2022 IEEE 42nd International Conference on Distributed Computing Systems (ICDCS) | 978-1-6654-7177-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICDCS54860.2022.00071

Defending against Thru-barrier Stealthy Voice Attacks via Cross-Domain Sensing on Phoneme Sounds

Cong Shi*, Tianming Zhao†, Wenjin Zhang*, Ahmed Tanvir Mahdad‡, Zhengkun Ye†,
Yan Wang†, Nitesh Saxena‡, Yingying Chen*
*Rutgers University, New Brunswick, NJ 08901, USA, †Temple University, Philadelphia, PA 19122, USA
‡Texas A&M University, TX 77843, USA
Email: *{cs1421, wz315, yingche}@scarletmail.rutgers.edu, †{tum94362, zhengkun.ye, y.wang}@temple.edu,
‡{mahdad, nsaxena}@tamu.edu

***Abstract*—The open nature of voice input makes voice assistant (VA) systems vulnerable to various acoustic attacks (e.g., replay and voice synthesis attacks). A simple yet effective way for adversaries to launch these attacks is to hide behind barriers (e.g., a wall, a window, or a door) and give unauthorized voice commands without being observed by legitimate users. In this work, we develop an automated, training-free defense system that can protect VA systems from such thru-barrier acoustic attacks. Our study finds that acoustic signals passing through the barriers generally present a unique frequency-selective effect in the vibration domain. Thus, we propose to devise a system to capture this unique effect of barriers by leveraging low-cost, cross-domain sensing available in users' wearables. The system replays the audio-domain signals with the wearable's speaker and captures the conductive vibrations caused by the audio sounds in the vibration domain via the built-in accelerometer. To improve the proposed system's reliability, we develop a unique vibration-domain enhancement method to extract the phonemes most sensitive to the frequency-selective effect of barriers. We identify effective vibration-domain features that capture the barriers' effects in the vibration domain. A 2D-correlation-based method is developed to examine the speech similarity between the recordings from the VA system and the user's wearable and detect thru-barrier attacks. Extensive experiments with various barriers and environments demonstrate that the proposed defense system can effectively defend random, replay, synthesis, and hidden voice attacks with less than 4% equal error rates.**

I. Introduction

In recent years, the adoption of voice assistant (VA) systems has become a rising trend as they provide a convenient interface for users to interact with their smart/IoT devices (e.g., smart speakers, home robots, computers, smartphones). With such widespread usage of VA systems, their inherent security risks have drawn increasing public attention. The open nature of voice access provides great convenience, but it also renders the VA systems susceptible to a variety of acoustic attacks, such as replay attacks [1], voice synthesis attacks [2], and even hidden voice attacks [3] that modulate voice commands into noise-like and unintelligent attack sounds. A particularly stealthy way for adversaries to launch these attacks is to hide behind room barriers (e.g., a window or a door) and issue unauthorized voice commands without being observed, as illustrated in Figure 1. Such thru-barrier attacks are easy to perform in practical environments (without requiring physical access to the VA devices), and do not require any modifications to the hardware or software of the victim VA systems. The attacks are also stealthy as the adversary remains invisible to the user, or they can be launched when the user is not present, causing severe security issues (e.g., disarming the smart locking system of a house to gain access) without being noticed by users. Therefore, adversaries may be generally well-position at launching thru-barrier acoustic attacks in practice, which becomes a common security problem for VA systems. In this work, we show that such thru-barrier attacks can succeed with a high probability, which motivates us to develop a fundamental defense system to against this attack.

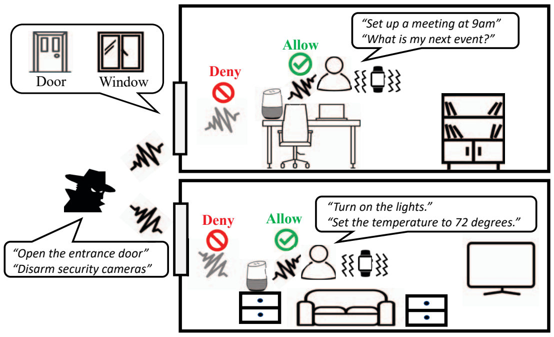


Fig. 1. Illustration of the application scenarios of the proposed defense system against thru-barrier attacks: an adversary is trying to hide behind a window or a door and give unauthorized voice commands to attack a VA system in a room. A user is using the proposed training-free defense system to detect the attack using cross-domain sensing on the user's wearable device.

Existing defense solutions for VA systems mainly focus on discriminating live speeches and replayed sounds. These approaches exploit machine learning models and various spectral features extracted from audio signals for attack detection, such as Mel Frequency Cepstral Coefficient [4], Constant Q Cepstral Coefficients [5], and Rectangular Frequency Cepstral Coefficient [4]. These works do not consider the impacts of barriers that significantly distort the voice sound. Thus, they

2575-8411/22/$31.00 ©2022 IEEE
DOI 10.1109/ICDCS54860.2022.00071

may not work well with the thru-barrier attacks. Although current VA devices are equipped with voice authentication functionalities [6], many people usually do not activate these functionalities at homes and offices. Such a practice renders VA systems vulnerable to thru-barrier acoustic attacks. Recently, second-factor authentication (e.g., via challenge questions [3], replay messages/calls [7]) has been proposed to add another layer of defense. However, it require users to confirm every voice commands, which incur great burdens and could be prone to users' careless behaviors [8].

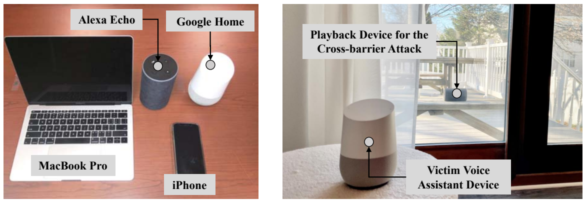
In this project, we conduct extensive experiments to explore the sound properties of barriers made out of different materials. We found that when a voice sound passes through a barrier, the voice sound intensity diminishes due to the attenuation of the barrier. *Our studies find that the attenuation effects are more significant in the high-frequency ranges (e.g., over 500Hz) for common barrier materials (e.g., glasses, wood), and the voice sound through the barrier is usually dominated by low-frequency components (e.g., 85Hz∼500Hz).* We refer to such frequency-selective attenuation of barriers as *barrier effect*. In contrast, a legitimate user's voice commands not passing through the barrier usually include high- and low-frequency components. Therefore, one approach to capture the barrier effect and detect thru-barrier attacks is to examine the high-frequency spectral energy of the voice sounds captured by the VA device. *However, we find that this approach is not reliable as some voice sounds inherently have low spectral energy in high-frequency ranges, leading to false detection of thru-barrier attacks.* Therefore, a reliable method to capture the barrier effect in voice commands and detect thru-barrier attacks is urgently needed for VA systems.

Toward this end, we design an automated system that can effectively detect thru-barrier attacks targeting VA systems. Unlike existing works, our system compares the voice commands recorded by the VA device and the user's wearable device in the vibration domain to capture the unique barrier effect and detect thru-barrier attacks. In particular, after being triggered by a wake word (e.g., "Hey, Siri" or "Alexa"), the proposed system records a voice command using the VA device and the wearable device, respectively. The recorded voice commands are aggregated at the wearable device, which replays them using the built-in speaker and captures the corresponding conductive vibrations via its accelerometer. We exploit such a replay process to ensure voice commands result in strong vibration signals, thereby effectively converting audio-domain signals into the vibration domain. The system adopts a correlation-based method to examine the non-linear frequency selectivity and detect the existence of thru-barrier attacks. The insight is that our cross-domain approach converts voice commands from the audio domain to the vibration domain, where the unique barrier effect becomes more significant. By examining the existence of the barrier effect, our system can effectively identify thru-barrier attacks without requiring any training efforts. The proposed system can serve as an additional layer on top of the existing voice authentication systems to enhance VA systems' security.

Realizing such a training-free thru-barrier attack detection system faces several challenges in practice. First, the barrier effect is minute and difficult to observe in the audio domain. To ensure robust detection of thru-barrier attacks in practice, we need to explore effective solutions to detect such minute incidents in the adversary's voice commands. Second, the accelerometers in wearable devices usually have low sampling rates (e.g., up to 200Hz), which results in aliased signals in the vibration signals converted from voice sounds. Such effects make it challenging to distinguish the adversary's and the legitimate user's vibration signals. Third, we find that not all phonemes inherently have all-spectrum energy to reveal the barrier effect and trigger wearables' accelerometers. We need to identify the phonemes to enable effective thru-barrier attack detection using cross-domain sensing.

To address the above challenges of detecting thru-barrier attacks, we develop a cross-domain sensing scheme that examines audio domain signals in the vibration domain. Our scheme can enhance the frequency-selective barrier effect in the adversary's voice commands and enable thru-barrier attack detection in practical environments. In addition, we develop a correlation-based detector to enable training-free thru-barrier attack detection by leveraging the fact that the adversary's voice sounds are noisier in the vibration domain because accelerometers generate more random noises when converting low-frequency signals [9]. Such an effect renders the vibration signals of the adversary have relatively lower correlations compared to those of the legitimate user, so as to enable detecting the adversary while addressing the ambiguity induced by signal aliasing. Furthermore, we propose a vibration domain enhancement method to enable effective and efficient thru-barrier attack detection based on the effective phonemes that are most sensitive to thru-barrier attacks. We summarize our contributions as follows:

- This is the first work to study the effects of thru-barrier voice attacks when an adversary stealthily launches an attack at a distance (e.g., outside a window or a room). Our study finds that the cross-domain vibrations of a wearable accelerometer can reveal unique frequency selectivity patterns of different kinds of barriers, facilitating the detection of thru-barrier stealthy voice attacks.
- We develop a training-free defense system to protect VA systems against various thru-barrier voice attacks, including both clear voice attacks (e.g., random, voice synthesis, and replay attacks) and hidden voice attacks. The defending system can serve as an additional layer on top of the existing voice authentication schemes to provide enhanced security.
- Our unique defense techniques are built upon the identification of phoneme sounds in speech. The wearable's built-in speaker and accelerometer are used to perform cross-domain sensing, and the most prominent phonemes are selected for efficient sound playback and effective attack defense.
- Extensive experiments are conducted with different kinds of wearable devices, room barriers, and sound volumes in four different room environments over a period of five months.


Alexa Echo
Google Home
MacBook Pro
iPhone
Playback Device for the Cross-barrier Attack
Victim Voice Assistant Device

(a) Voice assistant devices (b) Setup to study the thru-barrier attack

Fig. 2. Devices and experiment setup to validate the vulnerability of VA devices to thru-barrier acoustic attacks.

The results show that our system can effectively defend random attacks, replay attacks, voice synthesis attacks, and hidden voice attacks with less than 4% equal error rates.

II. Threat Model

This paper considers thru-barrier attacks where an adversary wants to issue unauthorized voice commands to a VA device behind a barrier (e.g., a wall, a window, or a door). We mainly consider the scenarios that a legitimate user (victim) of the VA device wears a wearable device and is inside the room when the attack is launched. Our defense system rejects voice commands at the VA if the wearable device is absent. The wearable device (e.g., a smartwatch) is equipped with accelerometers and connected to the VA system through the cloud service. Due to the barrier's blockage effect, the adversary's voice commands are significantly attenuated, making users hard to perceive the attacks. Yet, the VA system can still detect and recognize the adversary's voice commands. To avoid being noticed from the audible reply by the user, the adversary can first lower the volume of the VA device using voice commands. We consider the following three approaches that the adversary may use to launch the thru-barrier attacks:

Random Attack. When the prior knowledge of the user's voice is not available, the adversary can try to use his/her own voice to launch the attacks. Such simple attacks have considerable probabilities to bypass state-of-the-art defensive schemes (e.g., over 11% EER on detecting spoofing attacks [10]).

Replay Attack. The adversary can replay the victim's voice samples using playback devices (e.g., loudspeakers) to spoof VA systems [1]. The voice samples can be obtained from the victim's public speech collected through the Internet.

Voice Synthesis Attack. The adversary can also leverage voice synthesis models [2] to convert text into any target speeches of a victim based on either a pre-trained model or a small number of the victim's voice samples for training.

Hidden Voice Attack. The adversary can use obfuscated voice commands as attack sounds [3] to further increase its imperceptibility. The obfuscated voice commands are usually generated by converting clear voice commands into sound signals meaningful to VA systems' embedded speech recognition models but incomprehensible to human beings.

III. Attack Study

A. *Acoustic Attacks Across Room Barriers*

To understand the severity of the thru-barrier attacks, we conduct experiments by attacking four different commercial VA devices in an apartment, including two smart speakers (i.e., Amazon Echo, Google Home), a laptop (i.e., MacBook Pro), and a smartphone (i.e., iPhone). In the experiments, we use a loudspeaker (Razer Sound Bar RC30) to replay wake words (i.e., "OK Google", "Alexa", "Hey Siri") to activate the VA devices placed behind two barriers, a glass window and a wooden door. The loudspeaker is placed near the window (i.e., 10cm) to replay the wake words with normal conversation sound pressure levels (i.e., 65dB and 75dB). The VA devices are 2m away from the barrier. For the voice synthesis attack, we use a pre-trained voice synthesis model [11] to generate the attack sounds. We perform the attack 10 times for each VA device and calculate the number of attack attempts that trigger the VA device. The results are summarized in Table I.

For replay attacks, we find that the thru-barrier attacks can trigger the smart speakers and the laptop with moderate/high success rates for the two barriers, while the smartphone has relatively lower success rates. This is because smart speakers and laptops use high-sensitive microphone arrays (or far-field microphones) to cover large areas, making them more susceptible to acoustic attacks. Random attacks and voice synthesis attacks have similar success rates in attacking Google Home and Alexa Echo compared to random attacks. Table I does not have random and synthesis attack results for MacBook Pro and iPhone because Siri has an embedded voice recognition mechanism, and they did not respond to the voices they cannot recognize. In addition, we evaluate the hidden voice attack on Google home by replaying a publicly available command "OK Google" [3]. We find that with a sound volume of 65dB, 5 out of 10 and 10 out of 10 attack attempts succeed for the two barriers, respectively. All the attack attempts succeed if the sound volume increases to 75dB. We did not test the hidden voice attack on the other devices as the corresponding commands are not available. Note that for all the attack approaches, the adversary can achieve a considerable increase in the success probability if he/she repeat the attack (e.g., twice). In general, the results confirm the security issues of VA devices under thru-barrier attacks.

B. *Frequency-selective Barriers Effect*

When a sound wave travels through a medium, its intensity diminishes with the distance that results from acoustic attenuation [12]. The amplitude change of an attenuating sound wave can be expressed as:

$$P(x+\Delta d) = P(x)e^{-\alpha(f,\eta)\Delta d}, \tag{1}$$

where $P(x)$ is the unattenuated sound pressure at initial location x, $P(x+\Delta d)$ is the pressure after the sound wave has traveled a distance Δd from the initial location (i.e., the thickness of the medium). The parameter $\alpha(f,\eta)$ is the frequency-material-dependent sound attenuation/absorption coefficient.

Notably, α varies with barriers' materials and sound frequency [13], and the larger the coefficient the easier for sounds to penetrate the barrier. For glass windows and wooden doors, the coefficients corresponding to higher frequencies (e.g., 0.02 for the glass window and 0.04 for the wooden door) are

TABLE I
Study on thru-barrier attack against representative voice assistant devices.

Device Name	Command	Attack success out of 10 attempts (65dB; 75dB)					
		Glass window			Wooden door		
		Random Attack	Replay Attack	Voice Synthesis Attack	Random Attack	Replay Attack	Voice Synthesis Attack
Google Home	OK Google	9/10; 10/10	10/10; 10/10	4/10; 10/10	10/10; 10/10	10/10; 10/10	8/10; 10/10
Alexa Echo	Alexa	5/10; 10/10	4/10; 10/10	3/10; 10/10	9/10; 10/10	10/10; 10/10	3/10; 10/10
MacBook Pro	Hey Siri	-	4/10; 10/10	-	-	4/10; 10/10	-
iPhone	Hey Siri	-	0/10; 6/10	-	-	0/10; 7/10	-

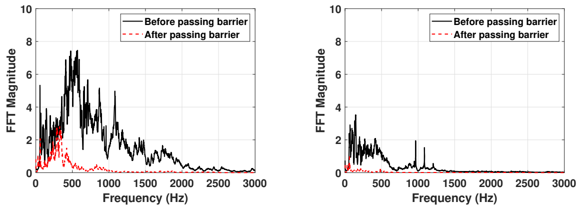
smaller than coefficients corresponding to lower frequencies (e.g., 0.10 for the glass window and 0.14 for the wooden door) indicating more absorptions for higher frequencies than lower frequencies. We refer to this phenomenon as the barrier effect. Therefore, when an adversary launches a thru-barrier attack, the attack sound passing through the barrier retains more energy in the lower-frequency band. We also notice that brick walls have low coefficients for all frequencies, indicating that brick walls can generally absorb sound, making it hard to launch the thru-barrier attack. As such, we focus on studying attacks through glass windows and wooden doors, where the adversary likely launches successful attacks.

We conduct an experiment to demonstrate the barrier effect of a glass window. Specifically, we use a loudspeaker outside a glass window to play a list of 63 phoneme sounds from the TIMIT dataset [14]. Two microphones are used to record the phoneme sounds before and after passing the glass window. Figure 2 (b) illustrates the setup. We randomly choose 100 sound segments from five males and five females for each phoneme and play them with a sound pressure level of 75dB. We apply Fast Fourier Transformation (FFT) to all the recorded sound samples and calculate the average of each phoneme's FFT magnitude. Figure 3 shows the comparison of the average FFT magnitude of a vowel phoneme sound (i.e., /ae/) and a consonant phoneme sound (i.e., /v/) before and after they pass through the window. We can see that high-frequency components (over $500Hz$) of both phoneme sounds are attenuated significantly by the glass window, which is consistent with our acoustic model. We also find that the vowel sound passing the barrier has a similar power spectrum to the consonant sound that did not pass the barrier. This suggests that it is not reliable to use the frequency-selective barrier effect in the audio domain to detect thru-barrier attacks.

IV. Defense System Design

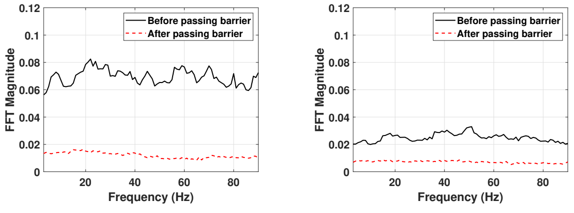
A. Cross-domain Sensing

Wearable devices (e.g., wrist bands, smartwatches) are usually equipped with an accelerometer to measure the device's movements. Recent studies [15] have shown that the accelerometer can pick up physical vibrations caused by speeches, which is referred to as cross-domain sensing. In this work, we find that such cross-domain sensing can reveal the unique barrier effect in the vibration domain more clearly than in the audio domain. The insights are the accelerometer can significantly attenuate low-frequency audio signals (e.g., below 500Hz) that co-exist in both the adversary's voices (i.e., passing a barrier) and the user' (i.e., without passing a barrier). Meanwhile, it captures the high-frequency audio signals (e.g., over 1000Hz) that only exist in the user's voices. Such characteristics make the adversary's voices, and the user's more distinguishable in the vibration domain. In addition, researchers have found that the amplifier of the accelerometer injects random noises when mapping low-frequency voice signals to the vibration domain [9]. We find that this effect helps us detect attack sounds through barriers, which are dominated by low-frequency components.



(a) /ae/ (b) /v/

Fig. 3. Comparison between phoneme sounds before and after passing the barrier in the audio domain.



(a) /ae/ (b) /v/

Fig. 4. Comparison between phoneme sounds before and after passing the barrier in the vibration domain.

Figure 4 presents the average FFT magnitudes of the vibrations caused by two phonemes, /ae/ and /v/, captured by the accelerometer in a commercial smartwatch (i.e., Fossil Gen 5). Different from the FFT magnitudes of the sounds of these two phonemes as shown in Figure 3, we find that the vowel sound /ae/ passing the barrier and the consonant sound /v/ not passing the barrier are distinguishable in the vibration domain. The results demonstrate that cross-domain sensing technology can facilitate detecting thru-barrier attacks.

B. Challenges

Weak Barrier Effect in the Audio Domain. While we find the unique barrier effect in thru-barrier attacks, this effect is weak in the audio domain. More specifically, the adversary's voices may have a similar spectral energy distribution as the user's in practice, although they are dominated by low-frequency components. As such, we propose to leverage cross-domain sensing to capture the barrier effect and detect thru-barrier attacks in the vibration domain.

Ambiguous Signal Conversion in Cross-domain Sensing. Commercial wearable devices usually have lower sampling rates (e.g., up to 200Hz), which causes aliasing effects where multiple high-frequency audio signals (e.g., higher than 200Hz) are mapped into the same low-frequency signals (e.g., less than 200Hz) in the vibration domain. Such aliasing effects make it challenging to discriminate the adversary's voices and the user's using cross-domain sensing.

Phoneme-dependent Ineffective Vibration Generation. Some consonant phonemes inherently have weak sound intensity due to lacking larynx vibration during sound production. Such phonemes cannot generate strong enough vibrations to trigger the accelerometer. Thus, they are not suitable for detecting the unique frequency-selective effects in the vibration domain. It is necessary to identify the phonemes sensitive to the barrier effect to facilitate thru-barrier attacks detection.

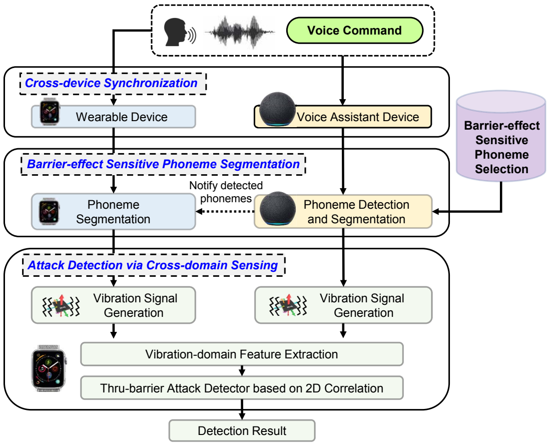

Voice Command
Cross-device Synchronization
Wearable Device
Voice Assistant Device
Barrier-effect Sensitive Phoneme Selection
Barrier-effect Sensitive Phoneme Segmentation
Notify detected phonemes
Phoneme Segmentation
Phoneme Detection and Segmentation
Attack Detection via Cross-domain Sensing
Vibration Signal Generation
Vibration Signal Generation
Vibration-domain Feature Extraction
Thru-barrier Attack Detector based on 2D Correlation
Detection Result

Fig. 5. System architecture.

C. System Overview

We develop a training-free defense system to detect thru-barrier attacks, which compares the voice commands recorded by the VA and the wearable devices in the vibration domain to examine the effects of barriers. The architecture of our system is illustrated in Figure 5. Our system first performs the *Cross-device Synchronization* to ensure the VA and the wearable devices record the same voice command. We assume that both the VA and wearable devices are connected to the same local WiFi network, where network communication delay is usually low and suitable for synchronization. Upon detecting a wake word on the VA device, our system notifies the wearable device to record the same voice command simultaneously. Then, the system utilizes an *Barrier-effect Sensitive Phoneme Segmentation* scheme based on deep learning to obtain the sound segments associated with the effective phonemes. The phoneme segmentation mechanism can reuse immediate results of the speech recognition pipeline (e.g., preprocessed audio, spectrum features, and hidden representations) on the VA system to reduce computational cost. Based on the segmentation information, our system separately extracts and concatenates all the sound segments in the voice commands of the wearable and VA devices for cross-domain sensing. To enable reliable attack detection, we employ the *Barrier-effect Sensitive Phoneme Selection* to choose a set of effective phonemes that can facilitate cross-domain sensing offline. Notably, we examine the frequency-selectivity attenuation of different barriers to identify individual effective phonemes.

Next, our system performs the *Attack Detection via Cross-domain Sensing*. It aggregates the recorded voice commands at the wearable device and leverages the built-in speaker and accelerometer of the wearable device to perform cross-domain sensing. We develop a high-pass filter that preprocesses the accelerometer measurements to remove the interference of body movements. The *Vibration Domain Feature Extraction* is employed to derive short-timer Fourier Transform (STFT) representations from the signals as vibration-domain features. To mitigate the impacts of varying distances between the user (with the wearable) and the VA device, we design a vibration-domain normalization scheme to resolve the scale differences between the features of the wearable and the VA device. Finally, a *Thru-barrier Attack Detector* is developed to examine the similarity of two sets of vibration-domain features using 2D correlation. The correlation score can reflect the intensity of noises, and the adversary's voice (dominated by low-frequency components) with noisy audio recordings will have relatively low correlation scores. It helps to address the ambiguity induced by signal aliasing. Thru-barrier attacks will be detected via a threshold-based method.

V. Barrier-effect Sensitive Phoneme Segmentation

A. Sensitive Phoneme Selection

Some consonant phonemes (e.g., /s/, /z/) inherently have low sound intensities, leading to weak vibration signals in the vibration domain no matter the phoneme sound is from the adversary or the user. In addition, we find that people usually pronounce a few vowels (e.g., /aa/, /ao/) with high volumes due to strong larynx vibrations. The sounds of these phonemes still contain strong high-frequency components after passing the barrier. All the phonemes mentioned above are not sensitive to the barrier effect and thru-barrier attacks. Our system adopts an offline phoneme selection method to identify phonemes sensitive to the barrier effect. Specifically, we develop two criteria to determine the barrier-effect-sensitive phonemes that the system uses to detect thru-barrier attacks. **Criterion I**: the phoneme sound cannot trigger the accelerometer after passing a barrier (i.e., the phoneme sound attenuated by the barrier should not have high energy, especially the high-frequency components). **Criterion II**: the phoneme sound can trigger the accelerometer if not passing a barrier (i.e., the phoneme sound has overall high energy that is enough to be captured by the accelerometer). To determine the barrier-effect-sensitive phonemes, we first study 63 phonemes of the English language [14] and determine 37 common phonemes that are frequently used in VA voice commands [16], [17] as shown in Table II. Then, we examine spectral energy

TABLE II
IDENTIFIED COMMON TIMIT PHONEMES (SELECTED BARRIER-SENSITIVITY PHONEMES ARE MARKED IN BOLD).

phoneme : # of appearance											
t	129	**l**	70	**w**	40	*b*	31	**y**	15	**sh**	8
n	108	**k**	70	*ae*	39	*ao*	29	**aw**	15	**uh**	6
ah	107	**ch**	69	*ey*	38	**f**	29	**jh**	14		
s	101	**iy**	65	**p**	37	**v**	28	**g**	13		
r	100	**m**	65	**ay**	36	**hh**	20	**ch**	13		
ih	99	*er*	58	*aa*	32	**ng**	17	**dh**	12		
d	83	**z**	49	**uw**	31	*ow*	17	**th**	10		

distributions of the common phonemes in the vibration domain to determine the barrier-sensitive phonemes.

Specifically, we first determine the phonemes satisfying the **Criterion I** by examining the spectral energy distributions of every common phoneme sound passing typical barriers (i.e., a glass window or a wooden door). Note that these barriers have relatively large sound absorption coefficients [13] and are easier for sounds to pass through. We believe the phoneme selection method also applies to other barriers with smaller sound absorption coefficients (e.g., a concrete wall with 0.02). The experimental setup is the same as our study in Section III-B. We play the phonemes at 75dB and 85dB sound pressure levels to simulate practical attack sound volumes. The recorded phoneme sounds are sent to a Fossil Gen 5 smartwatch to generate vibration signals. We apply FFT to the vibration signals and compute the third quartile $Q_3^{adv}(p, f)$ FFT magnitude for every common phoneme p and available frequency $f \in [0, \frac{f_s}{2}$ (i.e., 75% of the record sounds with energy over this value) An example of the third quartile FFT magnitude of the phoneme /er/ passing a glass window is shown in Figure 6 (a). Next, we determine that a phoneme p satisfies the **Criterion I** if its maximum third quartile FFT magnitude over all possible frequencies is small than a threshold as shown in the following equation:

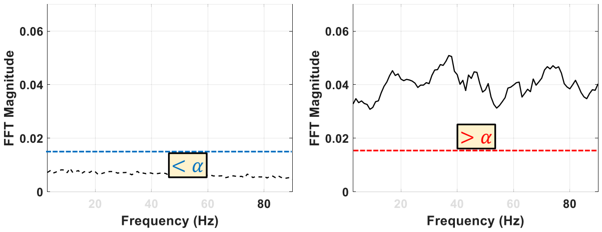
$$\arg\max_{f} Q_3^{adv}(p, f) < \alpha, \; f \in [0, \frac{f_s}{2}], \tag{2}$$

where f_s is the sampling rate of the accelerometer, we use a threshold $\alpha = 0.015$, which is empirically determined based on the FFT magnitude of ambient noises. We denote the set of phonemes satisfying the **Criterion I** as P_{adv}.

Then, we determine the phonemes satisfying the **Criterion II** by examining the spectral energy distributions of every common phoneme sound without passing a barrier. Similarly, we keep the same experimental setup as Section III-B but without the barrier. We also use the Fossil Gen 5 smartwatch to generate vibration signals and derive the third quartile $Q_3^{user}(p, f)$ FFT magnitude for the vibration signals of every common phoneme p and available frequency f. Figure 6 (b) shows an example of the third quartile FFT magnitude of phoneme /er/ without passing a barrier. We determine that a phoneme p satisfies the **Criterion II** if its minimum third quartile FFT magnitude over all possible frequencies is larger than the threshold α as shown in the following equation:

$$\arg\min_{f} Q_3^{user}(p, f) > \alpha, \; f \in [0, \frac{f_s}{2}]. \tag{3}$$

We denote the set of phonemes satisfying the **Criterion II** as P_{user}. With that, we determine the barrier-effect-sensitive phoneme set as the intersection of P_{user} and P_{adv}. In this work, we identify 31 phonemes out of 37 common phonemes are barrier-effect sensitive phonemes.



(a) Third quartile of sound energy of /er/ passing barrier

(b) Third quartile of sound energy of /er/ without barrier

Fig. 6. Demonstration of phoneme selections by applying two thresholds upon vibration signals with and without passing the barrier, respectively.

B. Deep-learning-based Phoneme Segmentation

To extract sound segments that only involve the aforementioned barrier-effective sensitive phonemes, we develop a phoneme detection scheme based on a recurrent neural network (RNN). The designed scheme operates on short frames of audio recordings sampled with a sliding time window and determines the frames containing only the effective phonemes. Compared to phoneme classification [18], [19], which aims to classify all 63 phonemes, our objective is to detect the barrier-sensitive phonemes (i.e., binary classification).

Phoneme Feature Extraction. We use Mel-frequency cepstral coefficients (MFCC) as features for phoneme detection, which have shown effective in prior work [18]. To enable fine-grained phoneme sound segmentation, we use a sliding time window with a short window length of 25ms to obtain frames (i.e., 400 data points per frame for the audio recording of 16KHz), shifting 10ms each frame. We compute Mel-frequency cepstral coefficient (MFCC) for each frame, which will be used as the inputs to the RNN model. The number of filterbank channels is set to 40, and 14-th order cepstral coefficients are computed in each frame, which represent frequency responses within 0Hz$\sim$ 900Hz. Note that we use MFCCs at relatively lower frequencies to ensure detecting phonemes in attack sounds across barriers, which usually have weak responses at high frequencies (discussed in Section III-B).

Effective Phoneme Detection with BRNN: Besides spectral patterns, a phoneme is also characterized by temporal patterns across multiple frames. It is necessary to fully leverage both time and frequency patterns to realize accurate phoneme detection. Therefore, we exploit bidirectional neural network (BRNN) architecture in our design, which has shown effectiveness in many sequential patterns learning tasks [20], [21]. Our BRNN design contains a forward and backward layer to simultaneously learn the MFCCs of the past and future. By denoting the extracted MFCCs as $X = \{x_1, x_2, ..., x_n\}$, where n denotes the total number of frames in the audio recording, we define the forward and backward operations as:

$$\begin{aligned}\overrightarrow{h}_t &= H(\overrightarrow{\omega}_1 x_t + \overrightarrow{\omega}_2 \overrightarrow{h}_{t-1} + \overrightarrow{b}),\\ \overleftarrow{h}_t &= H(\overleftarrow{\omega}_1 x_t + \overleftarrow{\omega}_2 \overleftarrow{h}_{t-1} + \overleftarrow{b}),\end{aligned} \tag{4}$$

where the derived temporal representation for the frame at t could then be computed as $h_t = h_t + h_t$. Specifically, $\overrightarrow{\omega}_1$, $\overrightarrow{\omega}_2$, $\overrightarrow{b}$ are learnable parameters for the forward layer, while $\overleftarrow{\omega}_1$, $\overleftarrow{\omega}_2$, $\overleftarrow{b}$ are the parameters of the backward layer. H is an activation function implemented with Long-Short Term Memory (LSTM) units. We empirically use 64 LSTM units in our design. To train the BRNN model, we use the TIMIT dataset that contains broadband recordings of 630 speakers of eight major dialects of American English [14]. The dataset contains audio data with time-aligned phonetic transcriptions involving 63 phonemes. Since the objective of our scheme is to detect effective phonemes (i.e., binary classification), instead of classifying phonemes, we label the effective phonemes as 1 and all other phonemes as 0. We attach a dense layer with 2 neurons to the BRNN for phoneme detection, and we train the whole model (i.e., BRNN and the dense layer) with an ADAM optimizer. To study the effectiveness of our phoneme detection scheme, we replay 6300 phoneme sound segments (100 sound samples per phoneme from five males and five females) of the TIMIT dataset using the setup in Figure 2 (b) and record the phoneme sounds before and after passing the barrier. We remove sound segments with a maximum magnitude below 0.01 since the low volume sounds would not trigger the VA device. The proposed phoneme detection scheme has an accuracy of 94% for the phoneme sounds without passing the barrier and 91% for the phoneme sounds passing the barrier, showing the effectiveness of our phoneme detection scheme. Given the detected phonemes, our system concatenates the voice sounds predicted as barrier-effect sensitive phonemes and extracts them for thru-barrier attack detection.

VI. Thru-barrier Attack Detection

A. Synchronization and Cross-domain Sensing

Our system compares the voice commands recorded by the wearable and the VA devices in the vibration domain to detect thru-barrier attacks. Compared to traditional data-driven approaches [4], [5], our comparison-based attack detection leveraging two devices enlarges the differences between vibration signals of the adversary and the legitimate user, making it possible to remove the need of collecting training data for attack detection. Such a comparison requires to trigger the voice command collection processes simultaneously on the wearable and VA devices. As wearable and VA devices are usually connected to the same local WiFi network, we design a WiFi communication-based method to coordinate the voice command recording on both devices. Upon detecting a wake word on the VA device, our system sends a triggering message through WiFi to notify the wearable device for recording the voice command. To address the residual synchronization errors caused by network delay (e.g., around 100ms), we design a cross-correlation-based method to estimate the delay:

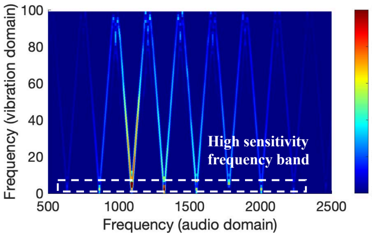


Fig. 7. Vibration response of a wearable's accelerometer (i.e., Fossil Gen 5) to an audio chirp signal (500Hz - 2500Hz). The accelerometer has a highly sensitive frequency range of 0 $\sim$ 5Hz.

$$Corr(\tau) = \sum_{n=0}^{N-1} s_v(n) s_w(n+\tau), \tag{5}$$

where $\tau_{est} = \arg\max_\tau(Corr(r))$. s_v and s_w are the audio signals of the voice commands recorded by the VA and the wearable device, respectively. N is the maximum length of the audio signals, and τ_{est} is the estimated delay in terms of data points. We remove the first τ_{est} data points from s_w to sure the starting point of the voice command is the same as s_v. Our system then performs barrier-sensitive phoneme detection on the VA device based on s_v and obtains the segments of effective phonemes. The segmentation information is sent to the wearable devices for phoneme segmentation for s_w. As mainstream VA devices do not come with an accelerometer, our system aggregates the two sets of phoneme sounds at the wearable device for cross-domain sensing. The system sequentially replays the phoneme sounds of the wearable and the VA device and then collects two sets of vibration signals with the accelerometer on the same wearable.

B. Vibration-domain Feature Extraction

Time-frequency Representation Derivation. To derive meaningful representations of voice sounds in the vibration domain for attack detection, we apply short-time Fourier transformation (STFT) to the vibration signal to derive time-frequency representations. Particularly, we apply FFT on the vibration signal within a sliding window to obtain frequency representations. We empirically determine the window size and the number of FFT points to be 64. We further compute the square of FFT magnitudes to obtain the power representations. By sliding the window across the time-series vibration signals and repeating the process, we can obtain the spectrogram representing the vibrations in time and frequency dimensions.

Accelerometer Artifact Mitigation. Accelerometers are sensitive to low-frequency vibrations due to their design purpose of capturing low-frequency body movements. We demonstrate this characteristic of the accelerometer in Figure 7, which shows the frequency responses of a smartwatch's accelerometer to a chirp signal of 500Hz$\sim$ 2500Hz. We can observe strong responses within 0Hz$\sim$ 5Hz. Such a high sensitivity artifact may amplify the attack sounds in the vibration domain even though they have low volumes, making them hard to detect in our system. As such, we propose to crop the spectrogram by removing the values corresponding to 5Hz and below. It not only ensures reliable defense but

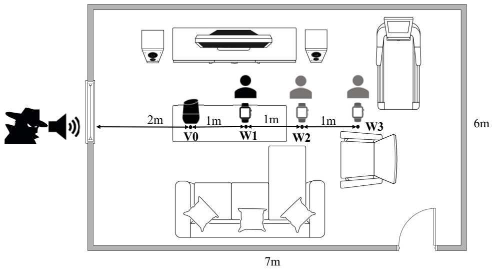


Fig. 8. Example of one of four room setups.

also mitigates the interference of body movements involved in daily activities, which generally impacts low-frequency sensor readings (e.g., 0.3Hz∼ 3.5Hz [22]). The cropped spectrogram values are used as vibration-domain features.

C. Thru-barrier Attack Detector based on 2D Correlation

To detect thru-barrier attacks, we leverage the fact that the attack sounds through barriers are noisier in the vibration domain. This is because the accelerometer injects random noises when mapping the low-frequency sounds into the vibration domain as we discussed in Section IV-A. It helps to address the ambiguity induced by signal aliasing as the adversary's (dominated by low-frequency sounds) and the user's voice (involving both high-frequency and low-frequency sounds) become more distinguishable. We exploit a 2D-correlation-based method upon the vibration-domain features to detect such randomness for attack detection. In real-world scenarios, the user may be at any distance from the VA device, which leads to different sound volumes and vibration signal magnitudes. Therefore, before computing the 2D correlation, we normalize the vibration-domain features by dividing all values of the spectrogram by its maximum value. Given the normalized features of the wearable and the VA device, we compute the 2D correlation between the two sets of features:

$$
\begin{aligned}
\mathcal{R}(stft_v, stft_w) &= \frac{W \times V}{\sqrt{W^2 \times V^2}}, \\
s.t., W &= \sum_t \sum_f (stft_w(t,f) - \overline{stft_w})), \\
V &= \sum_t \sum_f (stft_v(t,f) - \overline{stft_v})),
\end{aligned}
\tag{6}
$$

where $stft_v$ and $stft_w$ represent the normalized features of the VA device and wearable device, respectively. Finally, a threshold is then used to detect thru-barrier attacks.

VII. Evaluation

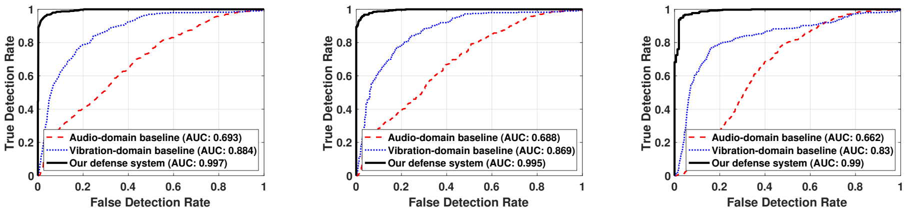
A. Experimental Methodology

Devices. To evaluate our defense system, two smartwatches, Fossil Gen 5 and Moto 360 2020 are involved to collect microphone recordings of voice commands. The microphone recordings are collected under a sampling rate of 16kHz. We use the wearable's built-in speaker and accelerometer to collect vibration signals of the voice commands. The sampling rates of the accelerometers on the two smartwatches are 200Hz. We use a Motorola Nexus 6 smartphone to emulate the VA device.

Experiment Setup. We evaluate the performance of our defense system in four different rooms (denoted as Room A, B, C, D), including one residential apartment and three university offices with the sizes of 7 × 6 m, 7 × 7 m, 6 × 4 m, and 5 × 3 m. The barrier of these four rooms is the glass window, wooden door, glass wall, and wooden door, and glass wall, respectively. For each room, we evaluate our defense system under normal situations (i.e., no attack present) and various attacks. Under normal situations, each participant in turn serves as the legitimate user to issues voice commands to the VA with an average sound pressure level of 65dB∼ 75dB at different distances as shown in Figure 8, mimicking real-world scenarios of using VA devices. We use a loudspeaker (i.e., Razer Sound Bar RC30) placed behind the room barrier (i.e., 10cm to the barrier) to generate the attack sounds. To examine the robustness of our system, we adopt different sound pressure levels in the attacks (i.e., 65dB, 75dB, 85dB), covering potential sound pressures that the adversary may issue in practice. Note that the legitimate user is present in the room under both normal situations and various attacks.

Data Collection. We collect the voice data under normal situations and attacks separately. For the normal situation, we recruit 20 participants to conduct the experiments. Among them, 10 participants conduct the experiments in both Room A and Room B, 5 participants are in Room C and the remaining 5 participants are in room D. The participants are asked to speak 20 voice commands at three distances to VA, respectively. In total, we collect 1200, 1200, 600, and 600 voice commands from Rooms A, B, C, and D, respectively. In addition, we evaluate our defense system under four types of attacks demonstrated in Section II. We replay the attack sounds through the barrier by using a loudspeaker and record the sounds with both the wearable and the VA device. We take turns considering each participant as the legitimate user and all the remaining participants as adversaries. Specifically, we leverage the voice commands of all the adversaries as attack sounds for *random attacks*, and we replay the recorded voice commands of the legitimate user to launch *replay attack*. For the *voice synthesis attack*, we train a speech synthesis model [11] using 20 voice commands of the legitimate user to generate the attack sounds. Additionally, we use 5 black-box hidden voice commands provided in prior work [3] to evaluate our defense system against hidden voice attacks. In total, we collect 26400, 3600, 3600, and 4800 samples for random, replay, voice synthesis, and hidden voice attacks, respectively.

Evaluation Metrics. *true detection rate (TDR)* is the percentage of the attack sounds that are correctly detected. *false detection rate (FDR)* is the percentage of voice commands of the user mistakenly detected as attack sounds; *equal error rate (EER)* depends on both TDR and FDR which corresponds to the threshold where two detection errors are approximately equal. *area under the ROC Curve (AUC)* measures the area under the receiver operating characteristics (ROC) curve. The ROC curve is obtained by plotting the TDR against the FDR under thresholds from 0 to 1 with a step of 0.01. The higher the AUC value, the better our system defends the attacks.



(a) Random attack

(b) Replay attack

(c) Voice synthesis attack

Fig. 9. Performance of our defense system under clear voice attacks.

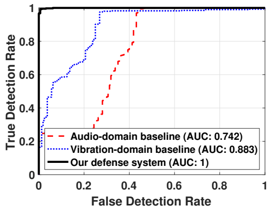


Fig. 10. Performance of our system under hidden voice attack.

B. Defense Against Clear Voice Attacks

Against Random Attack. We first evaluate the performance of our system against random attack. Specifically, we take turns considering each participant as the legitimate user and the remaining participants as adversaries. We leverage the voice commands of all the adversaries as attack sounds for random attacks. As shown in Figure 9 (a), each ROC curve corresponds to the average performance against random attack in a specific domain under different physical settings. We can observe that AUC reaches 0.693 under audio domain, 0.884 under vibration domain without phoneme selection, and 0.994 under vibration with phoneme selection, respectively. Moreover, EER is 37.4% under audio domain, 21% under vibration domain without phoneme selection, and 3.8% under vibration domain with phoneme selection. It is evident that simply adopting vibration domain can obviously improve the robustness (i.e., 0.19 AUC improvement and 16% EER reduction) against random attack. Notably, after adopting phoneme selection under vibration domain, the robustness can be further improved largely (i.e., 0.11 AUC improvement and 17% EER reduction). Those results demonstrate that vibration domain is much more effective when detecting random attacks compared to audio domain. And phoneme selection mechanism improve the performance against random attack.

Against Replay Attack. We next evaluate the performance of our system against replay attack. We use a loudspeaker to replay the recorded voice commands of the legitimate user for launching replay attack. As shown in Figure 9 (b), each ROC curve corresponds to the average performance against replay attack in each domain under different physical settings adopted in our experiment. We observe that our system respectively achieves AUC of 0.688 under audio domain, 0.869 under vibration domain without phoneme selection, and 0.995 under vibration with phoneme selection. Moreover, EER is 37.5% under audio domain, 20.7% under vibration domain without phoneme selection, and 3.5% under vibration domain with phoneme selection. It is obvious that vibration domain without phoneme selection already significantly improve the robustness (i.e., 0.18 AUC improvement and 17% EER reduction) against replay attack. After adopting phoneme selection under vibration domain, the robustness is further largely improved (i.e., 0.12 AUC improvement and 17% EER reduction). Those results show that vibration domain obviously outperform audio domain for defending replay attacks.

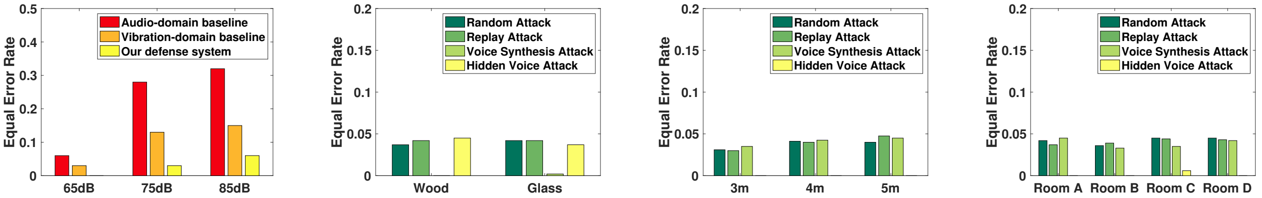
Against Voice Synthesis Attack. We further evaluate our system's performance against voice synthesis attack which can generate synthesized voice commands based on only several sound samples from the legitimate user. As shown in Figure 9 (c), each ROC curve corresponds to the average performance against voice synthesis attack in each domain. We can observe that AUC respectively reaches 0.662 under audio domain, 0.83 under vibration domain without phoneme selection, and 0.99 under vibration with phoneme selection. Moreover, EER is 37% under audio domain, 20.5% under vibration domain without phoneme selection, and 3.9% under vibration domain with phoneme selection. It is evident that simply adopting vibration domain can significantly improve the robustness (i.e., 0.17 AUC improvement and 17% EER reduction) against voice synthesis attack. After adopting phoneme selection under vibration domain, we observe that the robustness can be further largely improved (i.e., 0.16 AUC improvement and 17% EER reduction). Those results show that vibration domain without phoneme significantly outperforms audio domain when defending voice synthesis attack.

C. Defense Against Hidden Voice Attacks

Except clear voice attack, we further evaluate the performance of our system against hidden voice attacks which is more stealthy since they can only be recognized by machines and not by humans. As shown in Figure 10, each ROC curve corresponds to the average performance against hidden voice attack in each domain under different physical settings (e.g., barriers, device layouts, and rooms). We observe that AUC reaches 0.742 under audio domain, 0.883 under vibration domain without phoneme selection, and 1 under vibration with phoneme selection, respectively. Moreover, EERs are 0.35 under audio domain, 23.1% under vibration domain without phoneme selection, and 6% under vibration domain with phoneme selection. It is evident that simply adopting vibration domain can already obviously improve the robustness (i.e., 0.14 AUC improvement and 12% EER reduction) against hidden voice attack. Notably, after adopting phoneme selection under vibration domain, the robustness can be further improved largely (i.e., 0.12 AUC improvement and 22% EER reduction). Those results demonstrate that vibration domain is much more effective when detecting hidden voice attack compared to audio domain and our phoneme selection mechanism could further improve the robustness.

D. Defense Under Different Impacting Factors

Impact of Sound Pressure Levels. We study the robustness of our system under replay attacks with different sound



(a) Different attack sound volumes (b) Different barrier materials (c) Different barrier-to-VA distances (d) Different room environments.

Fig. 11. Performance of our defense system under different real-world impacting factors.

volumes, including 65dB, 75dB, and 85dB, which may be leveraged by the adversary in practice. We mainly consider replay attack as it is highly accessible to adversaries and it causes similar or higher error rates compared to the other attacks. As shown in Figure 11(a), our system can achieve less than 3.2% ERR under the sound volumes of 65dB and 75dB. In contrast, correlation in the audio domain has much higher EER, especially under the attacks with 85dB sound volume (i.e., 29.5% EER). We also find that the EER is much lower without phoneme selection. The results indicate both cross-domain sensing and our effective phoneme selection scheme helps in enabling robust defense.

Impact of Barrier Materials. To analyze the impacts of different barriers to our defense system, we compare the defense performance of barriers in rooms with two different barrier materials, i.e., wood and glass. Specifically, the barriers of Room A and D are made out of glass (i.e., a glass window and a glass wall) and the barriers of Room B and C are made out of wood. Figure 11(b) shows the EERs of our system under the four types of attacks. We can find that the EERs are similar across the two materials and are all below 4.2%. The results demonstrate our system has consistent performance in rooms with different barrier materials.

Impact of Distances. We further study the impacts of different distances between the barrier and the VA device. Specifically, we keep the barrier-to-wearable as 2m and change the barrier-to-VA distance. Figure 11(c) shows the EERs of our system under three different barrier-to-VA distances. We can find that our system has lower than 4.6% EERs under all the distances. We can also find that the EERs are slightly higher at the distance of 5m, mainly due to the lower sound quality of the user due to long distance to VA (i.e., 3m). Several voice commands of the user are mistakenly detected as attack sounds. We also conduct another set of experiments by changing the barrier-to-wearable distance, and our system has a similar performance. In general, our system is reliable under different barrier-to-VA and barrier-to-wearable distances.

Impact of Room Environments. We analyze the performance of our system in rooms with different sizes. The results of the four rooms are shown in Figure 11(d). We can find that for all room environments, the EER are below 5%, showing the scalability of our system to different room environments. We also find that our system is highly effective against hidden voice attacks, with close to 0% EERs. This is because the hidden voice commands reside in a wider frequency range (e.g., $0 \sim 6$kHz) compared to clear voice attacks, making the frequency-selectivity attention of the barrier more obvious.

VIII. Related Work

Audio-Domain Voice Authentication. The conventional user authentication methods mainly rely on training a model to extract the voice characteristics in the audio domain to identify the user [1], [4], [23]. Voice authentication systems usually examine the unique voice features, for instance, Mel-Frequency Cepstral Coefficients (MFCCs) [4] and Spectral Subband Centroids (SSCs) [23], to differentiate people's voices. Nevertheless, these voice authentication methods solely relying on audio-domain features are vulnerable to acoustic-based attacks (e.g., replay attacks [1]).

Vibration Domain Speech Recognition. Motion sensors (i.e., accelerometers and gyroscopes) have been shown to be able to pick up speech in the vibration domain [15]. WALNUT [24] simulates the physical attack of sound injection on an accelerometer and shows that the accelerometer can be affected by acoustic interference. In addition, the sound from an external speaker has been proven to impact the motion sensor. By way of illustration, Gyrophone [25] has shown that gyroscopes can be used in attacks (i.e., deriving voice content) to examine speakers' voices when the gyroscopes are deployed to the same solid surface of the speakers. In addition, EchoVib [26] verifies the speech played back by the device's loudspeaker and further performs user verification by examining the human speech's unique effects on built-in motion sensors. However, none of those works explore vibration-domain information for thru-barrier attack detection.

Voice Authentication Using Second Factors. Recently, two-factor authentication schemes are gaining attention. For instance, a two-microphone authentication (2MA) [27] system takes advantage of the presence of multiple microphones being present in an ecosystem to authenticate the source of the command. 2MA authentication framework demonstrates that such construction works using independent devices (e.g., a mobile phone and a voice assistant) increase the effort required by an attacker to inject such commands successfully. Moreover, Listening-Watch [28] has been proposed as a low-effort two-factor authentication system using speech signals based on a wearable device and active sounds that are resistant to co-located and remote attack. Furthermore, VAuth [29] requires the user to wear an additional device that is in continuous contact with the user's body provides continuous authentication for voice assistants with high accuracy and very low false-positive rate. However, those two-factor authentication approaches require more cumbersome operations to confirm each voice command, which can easily lead to wrong operations. The closest research to this work is WearID [30]. It leverages accelerometers in users' wearables to directly capture and verify users' voice commands. Such an approach is not convenient in practice as it requires users to speak in close proximity to their wearables (i.e., less than 30cm).

IX. Conclusion

This paper presents a training-free cross-domain defense system that protects VA systems against stealthy thru-barrier attacks. Our system utilizes the cross-domain sensing technology in wearables to convert voice commands from the audio domain to the vibration domain, which enhances the frequency-selective attenuation effects of barriers and facilitates the thru-barrier attack detection. We identify the phonemes sensitive to the barrier effects and develop the sensitive phoneme detection method to ensure robust thru-barrier attack detection. Our correlation-based attack detection method is training-free and can effectively detect various attacks. Extensive experiments show that the proposed system can effectively defend against thru-barrier attacks implemented with various attack approaches.

X. Acknowledgment

This work was partially supported by the National Science Foundation Grants CCF1909963, CCF2000480, CCF2028858, CCF2028876, CNS1526524, CNS1714807, CNS1801630, CNS2030501, CNS2120276, CNS2120396, CNS-2145389, and OAC-1547350

References

[1] J. Lindberg and M. Blomberg, "Vulnerability in speaker verification-a study of technical impostor techniques," in *6th European Conference on Speech Communication and Technology (ISCA Eurospeech)*, 1999.

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE ICASSP)*, 2018, pp. 4779–4783.

[3] N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou, "Hidden voice commands," in *Proceedings of USENIX Security Symposium*, 2016, pp. 513–530.

[4] C. Hanilçi, "Features and classifiers for replay spoofing attack detection," in *2017 10Th international conference on electrical and electronics engineering (ELECO)*. Ieee, 2017, pp. 1187–1191.

[5] M. Todisco, H. Delgado, and N. W. Evans, "A new feature for automatic speaker verification anti-spoofing: constant q cepstral coefficients." in *Odyssey*, vol. 2016, 2016, pp. 283–290.

[6] WeChat, "Voiceprint," 2017, https://thenextweb.com/apps/2015/03/25/wechat-on-ios-now-lets-you-log-in-using-just-your-voice/.

[7] Google, "How you sign in with 2-step verification," 2019, https://support.google.com/accounts/answer/1085463?hl=en.

[8] A. P. Felt, E. Ha, S. Egelman, A. Haney, E. Chin, and D. Wagner, "Android permissions: User attention, comprehension, and behavior," in *Proceedings of the 8th Symposium on Usable Privacy and Security (ACM SOUPS)*, 2012, pp. 1–14.

[9] P.-C. Wu, C.-Y. Yeh, H.-H. Tsai, and Y.-Z. Juang, "Low-frequency noise reduction technique for accelerometer readout circuit," in *2016 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*. IEEE, 2016, pp. 483–486.

[10] M. E. Ahmed, I.-Y. Kwak, J. H. Huh, I. Kim, T. Oh, and H. Kim, "Void: A fast and light voice liveness detection system," in *Proceedings of USENIX Security Symposium*, 2020, pp. 2685–2702.

[11] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, 2018, pp. 4485–4495.

[12] T. L. Szabo and J. Wu, "A model for longitudinal and shear wave propagation in viscoelastic media," *The Journal of the Acoustical Society of America*, vol. 107, no. 5, pp. 2437–2446, 2000.

[13] M. Vorländer and J. E. Summers, "Auralization: Fundamentals of acoustics, modelling, simulation, algorithms, and acoustic virtual reality," *Acoustical Society of America Journal*, vol. 123, no. 6, p. 4028, 2008.

[14] J. S. Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium, 1993*, 1993.

[15] S. A. Anand and N. Saxena, "Speechless: Analyzing the threat to speech privacy from smartphone motion sensors," in *2018 IEEE Symposium on Security and Privacy (IEEE SP)*, 2018, pp. 1000–1017.

[16] "Every alexa command for your amazon echo speaker or display available now," https://www.cnet.com/home/smart-home/every-alexa-command-you-can-give-your-amazon-echo-smart-speaker-or-display/, 2021.

[17] "101 google assistant commands: The best things to ask your google home," https://www.the-ambient.com/guides/best-google-assistant-commands-382, 2021.

[18] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[19] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang *et al.*, "A comparative study on transformer vs rnn in speech applications," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (IEEE ASRU)*, 2019, pp. 449–456.

[20] C. Shi, X. Guo, T. Yu, Y. Chen, Y. Xie, and J. Liu, "Mobile device usage recommendation based on user context inference using embedded sensors," in *2020 29th International Conference on Computer Communications and Networks (IEEE ICCCN)*, 2020, pp. 1–9.

[21] A. Shewalkar, D. Nyavanandi, and S. A. Ludwig, "Performance evaluation of deep neural networks applied to speech recognition: Rnn, lstm and gru," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 9, no. 4, pp. 235–245, 2019.

[22] G. Plasqui, A. Bonomi, and K. Westerterp, "Daily physical activity assessment with accelerometers: New insights and validation studies," *Obesity reviews : an official journal of the International Association for the Study of Obesity*, 2013.

[23] T. Kinnunen, B. Zhang, J. Zhu, and Y. Wang, "Speaker verification with adaptive spectral subband centroids," in *International Conference on Biometrics (ICB)*. Springer, 2007, pp. 58–66.

[24] T. Trippel, O. Weisse, W. Xu, P. Honeyman, and K. Fu, "Walnut: Waging doubt on the integrity of mems accelerometers with acoustic injection attacks," in *Security and Privacy (EuroS&P), 2017 IEEE European Symposium on*. IEEE, 2017, pp. 3–18.

[25] Y. Michalevsky, D. Boneh, and G. Nakibly, "Gyrophone: Recognizing speech from gyroscope signals," in *Proceedings of USENIX Security Symposium*, 2014, pp. 1053–1067.

[26] S. A. Anand, J. Liu, C. Wang, M. Shirvanian, N. Saxena, and Y. Chen, "Echovib: Exploring voice authentication via unique non-linear vibrations of short replayed speech," in *Proceedings of the Asia Conference on Computer and Communications Security (ACM AsiaCCS)*, 2021, pp. 67–81.

[27] L. Blue, H. Abdullah, L. Vargas, and P. Traynor, "2ma: Verifying voice commands via two microphone authentication," in *Proceedings of the 2018 on Asia Conference on Computer and Communications Security (ACM AsiaCCS)*, 2018, pp. 89–100.

[28] P. Shrestha and N. Saxena, "Listening watch: Wearable two-factor authentication using speech signals resilient to near-far attacks," in *Proceedings of the 11th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, 2018, pp. 99–110.

[29] H. Feng, K. Fawaz, and K. G. Shin, "Continuous authentication for voice assistants," in *Proceedings of Annual International Conference on Mobile Computing and Networking (ACM MobiCom)*, 2017, pp. 343–355.

[30] C. Shi, Y. Wang, Y. Chen, N. Saxena, and C. Wang, "Wearid: Low-effort wearable-assisted authentication of voice commands via cross-domain comparison without training," in *Annual Computer Security Applications Conference (ACSAC)*, 2020, pp. 829–842.