/issue-16/internet-of-things-wearable-technology.html. (2018).

[12] 2018. Samsung Smart TV. http://www.samsung.com/ph/smarttv/voice$_{c}$ontrol.html. (2018).

[13] 2018. Security Tokens. https://en.wikipedia.org/wiki/Security$_{t}$oken. (2018).

[14] 2018. Speech Privacy Systems. https://www.speechprivacysystems.com. (2018).

[15] 2018. Using AES-CCM and AES-GCM Authenticated Encryption in the Cryptographic Message Syntax (CMS). https://tools.ietf.org/html/rfc5084. (2018).

[16] Xavier Anguera, Chuck Wooters, and Javier Hernando. 2007. Acoustic beamforming for speaker diarization of meetings. *IEEE Transactions on Audio, Speech, and Language Processing* 15, 7 (2007), 2011–2022.

[17] Noah Apthorpe, Dillon Reisman, Srikanth Sundaresan, Arvind Narayanan, and Nick Feamster. 2017. Spying on the smart home: Privacy attacks and defenses on encrypted iot traffic. *arXiv preprint arXiv:1708.05044* (2017).

[18] Mossab Baloul, Estelle Cherrier, and Christophe Rosenberger. 2012. Challenge-based speaker recognition for mobile authentication. In *Biometrics Special Interest Group (BIOSIG), 2012 BIOSIG-Proceedings of the International Conference of the*. IEEE, 1–7.

[19] Adel Belouchrani, Karim Abed-Meraim, J-F Cardoso, and Eric Moulines. 1997. A blind source separation technique using second-order statistics. *IEEE Transactions on signal processing* 45, 2 (1997), 434–444.

[20] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. 2016. Hidden Voice Commands.. In *USENIX Security Symposium*. 513–530.

[21] Amitava Das, Ohil K Manyam, Makarand Tapaswi, and Veeresh Taranalli. 2008. Multilingual spoken-password based user authentication in emerging economies using cellular phone networks. In *Spoken Language Technology Workshop, 2008. SLT 2008. IEEE*. IEEE, 5–8.

[22] Wenrui Diao, Xiangyu Liu, Zhe Zhou, and Kehuan Zhang. 2014. Your voice assistant is mine: How to abuse speakers to steal information and control your phone. In *Proceedings of the 4th ACM Workshop on Security and Privacy in Smartphones & Mobile Devices*. ACM, 63–74.

[23] Aarthi Easwara Moorthy and Kim-Phuong L Vu. 2015. Privacy concerns for use of voice activated personal assistant in the public space. *International Journal of Human-Computer Interaction* 31, 4 (2015), 307–335.

[24] Huan Feng, Kassem Fawaz, and Kang G Shin. 2017. Continuous authentication for voice assistants. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*. ACM, 343–355.

[25] Rafik Goubran and Radamis Botros. 2003. Adaptive sound masking system and method. (June 5 2003). US Patent App. 09/998,191.

[26] Malcolm Hawkes and Arye Nehorai. 1998. Acoustic vector-sensor beamforming and Capon direction estimation. *IEEE Transactions on Signal Processing* 46, 9 (1998), 2291–2304.

[27] Keith Johnson, David B Pisoni, and Robert H Bernacki. 1990. Do voice recordings reveal whether a person is intoxicated? A case study. *Phonetica* 47, 3-4 (1990), 215–237.

[28] Chaouki Kasmi and Jose Lopes Esteves. 2015. IEMI threats for information security: Remote command injection on modern smartphones. *IEEE Transactions on Electromagnetic Compatibility* 57, 6 (2015), 1752–1755.

[29] Max Kunz, Klaus Kasper, Herbert Reininger, Manuel Möbius, and Jonathan Ohms. 2011. Continuous Speaker Verification in Realtime.. In *BIOSIG*. Citeseer, 79–88.

[30] Arnold M McCalmont. 1980. Voice privacy system with amplitude masking. (March 25 1980). US Patent 4,195,202.

[31] Meinard Müller. 2007. Dynamic time warping. *Information retrieval for music and motion* (2007), 69–84.

[32] Giuseppe Petracca, Yuqiong Sun, Trent Jaeger, and Ahmad Atamli. 2015. Audroid: Preventing attacks on audio channels in mobile devices. In *Proceedings of the 31st Annual Computer Security Applications Conference*. ACM, 181–190.

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

[34] Nirupam Roy, Sheng Shen, Haitham Hassanieh, and Romit Roy Choudhury. 2018. Inaudible Voice Commands: The Long-Range Attack and Defense. In *15th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 18)*. 547–560.

[35] Tavish Vaidya, Yuankai Zhang, Micah Sherr, and Clay Shields. 2015. Cocaine noodles: exploiting the gap between human and machine speech recognition. *WOOT* 15 (2015), 10–11.

[36] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. 2017. DolphinAttack: Inaudible voice commands. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 103–117.