WearID: Low-Effort Wearable-Assisted Authentication of Voice Commands via Cross-Domain Comparison without Training

Cong Shi
WINLAB, Rutgers University
New Brunswick, NJ, US
cs1421@winlab.rutgers.edu

Yan Wang
Temple University
Philadelphia, PA, US
y.wang@temple.edu

Yingying Chen
WINLAB, Rutgers University
New Brunswick, NJ, US
yingche@scarletmail.rutgers.edu

Nitesh Saxena
University of Alabama at Birmingham
Birmingham, AL, US
saxena@uab.edu

Chen Wang*
Louisiana State University
Baton Rouge, LA, US
chenwang1@lsu.edu

ABSTRACT

Due to the open nature of voice input, voice assistant (VA) systems (e.g., Google Home and Amazon Alexa) are vulnerable to various security and privacy leakages (e.g., credit card numbers, passwords), especially when issuing critical user commands involving large purchases, critical calls, etc. Though the existing VA systems may employ voice features to identify users, they are still vulnerable to various acoustic-based attacks (e.g., impersonation, replay, and hidden command attacks). In this work, we propose a training-free voice authentication system, *WearID*, leveraging the *cross-domain* speech similarity between the audio domain and the vibration domain to provide enhanced security to the ever-growing deployment of VA systems. In particular, when a user gives a critical command, WearID exploits motion sensors on the user's wearable device to capture the aerial speech in the vibration domain and verify it with the speech captured in the audio domain via the VA device's microphone. Compared to existing approaches, our solution is low-effort and privacy-preserving, as it neither requires users' active inputs (e.g., replying messages/calls) nor to store users' privacy-sensitive voice samples for training. In addition, our solution exploits the distinct vibration sensing interface and its short sensing range to sound (e.g., 25cm) to verify voice commands. Examining the similarity of the two domains' data is not trivial. The huge sampling rate gap (e.g., 8000Hz vs. 200Hz) between the audio and vibration domains makes it hard to compare the two domains' data directly, and even tiny data noises could be magnified and cause authentication failures. To address the challenges, we investigate the complex relationship between the two sensing domains and develop a spectrogram-based algorithm to convert the microphone data into the lower-frequency " motion sensor data" to facilitate cross-domain comparisons. We further develop a user authentication scheme to verify that the received voice command originates from the legitimate user based on the cross-domain speech similarity of the received voice commands. We report on extensive experiments to evaluate the WearID under various audible and inaudible attacks. The results show WearID can verify voice commands with 99.8% accuracy in the normal situation and detect 97.2% fake voice commands from various attacks, including impersonation/replay attacks and hidden voice/ultrasound attacks.

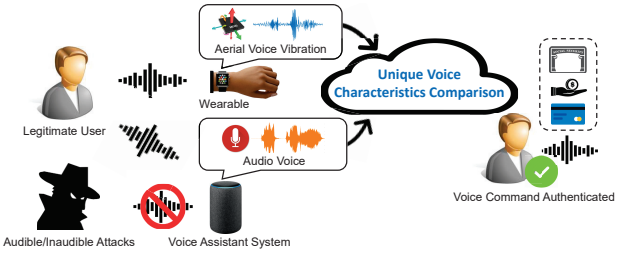


Figure 1: Illustration of the proposed idea in WearID - exploring the complement of the vibration domain to defend against audio-based attacks, e.g., impersonation, replay, hidden voice command and ultrasound attacks.

ACM Reference Format:
Cong Shi, Yan Wang, Yingying Chen, Nitesh Saxena, and Chen Wang. 2020. WearID: Low-Effort Wearable-Assisted Authentication of Voice Commands via Cross-Domain Comparison without Training. In *Annual Computer Security Applications Conference (ACSAC 2020), December 7–11, 2020, Austin, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3427228.3427259

1 INTRODUCTION

In recent years, smart devices (e.g., Google Home and Amazon Alexa) have incorporated advanced speech recognition technologies that enable the devices to understand natural language and take voice commands. By using voices as inputs, users can smoothly and conveniently interact with their voice assistant (VA) systems to accomplish numerous daily tasks, such as playing music, managing calendar events, shopping online and controlling smart home appliances. With the growing trend of using VA systems, more and more people tend to use voice commands to complete important tasks. For example, making a big purchase (e.g., over 100 dollars), unlocking the entrance door to a house, or making a critical call

*This work was done when Chen Wang was a graduate student at Rutgers University.

ACSAC 2020, December 7–11, 2020, Austin, USA

ACM ISBN 978-1-4503-8858-0/20/12…$15.00
https://doi.org/10.1145/3427228.3427259

Table 1: Comparing WearID with potential security solutions to secure critical voice commands.

	Requiring user active input	Requiring training and storing voice templates	Requiring audio playback	Requiring dedicated sensor	Vulnerable to audible attacks [30, 42]	Vulnerable to inaudible attacks [12, 51]
WearID (Our Solution)	×	×	×	×	×	×
One-tap two-factor authentication [37]	✔	×	×	×	×	×
SMS/call-based two-factor authentication [21]	✔	×	×	×	×	×
Audio CAPTCHA (suggested in [12])	✔	×	×	×	×	×
Voice biometric-based authentication [39, 44]	×	✔	×	×	✔	✔
Two microphone authentication (2MA) [10]	×	×	✔	×	✔	×
Defenses against inaudible attacks [12, 51]	×	✔	×	×	✔	×
Defense with smartphone motion sensor [46]	×	✔	✔	×	✔	×
VAuth [20]	×	×	×	✔	×	×

(e.g., calling a bank for conducting transactions [45]). We call these voice commands as *highly critical commands* since the commands could access highly sensitive information and functionalities (e.g., credit card numbers, passwords, and payments). The significant financial benefits of using such highly critical commands lure adversary into faking the user's commands and put the user's privacy and property under high risks. For instance, the adversary can get a user's credentials for accessing personal devices by asking, "OK, Google, what is my password?" [8] The adversary can also make a significant amount purchase through the user's associated account [9] by telling the VA system "Alexa, Order a MacBook from Prime Now." When the adversary can access the VA system at home remotely (e.g., through a hacked Smart TV), the adversary can even use critical commands to control security critical IoT devices [15], such as disarming a smart locking system and gain entry into the house. To ensure the successful large-scale deployment of VA systems, it is critical to address these inherited security vulnerabilities in VA systems and bring trustworthiness to users. In this work, we thus aim to design a low-overhead system with enhanced security that could protect highly critical commands in VA systems.

Existing Solutions. Existing authentication and defense mechanisms for VA systems relying on voice biometric technologies [3, 22, 26, 39, 44, 46] use users' unique sound characteristics and machine learning-based models for user authentication. These solutions solely rely on acoustic features in the *audio domain* (i.e., extracting information from the data captured by microphones). Thus they are vulnerable to acoustics attacks, either audible attacks (e.g., replay attacks [30] and impersonation attacks [42]) or more surreptitious inaudible attacks (e.g., hidden voice commands [12] and ultrasound attacks [51]). To add another layer of defense, some VA systems exploit a second factor to secure voice commands, such as challenge questions via audio CAPTCHA [12], replay messages/calls [21], or virtual buttons [37] on the user's mobile device (e.g., smartphone). However, these approaches require significant user efforts to confirm the authenticity of each single voice command. Furthermore, they could be prone to user careless behaviors [19] of habituations of confirming, meaning the attack attempts may be accepted without paying attention. Recently, VAuth [20] develop a system that utilizes the user's facial vibrations captured by accelerometers embedded in a pair of glasses for user authentication. The dedicated sensors requiring a high sampling frequency of 11kHz entail additional costs, making the system not practical. While 2MA [10] needs to use multiple spatially distributed microphones, which leads to extra cost and considerable energy consumption. Moreover, this approach only works in the audio domain alone so that they are still vulnerable to the attacks in the audio domain. We summarize the weaknesses of the state-of-the-art voice authentication techniques in Table 1.

Our Approach. In this paper, we explore the feasibility of leveraging wearables' accelerometers to harness the aerial voice vibrations corresponding to live human speeches for user authentication. We propose a low-effort training-free user authentication system, *WearID*. It utilizes the wearable as a personal identity token and performs cross-domain authentication (audio vs. vibration) to verify the identity of the person who gives the voice command. WearID provides a scalable solution that would enable using VA under high-security-level scenarios (e.g., nuclear power stations, stock exchanges, and data centers), where all voice commands are critical and desire around-the-clock authentication. It is also compatible with existing voice-based authentication methods in VA systems (e.g., Google Voice Match and Amazon Alexa Voice Profile), where WearID could be invoked when critical commands are detected. Compared with existing voice biometric technologies [39, 44], WearID does not require extra user efforts (e.g., answering challenging questions and replying messages/calls) or additional training using privacy-sensitive voice samples. In addition, WearID reuses wearable devices that have already been widely accepted worldwide (i.e., 593 million in 2018 [40]), making it low-cost and more practical than the two most similar approaches, VAuth [20] and 2MA [10]. Moreover, our solution is different from existing two-factor authentication methods using co-location information (e.g., WiFi [29], Bluetooth [38], and ambient sound and light [23]), since it can resist the acoustic attacks as mentioned above.

The basic idea of WearID is examining the similarity between the unique voice characteristics in the aerial speech vibration and the audio voice for user authentication. As shown in Figure 1, triggered by a wake word detected at the VA device, WearID exploits the wearable's accelerometer and VA's microphone to capture voice commands in the vibration domain and audio domain at the same time, respectively. The voice commands recording data are sent to a cloud server for user authentication. To realize the cross-domain similarity comparison, we develop a training-free algorithm that converts high-fidelity microphone data into a low-fidelity aliasing form and correlates the time-frequency characteristics of the

speech signals in the vibration domain and the audio domain to verify the voice command. The algorithm could be easily integrated with existing VA systems and wearables without any hardware modification.

Recent studies [32, 52] have shown the initial success of using motion sensors on smartphones to capture the speaker's voice. However, examining the cross-domain similarity in practical scenarios using aerial speech vibrations captured by motion sensors in wearables is nontrivial. *First*, the unique response of the wearable's motion sensor to aerial speech and the associated acoustic characteristics in the vibration domain remains unclear. *Second*, the heterogeneous hardware designs and the huge sampling rate gap (e.g., $8000Hz$ vs. $200Hz$) make it hard to compare the acoustic characteristics from the vibration and audio domains directly. Thus, we must quantify the relationship between two distinct domains to support a training-free user authentication approach. *Third*, the synchronization of the data collection in totally different hardware is difficult. *Fourth*, the proposed system should defend against various audible [30, 42] and inaudible attacks [12, 51].

To ensure reliable cross-domain comparison, we extensively study response distances and unique characteristics of aerial speech vibrations captured by wearables. We develop a spectrogram-based method to model the complex relationship between the voice command signals in the vibration and the audio domains and enable similarity comparison between them. Particularly, we propose to convert the spectrogram of high-frequency microphone data to the low-frequency aliasing one that is comparable to the accelerometer spectrogram. To enhance the reliability, we quantify the frequency selectivities of the accelerometer and the microphone and select the frequency components that are sensitive for both sensing modalities for comparison. Moreover, to address the residual synchronization errors caused by network delay, we develop a 2D-correlation based method to align the spectrograms of the two sensing domains through searching for an offset that results in the maximum correlation.

Our Contributions:

- We show that the aerial speech vibrations of human voices can be captured by the accelerometer embedded in wearable devices. This could serve as an additional domain (i.e., vibration domain) to the original audio domain to verify the highly critical commands of the user and provide enhanced security for the VA system.
- We propose a unique voice command authentication system, WearID, which can be easily integrated with the existing VA systems and wearable devices without making any hardware modifications. The system is low-effort and privacy-preserving as it does not require any prior training, and therefore does not need to store privacy-sensitive voice sample templates.
- We leverage the accelerometer's short response distance to voice to effectively prevent the impersonation/replayed sounds from accessing the wearable. We derive the unique spectral relationship between the aerial speech vibrations captured by wearables' accelerometers and the audio recorded by VA's microphones, we propose cross-domain comparison that can effectively examine the similarity of weak and low-resolution signals in the vibration domain the and speech signals in the audio domain.
- We conduct extensive experiments and user studies with different smartwatches models and participants, which result in 600 human voice segments. The results show that WearID can authenticate user's voice commands with 99.8% accuracy in the normal situation and detect 97.2% of various impersonation and replay attacks with a low false negative rate of 2%. When under the hidden voice and ultrasound attacks [51], WearID achieves close to 100% accuracy of authenticating the users.

2 RELATED WORK

Audio-domain Voice Authentication and Security Issues. The traditional user authentication methods designed for voice access systems mainly extract each individual's voice features in the audio domain to identify users [11, 25, 26, 36, 44, 48]. Mel-Frequency Cepstral Coefficients (MFCCs) [33] and Spectral Subband Centroids (SSCs) [28] describe a voice's timbre and vocal-tract resonances and are widely used as unique voice features to distinguish users. The modulation frequency [6] capturing formant and energy transition details of a voice sound contains speaker-specific information for user identification. However, only relying on the audio-domain features has been shown to be vulnerable to acoustic-based attacks. For example, an adversary can spoof the legitimate user to pass a voice authentication system by recording and replaying a user's voice sound [30]. Moreover, the adversary can study the user's daily speech to impersonate or synthesize the user's voice to pass the voice authentication [17, 30, 42].

WearID Versus Other Authentication Methods. Rather than using voice features, recent research studies propose to defend against replay attacks by determining the liveness of sound source [14, 53, 54]. Specifically, Chen *et al.* [14] examine the unique magnetic field patterns generated by electro-acoustic transducers to detect loudspeaker-generated voice. VoiceLive [54] and VoiceGesture[53] detect the dynamic acoustic characteristics (via time-difference-of-arrival and Doppler shifts) that only occur in human voices to identify liveness. However, these approaches focus on smartphones and require users to place the smartphone's microphone close to the mouth. Thus, they are not applicable to the VA systems (e.g., Google Home and Amazon Alexa) that allow users to give voice commands freely from a distance. Feng *et al.* [20] develop a user verification system for the VA systems by capturing the user's facial vibrations via an accelerometer with high-sampling rate (i.e., 11kHz) embedded in a pair of glasses. The vibrations are then compared with the voice recorded by the VA system to verify whether the voice command is given by the legitimate user wearing the glasses. In contrast, WearID addresses a more challenging problem as it studies much weaker and low-resolution aerial speech vibration signals sampled by wearable accelerometers at 100Hz. With such a capability, WearID can be seamlessly integrated into wrist-worn wearable devices (e.g., smartwatches, fitness trackers) that are already widely accepted worldwide.

Vibration-domain Voice Recognition. Recent studies show that the MEMS motion sensors (e.g., accelerometer and gyroscope) are able to capture acoustic sounds [5, 16, 24, 32, 52]. Gyrophone [32] utilizes the gyroscope in a smartphone to recognize the speaker's information (e.g., gender and speaker identity) from the speech played by a loudspeaker. Accelword [52] leverages the accelerometer in

a smartphone rather than a microphone to recognize the user's wake word sound(e.g., Siri), which reduces the energy consumption. Speechless [4] further analyzes the speech privacy leakage, including the speech content from the smartphone motion sensors under various attacking scenarios. These works require much effort to train the system with motion sensor data and do not reveal the relationship between the sensor readings and real voice recorded by microphones. Spearphone [5] uses the accelerometer of the smartphone to eavesdrop speeches from the vibrations generated from the built-in loudspeaker, which requires the accelerometers and the loudspeakers on the same device. The impacts of aerial speech vibrations on the motion sensors in wearables are not yet clear.

3 ATTACK MODEL

We consider an adversary who is interested in obtaining the user's private/sensitive information or exerting an unpermitted operation through critical voice commands on the VA device shared among multiple users (e.g., at an office or home). We assume that the adversary cannot physically break the VA device, take control of the VA cloud service, or get the possession of the user's wearable device. We summarize the potential attacks in two major categories:

Attack on User's Absence. This type of attacks can be launched when the user is away from the VA device. The adversary tries to get close to the VA device and fool the VA system by using his own voice or audio playback techniques:

- *Random Attack.* The adversary does not have the prior knowledge of the victim's voice and attempts to fool the VA system with his own voice. Despite the simple approach, such attacks can achieve a considerable attack success rate of about 3.5% [50] on state-of-the-art speaker verification approaches.
- *Impersonation Attack.* An experienced adversary that has the knowledge of the victim's voice could try to spoof the VA system by mimicking the victim's voice. The adversary can produce the voice sound by using speech synthesis techniques and playback devices (e.g., loudspeaker).
- *Replay Attack.* The adversary tries to capture the victim's voice commands via a recording device (e.g., the microphone of a smartphone) and replay the recorded voice via a loudspeaker, attempting to fool the VA system.

Co-location Attack. This type of attacks can be launched surreptitiously even when the victim is present near the VA device:

- *Hidden Voice Command Attack.* The adversary could inject the recorded user's voice commands into the background music or the audio channel of video streams [49]. He could also provide hidden voice commands that exploit the underlying mechanisms (e.g., GMM-HMM models [12]) of VA systems. Such attacks could stealthily spoof the VA systems without being perceived by human subjects. To avoid being noticed from the audible reply, an adversary can first control the volume to mute the VA device via hidden voice commands.
- *Ultrasound Attack.* The adversary could modulate the recorded voice commands of a victim to the ultrasound frequency band (i.e., $\geq 20KHz$), and use the modulated sound to fool the VA system stealthily. Although human ears cannot hear the modulated voice commands, they could be recognized by existing VA systems due to the non-linearity of the microphone [51].

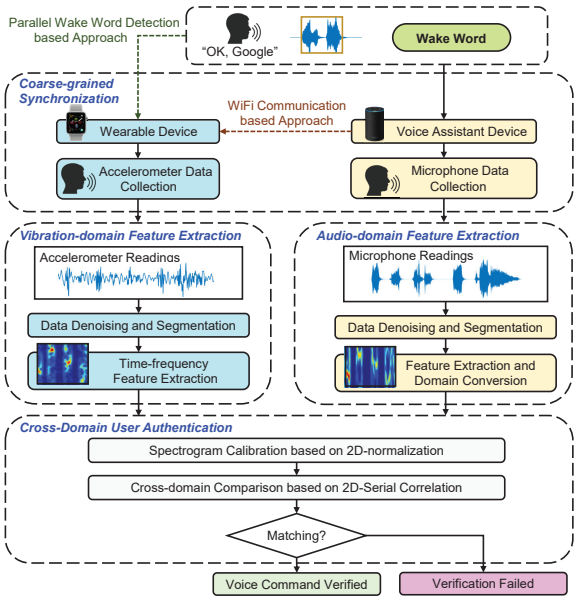


Figure 2: User authentication overview.

4 USER AUTHENTICATION DESIGN

4.1 Why Wearable? Why Motion Sensor?

Since the number of wearable users has reached half a billion worldwide [40], it is natural for us to explore such pervasiveness and use wearable devices in our design. These devices are usually worn on the user body and rarely left unattended, making it eligible as a trusted device. For example, smart wristbands have been used as a replacement to student ID card [1] since they are hard to forget to carry. As another example, smartwatches have been accepted as a convenient and valid security token for contactless payment [34]. In this paper, we propose to utilize motion sensors in commercial wearable devices (e.g., smartwatches, smart wristbands, and activity trackers) to capture users' voice commands for user authentication. We choose to use motion sensors because it captures distinct characteristics of voice sound in the vibration domain. Such unique characteristics are harder to forge compared to the voice sound captured by microphones. As a result, our system enables VA systems to resist acoustic-based attacks, including audible and inaudible attacks, which can effectively attack existing user authentication methods for VA systems. The effectiveness of WearID on defending the audible attacks and the inaudible attacks are discussed in Section 7.3 and Section 7.4, respectively.

4.2 Challenges

In order to conduct cross-domain comparison for user authentication, a number of challenges need to be addressed.

- **Weak Response to Human Speech**. Due to the design purpose of measuring acceleration force, wearables' accelerometers have weak responses to aerial speech vibrations caused by human speeches while being sensitive to human motions that are considered as noises in our system. Such inherited characteristics of accelerometers make it difficult to determine speech

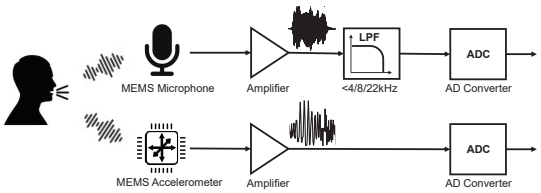
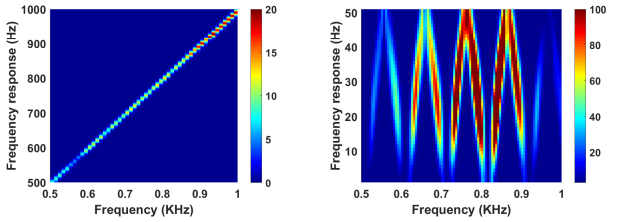


Figure 3: Hardware flow of microphone and motion sensor.



(a) VA's microphone (b) Accelerometer on Huawei watch 2 sport

Figure 4: Frequency responses of a microphone and accelerometer (Z axis) to a chirp signal ($500Hz \sim 1000Hz$).

segments and disentangle aerial speech vibrations from noisy accelerometer readings.

- **Complex Cross-domain Relationship.** Comparing human speech representations in the vibration domain and those in the audio domain is challenging. The heterogeneous hardware designs lead to distinctive frequency-selectivity patterns in accelerometer and microphone readings, making it hard to find similar acoustic characteristics through cross-domain comparisons. Moreover, the huge sampling rate gap (i.e., $8000Hz$ versus $200Hz$) of the two different sensors render any direct comparison between the vibration signals and the audio signals impossible.
- **Coarsely synchronized acoustic signals in two domains.** Network delay introduces unpredictable offsets between the vibration signals captured with the wearable and the audio signals recorded with the VA device. It is necessary to align the signals from the two sensors for a reliable comparison.

4.3 System Flow

Toward this end, we develop a wearable-assisted low-effort user authentication system, WearID, which verifies the authenticity of critical voice commands by examining the cross-domain similarity of the unique voice characteristics captured with accelerometer of the wearable device and microphone of the VA device. As illustrated in Figure 2, after a critical command/wake word is detected by the VA system, the system performs the *Coarse-grained Synchronization* to ensure that the VA and wearable devices start the data collection process simultaneously. Depending on the network condition (e.g., WiFi network delay), we develop two approaches for the *Coarse-grained Synchronization*. When the network delay is low and suitable for synchronization, the *WiFi Communication-based Approach* allows the VA device to detect the critical command/wake word, start its data collection, and send a notification to trigger the data collection on the wearable via the WiFi connection. Since there is a growing trend of having the motion sensors always activated on a wearable device (e.g., for fitness tracking), we propose an alternative solution, the *Parallel Wake-word Detection-based Approach*, for the cases of high network delay. In particular, the system exploits the accelerometers on the wearable device to detect the wake word in parallel with the VA device, which triggers the data collection on both devices separately. When the wearable and the VA device are coarsely synchronized, the voice command right after the detected wake word is recorded by both devices for user authentication using cross-domain comparison.

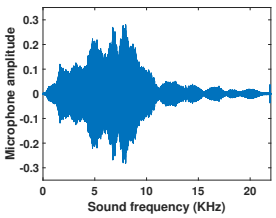
Next, WearID exploits the *Vibration Domain Feature Derivation* and *Audio Domain Feature Derivation* to derive time-frequency features from the voice command captured in the vibration domain and the audio domain, respectively. The *Vibration Domain Feature Derivation* first removes the noises caused by human motions in the accelerometer readings by using a high-pass filter and then segments the vibration signals of the voice command by examining its moving variance. Next, the time-frequency representations (i.e., spectrogram) of the voice segment are extracted and used as the vibration domain features. Similarly, the *Audio Domain Feature Derivation* denoises the microphone data and computes the spectrogram of the audio segment. Due to the huge sampling rate gap between microphone and accelerometer (e.g., 8000Hz vs. 200Hz), directly comparing the spectrograms in the two sensing domains is nearly infeasible. Therefore, we propose the *Feature Extraction and Domain Conversion*, which extracts and converts the spectrogram in the audio domain to the low-frequency aliased representations, which are comparable to the spectrogram in the vibration domain.

Finally, WearID performs *Cross-Domain User Authentication* via examining the similarity between the spectrogram of the wearable and the converted spectrogram of the microphone. The proposed system exploits *Spectrogram Calibration based on 2D-normalization* to further calibrate the spectrogram of the two sensors by normalizing their time lengths and magnitudes, which addresses the scale mismatches. Due to the time difference of triggering the microphone and the accelerometer, there exists an unpredictable relative time offset between the two spectrograms. To addresses this, we propose *Cross-domain Comparison based on 2D-Serial Correlation*, which quantifies the cross-domain similarity by finding the maximum 2D-correlation coefficient between the spectrograms by sliding one spectrogram across the other. The authentication succeeds if the maximum 2D-correlation coefficient is over a predefined threshold. Otherwise, it fails and rejects the voice command.

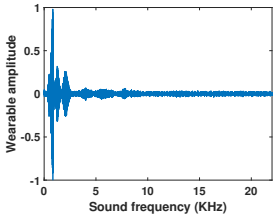
5 CAPTURING VOICE COMMANDS THROUGH VIBRATION

5.1 Relationships and Differences between Microphone and Accelerometer

Both microphone and accelerometer are Micro Electro Mechanical System (MEMS) sensors. MEMS microphones exploit a pressure-sensitive diaphragm to capture sound waves as analog signals [47], which are amplified and fed to a Low Pass Filter (LPF) with a cutoff frequency of half of the sampling frequency. An Analog-to-Digital Converter (ADC) is then applied to digitize the analog signals. Differently, MEMS accelerometers in wearable devices measure sound



(a) Amplitude of microphone



(b) Amplitude of wearable accelerometer

Figure 5: Responses of the microphone and the accelerometer (Z axis) to a chirp from $0Hz$ to $22kHz$.

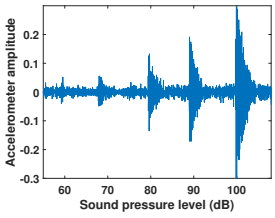


Figure 6: Response of accelerometer (Z axis) under different sound pressure levels.

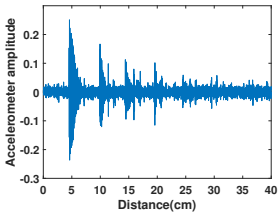


Figure 7: Response of accelerometer (Z axis) under different subject-to-wearable distances.

in terms of the vibration of the inertial mass, which is originally designed to capture the device's acceleration caused by human motion. The hardware comparison between a microphone and an accelerometer is shown in Figure 3. The accelerometer does not contain an LPF, and thus, it can capture vibration signals approaching its sensing limit, e.g., up to $4KHz$ for Invensense M6515 on LG Urbane watch 150. Such sensing capability is sufficient for capturing human voices, which typically ranges from $85Hz \sim 255Hz$. However, the missing LPF in accelerometer design results in distinctive frequency sensitivity to human voices compared with the microphone. Furthermore, vendors of wearable devices usually limit the sampling rate to below $200Hz$ (e.g., $100Hz$ on Huawei watch 2 sport) to reduce power consumption, which causes significant signal aliasing and make it even harder for cross-domain comparison.

5.2 Acoustic Response in Vibration Domain

Aliased Signal. The low sampling rate of wearable's accelerometer causing the captured speech vibrations aliased, where multiple frequencies of the vibration signals mapped to a signal frequency [32]. Figure 4 compares the spectrograms of a microphone and an accelerometer under a chirp sound from $0.5kHz \sim 1kHz$, where the accelerometer's spectrogram shows a "Zigzag" curve. This validates that a frequency in the vibration domain could correspond to multiple frequencies in the audio domain (i.e., aliased). Such aliasing effects render a direct comparison between the speech signals in the vibration domain and those in the audio domain almost impossible. Note that the aliasing effects are usually removed on the microphone with the LPF, which is missing in the accelerometer's hardware design. We model the frequency relationship between the vibration signal and the audio signal as:

$$f_{alias} = |f - Nf_s|, N \in Z, \tag{1}$$

where f_{alias}, f and f_s denotes the aliasing vibration signal frequency, audio signal frequency, and sampling rate of the accelerometer. We discuss how to perform the cross-domain comparison with accelerometer and microphone data in Section 6.3.

Unique Response to the Aerial Speech Vibrations. Due to the heterogeneous sensing mechanisms and hardware design (e.g., LPF), accelerometer and microphone show distinctive acoustic response to human speeches. To study unique acoustic response in the vibration domain, we conduct an experiment by playing a chirp using a loudspeaker and studying the response of the wearable's accelerometer. Specifically, we play an audio that sweeps from $0Hz \sim 22kHz$ by using a Logitech loudspeaker and use an wearable's accelerometer (i.e., on LG Urbane W150) and a smartphone microphone (on Nexus 6) to record the sound, where the distance between the loudspeaker and the recording devices is $10cm$. As shown in Figure 5, we find that the accelerometer has response for the sound between $400Hz$ and $3400Hz$, whereas the microphone captures sound between $80Hz$ and $15kHz$. Compared to the microphone, the accelerometer is only sensitive to sound reside in a lower frequency band. Furthermore, we find that even for the same frequency, the accelerometer has unique responses in terms of amplitude compared to that on the microphone. Such frequency selectivity makes the audible and inaudible attacks fail to reproduce a user's acoustic characteristics on the accelerometer readings, though they may succeed in synthesizing the user's voice on microphone recordings. The distinct acoustic characteristics in vibration domain thus add a layer of protection against the acoustic attacks, even the state-of-the-art audio adversarial attacks [13, 35, 49].

Recording Live Human Speech Using Wearables. We conduct an experiment to further study the sensitivity of the wearable's accelerometer on live human speeches. Particularly, we use a smartwatch (Huawei watch 2 sport) to record a voice command (i.e., "calendar") spoken by a human subject with the sound pressure levels (SPL) of $60dB$, $70dB$, $80dB$, $90dB$, and $100dB$, under an ambient noise level of $37dB$. The distance between the subject's mouth and the smartwatch is $10cm$, with the smartwatch worn on his left hand. Figure 6 shows the response of the Z-axis of the accelerometer. We can observe that the wearable can capture speech vibrations with SPL over $70dB$ and the amplitude grows with the SPL. Particularly, when the SPL of speech vibration reaches $80dB$ (presentation-level volume), the accelerometer can clearly reveal the speech vibrations, with a signal-to-noise ratio of over 9.71. This means that an SPL of $80dB$ could inject sufficient voice characteristics into the vibration readings for cross-domain comparison. We confirm this with extensive experiments shown in Section 7. We also test the capability of the wearable's accelerometer on picking up voice under various subject-to-wearable distances from $5cm$ to $35cm$ (with a $5cm$ gap). The subject speaks the same voice command (i.e., "calendar") to the smartwatch using an average SPL of $80dB$. We can observe in Figure 7 that when the distance increases to $30cm$, the response can be barely observed (with a low SNR of 2.0). Such short response distance of the accelerometer can facilitate WearID to shield against many acoustic attacks.

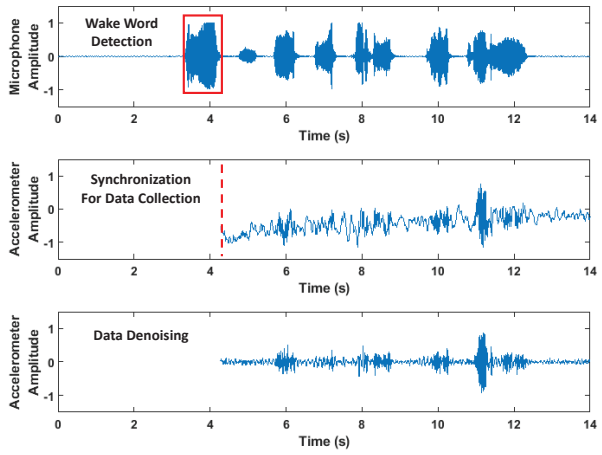


Figure 8: Synchronization of the microphone data (8000Hz) and accelerometer data (200Hz) at Z axis, and the calibrated accelerometer data (i.e., with hand vibration noise removal).

6 CROSS-DOMAIN USER AUTHENTICATION

6.1 Coarse-grained Synchronization

To examine the cross-domain similarity for user authentication, WearID needs to simultaneously capture a same voice command from a subject using the wearable's accelerometer and the VA device's microphone. This requires the system to trigger the data collection on both devices with a low relative time delay so as to record the same speech. We develop two alternative synchronization approaches: *WiFi Communication-based Method* and *Parallel Wake-word Detection Method*. WearID uses the WiFi Communication-based Method if the wearable is equipped with a WiFi module. The VA device detecting a wake word sends a triggering message through WiFi to notify the wearable device for collecting vibration signals. If a WiFi module is not equipped (e.g., activity tracker), the wearable device can receive the triggering message via the Bluetooth link with a paired smartphone that connects to the VA device. Figure 8(a) and (b) show the results of the WiFi communication-based synchronization between a VA device (i.e., Nexus 6) and a wearable (i.e., Huawei 2 sport) given a voice command "Alexa, What's on my calendar for tomorrow". We can find that the data of the microphone and the accelerometer are roughly synchronized.

As an alternative approach, the Parallel Wake-word Detection Method is used when the WiFi network delay is high and not suitable for synchronization. In such situations, the wearable device needs to recognize the wake word using its accelerator in parallel with the VA device and triggers the data collection. In particular, we build a machine learning model (e.g., SVM, random forest) based on the speech characteristics in the vibration domain for detecting wake words. Given that wearable's accelerometers usually run in the background for monitoring user's activities around-the-clock, this method would not introduce additional energy consumption on the wearable. Our study shows that using a random forest model can sufficiently recognize 10 wake words with 83% accuracy by using accelerometers in a Huawei 2 sport smartwatch.

6.2 Data Denoising and Segmentation

The accelerometer readings collected with wearables contain substantial noises caused by human motions (e.g., walking, hand tremor). These motions are unpredictable and can significantly distort the speech vibration patterns in accelerometer readings. Previous work [27, 43] found that human motion-related accelerations usually have frequencies lower than $20Hz$. Therefore, we adopt a high-pass Butterworth filter with a cut-off frequency of $20Hz$ to remove the impacts of human motions and reveal speech vibrations for cross-domain comparison. Figure 8(c) illustrates the accelerometer readings after our denoising. Compared with the raw accelerometer readings shown in Figure 8(b), the denoised accelerometer readings present more obvious patterns that are similar to the acoustic signal shown in Figure 8(a).

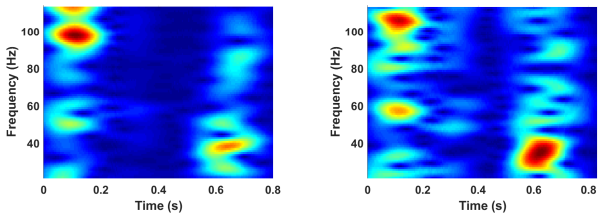
Next, we calculate the moving variance of the signals in the audio domain and determine the segment associated with human speeches based on an empirical threshold of 0.1, which sufficiently differentiates ambient noises and human speeches. Segmentation on accelerometer readings is particularly challenging due to its low sensitivity to aerial voice. Therefore, we use the segmentation results of the microphone recordings to assist the segmentation of the accelerometer readings. Since both data has been coarsely synchronized, we search for the starting point of voice segment on the accelerometer reading within a time window $W_T = 0.5$ after the onset of the microphone segment. The window is determined by an empirical study on the relative time offset between the onsets of microphone and accelerometer segments. We then determine the ending point of the segment in the accelerometer readings based on the length of the microphone segment.

6.3 Feature Extraction in Vibration and Audio Domains

Time-frequency Feature Extraction. In order to derive meaningful features for cross-domain comparison, we resort to time-frequency analysis which has shown great successes in both speech and speaker recognition tasks. Our preliminary study validates that solely relying on time-series correlation between the accelerometer and the microphone readings fails to effectively compare cross-domain similarity. We demonstrate the results of time-series comparison in Appendix A.1. To extract time-frequency features, we explore spectrograms that represent vibration/audio signals' energy distribution over a range of frequencies in short time frames. The spectrogram is derived by computing the Discrete Time Short Time Fourier Transform (DT-STFT) representations of the acoustic signals in vibration/audio domain with a sliding window, which is defined as following:

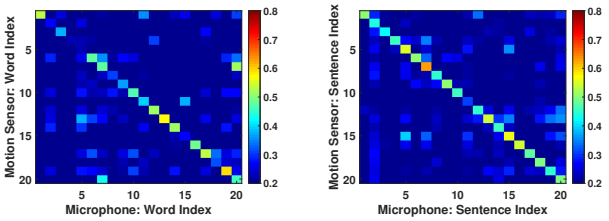
$$DTSTFT(t,f) = \sum_{n=t}^{t+N-1} x(n)w(n-t)e^{-jfn}, \tag{2}$$

where t and f are the time index and frequency index of the two-dimension spectrogram. $x(n)$ is a sample of the acoustic signal in the sliding window, and N is the size of the sliding window/FFT. We empirically determine N to be 2048 and 64 for microphone and accelerometer data, respectively. $w(n)$ is a Hamming window with length N. We then compute the magnitude squared of DT-STFT representations in at t: $P_t = [|DTSTFT(t,1)|^2, ..., |DTSTFT(t,F)|^2]$,



(a) Spectrogram of accelerometer (b) Aliasing spectrogram of microphone

Figure 9: Comparison of the accelerometer spectrogram (Z axis) with the converted microphone spectrogram ("Alexa").



(a) Voice sound of words (b) Voice sound of sentences

Figure 10: The spectrogram correlation based on our method.

Algorithm 1 Spectrogram-based Frequency Conversion Algorithm

function CONVERSION(S_{mic})
Input: S_{mic}-original microphone spectrogram
ω-sampling rate of accelerometer
Output: $\hat{S_{mic}}$-converted microphone spectrogram
$|\hat{S_{mic}} = zeros(T, F)|, \omega_{ws} = 2\pi \times \omega$
for $t = 1 : T$ **do**
for $f_mic = 700 : 3300$ **do**
// Frequency selection
for $N_{shift} = -10 : 10$ **do**
$f_w = |f_{mic} - N_{shift} \times \omega|$
if $|S_{mic}(t_n, f_m)| > 70dB$ & $f_w \leq f_s$ & $f_w > 0$ **then**
// Amplitude selection
$\hat{S_{mic}}(t_n, f_w) = \hat{S_{mic}}(t_n, f_w) + |S_{mic}(t_n, f_m)|$
// Spectrogram-based frequency conversion
end if
end for
end for
end for
end function

where $F = N/2$ following Nyquist theorem. Next, we slide the window by a step of p samples and repeat the same steps to derive the DT-STFT representation for each window. The time-frequency features S (i.e., spectrogram) are then obtained by combining the DT-STFT representations ordered in time: $S = [P_0, .., P_T]$.

Feature Domain Conversion. To mitigate the impacts of huge sampling rate gap between the microphone and the accelerometer, we develop a feature domain conversion method to transform the spectrograms in high-frequency audio domain to those in low-frequency vibration domain. The conversion method takes components in the audio domain spectrogram $S_{mic}(t, f_m)$ as input and calculates its new position (t, f_w) in the vibration domain. The original microphone frequency component f_m is then mapped to the low-frequency component f_w based on Equation 1, with the time index unchanged. If multiple spectrogram components are overlapped at the same point, we accumulate their energy in the new converted spectrogram. The conversion function is defined as:

$$\hat{S}_{mic}(t, f_w) = \sum_{n=-\inf}^{\inf} S_{mic}(t, win(|f_m + n \times \omega|)), \tag{3}$$

where ω is the sampling rate of the accelerometer. Such conversion maps each frequency component in the audio domain to an appropriate frequency in the vibration domain, which makes the cross-domain comparison possible.

Sensitive Feature Selection. In order to achieve reliable cross-domain comparison, we study the frequency selectivity differences of microphone and accelerometer and select the most sensitive time-frequency features across the two sensors. Compared to microphone, wearable has different sensitivities to human speeches across different frequency bands. We explore this phenomenon by recording a chip signal of $0Hz \sim 4kHz$ with a wearable (Huawei 2 sport) and a VA device (Nexus 6) and compare the similarity between the spectrograms of the accelerometer and the microphone readings. Note that we have applied the feature domain conversion method on the microphone's spectrogram. We find that the spectrograms of the accelerometer and the microphone only show high similarity within $700Hz \sim 3300$Hz, where the harmonics of human speeches reside. We illustrate the comparison results in Appendix Figure 3. To cope with the frequency selectivity differences, we only use the spectrogram of microphone from $700Hz$ to $3300Hz$ for generating the low-frequency aliasing spectrogram. Furthermore, since the wearable's accelerometer can only be triggered with sound waves over $70dB$ as shown in Section 5.2 (also confirmed in Accelword [52]), we exclude the frequency components of the microphone spectrogram with energy below $70dB$ for feature domain conversion. The algorithm integrating feature domain conversion and sensitive feature selection methods is presented in Algorithm 1. Figure 9(b) shows an aliasing spectrogram of a voice command "Alexa" derived from the microphone readings. We can observe that the aliasing spectrogram has an "equivalent" form with the spectrogram of accelerometer shown in Figure 9 (a).

6.4 User Authentication Using Cross-domain Similarity

Spectrogram Calibration based on 2D-normalization. The scales of measurements are greatly different on the accelerometer and the microphone. To resolve such scale differences, we develop a 2D-normalization scheme to normalize the energy values of the spectrograms across different frequencies. The normalization operation is defined as:

$$S_{norm}(t, f) = \frac{S(t, f) - S_{min}(t)}{S_{max}(t) - S_{min}(t)}, \tag{4}$$

where $S(t, f)$ is a spectrogram component at time t and frequency f. This normalization process is applied to spectrograms in both vibration and audio domains.

Cross-domain Comparison based on 2D-Serial Correlation. WearID authenticates users through comparing the 2D correlation between the spectrogram of the accelerometer and aliasing spectrogram of the microphone. We refer to the 2D correlation coefficient

Table 2: The specifications of the accelerometers in the tested wearable devices.

Model	Accelerometer	Programmable Measurement Range	Sensor sampling frequency	System sampling rate
LG W150	Invensense M6515	$\pm 2g, \pm 4g, \pm 8g, \pm 16g$	4-4000Hz	200Hz
Huawei watch 2 sport	STMicroelectronics LSM6DS3	$\pm 2g, \pm 4g, \pm 8g, \pm 16g$	4-1600Hz	100Hz

as cross-domain similarity which is defined as:

$$
\begin{aligned}
Corr(\hat{S}_{mic}, S_{acc}) &= \frac{A \times V}{\sqrt{A^2 \times V^2}}, \\
s.t., A &= \sum_t \sum_f (\hat{S}_{mic}(t,f) - \overline{\hat{S}_{mic}}), \\
V &= \sum_t \sum_f S_{acc}(t,f) - \overline{S_{acc}},
\end{aligned}
\tag{5}
$$

where $\overline{S}$ represents the mean of a spectrogram, either in the audio domain or in the vibration domain. S_{acc} represents the spectrogram of accelerometer. In practice, directly computing frame-wise correlation does not yield good similarity comparison performance due to the unpredictable offsets caused by coarsely synchronized data collection processes on the wearable and the VA device. Thus, we propose a 2D-serial correlation algorithm that searches for an optimal offset associating with the maximum correlation between the spectrograms in the vibration and the audio domains. Particularly, we fix the aliasing spectrogram in the audio domain and shift the spectrogram in the vibration domain frame by frame to calculate the correlation coefficient. The maximum 2D-correlation coefficient can then be found and used as the correlation score. Finally, a threshold-based method is applied to the correlation score and authenticate the user if the score is over an empirical threshold. Figure 10(a) shows the pairwise correlation scores of 20 spoken words provided in Table B. Most of the diagonal comparisons (i.e., same words) show the highest correlation scores. Figure 10(b) further confirms the effectiveness of our method on differentiating 20 representative voice commands shown in Table C, which shows better performance. This is reasonable since sentences contain much more speech information than single words.

7 PERFORMANCE EVALUATION

7.1 Experimental Methodology

Devices. To evaluate WearID, two smartwatch models, Huawei 2 sport ($100Hz$) and LG W150 ($200Hz$) are involved to collect accelerometer readings. The accelerometer specifications of the two smartwatches are listed in Table 2. Specifically, LG W150 is equipped with Invensense M6515 which supports sampling frequencies within $4Hz \sim 4000Hz$. The maximum acceleration that can be measured with this accelerometer is $\pm 16g$.Huawei watch 2 sport has the same programmable measurement range as the LG smartwatch, but it supports lower sampling frequencies, up to $1600Hz$. Although the accelerometers can record vibrations of $1.6KHz \sim 4KHz$, the vendors constrain the sampling rates to ensure low power consumption. Both smartwatches run Android Wear OS 2.0. In addition, as mentioned in Section 6.2, we use a high-pass filter of $20Hz$ to remove the impacts of body movements (e.g., typing on a keyboard, walking) on accelerometer readings. We use an Android smartphone (Motorola Nexus 6) to emulate the VA device recording voice commands at a sampling rate of $8kHz$.

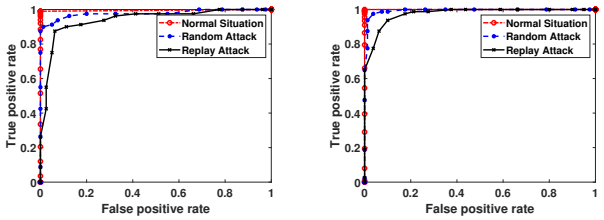
Experimental Setup. We evaluate the performance of WearID in a typical office environment. Compared with the home environment, the office environment has more dynamic ambient noises (e.g., air condition, people walking, and conversations). Each participant wears a smartwatch when he/she speaks voice commands to a Motorola Nexus 6 smartphone at $1m$ distance. The average SPL of the spoken speech commands is $80dB$ (i.e., typical presentation-level volume), which is reasonable as most users subconsciously increase their volume when issuing voice assistant commands, usually from a distance. Because people may wear watches differently (e.g., upside-down, loose around the wrist), we evaluate the impacts of different wearing positions on our system. Particularly, we test horizontal and vertical positions, which have the smallest and the largest impact angles between acoustic waves and smartwatches' screens, respectively. We use a Logitech S120 speaker [31] to conduct replay attacks and hidden voice commands, with the volume set to maximum. To imitate ultrasound attacks, we use a function generator (i.e., Keysight Technologies 33509B [41]) and a tweeter speaker [18] to generate ultrasound. The distance between the loudspeaker/tweeter speaker and the smartwatch is $30cm$.

Data Collection. We involve 10 participants to test WearID under the normal situation (i.e., no attack present) and various attacks over a six-month period. The participants are asked to speak 20 representative critical voice commands as listed in Appendix Table C. From each participant, 40 voice command samples with the smartwatch worn in horizontal and vertical positions are collected. In addition, we record 40 samples of ambient noises by using the smartwatch's accelerometer to examine WearID under the situation where the legitimate user is not issuing critical commands. Besides, 100 samples of 10 hidden voice commands are utilized to evaluate WearID against hidden voice command attacks [2].

Evaluation Metrics. To evaluate WearID, we use the following four metrics: *true positive rate (TPR)* is the percentage of critical commands of the legitimate user being correctly authenticated; *false positive rate (FPR)* is the percentage of the adversaries' critical voice commands that pass WearID; *receiver operating characteristics (ROC) curve* is generated by plotting the TPR against the FPR under thresholds from 0 to 1 with a step of 0.01; *Area under the ROC Curve (AUC)* measures how well the WearID correctly authenticating the legitimate users while rejecting the adversaries.

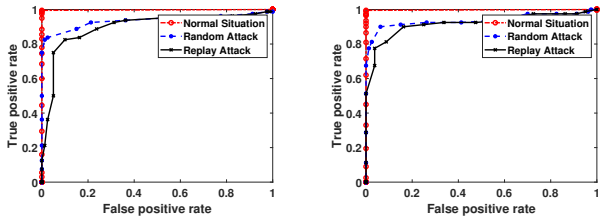
7.2 Authenticating Legitimate Users

We first evaluate WearID in normal situations, where the attacker does not present. While there is no malicious attack, WearID can still be mistakenly triggered by friendly users' (e.g., family members, colleagues) conversation that is similar to the critical commands, but the wearable device of the legitimate user only records ambient



(a) Horizontal (b) Vertical

Figure 11: Average ROC curve of verifying the user using Huawei watch 2 under normal situation, random attack and impersonate/replay attacks.



(a) Horizontal (b) Vertical

Figure 12: Average ROC curve of verifying the user using LG Urban W150 under normal situation, random attack and impersonate/replay attacks.

noises. For evaluation, we use the participants' critical commands recorded by the VA device against the corresponding wearables' vibration data to simulate the legitimate user using the VA device. Furthermore, we use the participants' critical commands recorded by the VA device against the wearables' ambient vibration data to simulate the cases that WearID is triggered mistakenly. We use these data to derive ROC curves and study the performance of WearID. The red curves in Figure 11 and Figure 12 present the ROC of authenticating the legitimate users when the users are wearing Huawei Watch 2 and LG W150 with two typical poses, respectively. We can observe that WearID achieves 99.8% TPR and 0% FPR on authenticating the legitimate users on Huawei Watch 2. Similarly, WearID can achieve 99.6% TPR and 0% FPR on LG W150. The 0% FPRs indicate that the voice commands from friendly users will not pass WearID, meaning that our system can be used in typical environments with multiple people. The AUCs of these cases are all around 100% no matter the smartwatch is worn horizontally or vertically, indicating that WearID can authenticate legitimate users' critical voice commands accurately and robustly with different wearables devices and their poses.

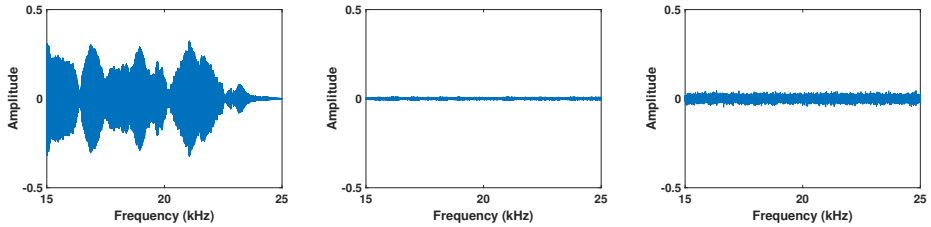
7.3 Attack on User's Absence

Against Random Attack. Under the random attack, an adversary tries to use his/her own voice to bypass the VA system. Although the user is not co-located with the VA device, when the WearID is triggered by the adversary, the user's smartwatch may still record the user's speeches (e.g., conversation). To evaluate WearID under such random attacks, we take turns considering each participant as the legitimate user and the remaining 9 participants as adversaries. We use the adversaries' critical command speeches recorded by the VA device against the vibration data of the legitimate user's voice commands for evaluating random attacks. In addition, we use the legitimate user's audio critical commands against his/her vibration data to simulate the legitimate use of the VA device. Figure 11 and Figure 12 show the average ROC curves of authenticating the legitimate user with two smartwatches under horizontal and vertical poses. We observe that WearID can authenticate the legitimate user and reject random attacks with high accuracy for both poses. In particular, the AUCs for Huawei watch 2 and LG Urbane W150 are 94.5% and 88.9% under the two poses. The vertical position shows slightly higher AUCs as it has the largest impact angles between acoustic waves and smartwatches' screens. Given an FPR of 5%, WearID can achieve high TPRs of 95.2% and 98.5% for Huawei watch 2 held in both positions. The results indicate that WearID is effective in defending against random attacks.

Against Impersonation and Replay/Synthesis Attack. Next, we evaluate WearID under more sophisticated impersonation and replay/synthesis attacks which reproduce a user's voice characteristics on the VA device. In this case, the wearable is attached to the absent user and out of the adversary's control, and it seldom happens when the two separated devices (i.e., VA device and the smartwatch) receive the exact same speech. However, the smartwatch may still record the user's speeches but with other content. For evaluation, we alternatively set each participant as the legitimate user. We use each legitimate user's critical voice command recorded by two smartwatches against other 19 audio recordings. To simulate legitimate use of critical voice commands, we use the legitimate user's vibration data of each command against the corresponding audio data. Figure 11 and Figure 12 show the ROC curves (i.e., black curves) when authenticating the user under impersonation/replay attacks. We find that WearID successfully reject the adversaries by using both Huawei watch 2 and LG Urbane W150 under both horizontal and vertical poses. In particular, WearID achieves 89.1% and 86.8% for Huawei Watch 2 and LG Urbane W150 under horizontal position. The AUCs are 91.23% and 88.34% under vertical pose. For a FPR of 10%, WearID can obtain the TPRs of 91.2% and 93.3% when Huawei watch 2 is held in horizontal and vertical directions. We find the performance of WearID under impersonation/replay attacks are slightly lower than that under random attacks. This is because the adversary has obtained the user's voice samples to improve the attack. While in the practical scenarios, a legitimate user's wearable device does not usually record the user's speeches, which make the performance of WearID approaching to that under the normal situation.

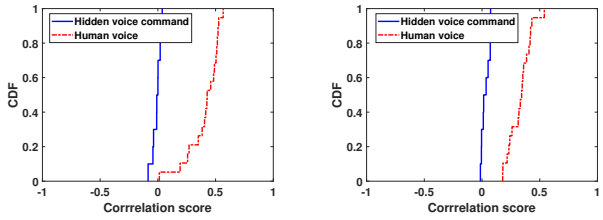
7.4 User Authentication under Co-location Attack

Against Hidden Voice Command. To evaluated WearID under hidden voice command attacks, we compare the vibration data and the audio data for each of the 100 recorded hidden commands. In addition, we alternatively set each of the 10 participants as the legitimate user and compare the vibration and the audio data of each critical voice command. Figure 14 depicts the CDFs of the cross-domain similarities of the hidden commands recorded by



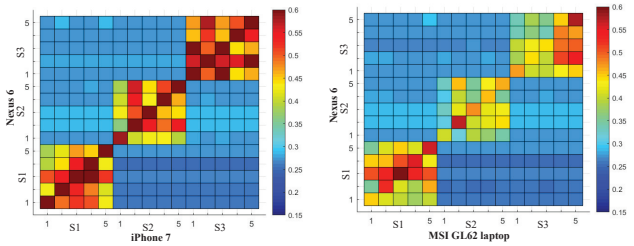
(a) Microphone (b) Huawei watch 2 (c) LG Urban W150

Figure 13: The frequency responses of the VA system and the wearables (i.e., microphone, Huawei watch 2, LG Urban W150 from left to right) under ultrasound attacks.



(a) Huawei watch 2 (b) LG Urban W150

Figure 14: CDF of the cross-domain 2D correlations to distinguish the hidden voice commands and the legitimate user's voice commands.



(a) Nexus6 vs. iPhone7 (b) Nexus6 vs. Laptop

Figure 15: Normalized 2D cross-correlation between spectrogram of different recording devices .

the VA device and the legitimate user's critical voice command captured by the smartwatches' accelerometers. We observe that the similarities between the two sensor readings are low for the hidden voice commands, which can be differentiated well from the legitimate user's critical commands. In particular, the median of the cross-domain similarities for the hidden voice commands is around 0 for Huawei watch 2 and 0.05 for LG Urban W150. In comparison, the median similarities for the legitimate user's voice commands are around 0.5 for Huawei watch 2 and 0.4 for LG Urban W150. This is because the accelerometers on the wearables have short response distances (i.e., less than $25cm$) and unique frequency selectivity patterns to sound. Thus, with our cross-domain user authentication approach, the hidden voice attacks can be defended.

Against Ultrasound Attack. Under the ultrasound attack, an adversary modulates the recorded user voice command to an inaudible frequency and replays it using an ultrasound speaker. In this scenario, both the VA's microphone and the user's smartwatch are exposed to this inaudible sound. We evaluate WearID by comparing the accelerometer's and the smartwatches' responses under a nearly inaudible chirp signal. In particular, we use a function generator (i.e., Keysight Technologies 33509B [41]) to generate a chirp of $15kHz \sim 25kHz$ and play the chirp using a tweeter speaker, which is placed $30cm$ away from the smartwatch. Figure 13 shows the frequency responses of the microphone and the two accelerometers. We can find that though the microphone show responses from $15kHz \sim 24kHz$, we do not observe any responses on the two smartwatches. The experimental results show that the smartwatch's accelerometer could shield the VA system from ultrasound attacks.

7.5 Scalability to Different VA Devices

To demonstrate the scalability of WearID to various VA devices (e.g., phones, laptops), we compare the voice signals recorded by Nexus 6 smartphone used to test WearID with the sound recorded with two other devices, a iPhone 7 and a MSI GL62 laptop. Since microphones in different VA devices share similar hardware components (e.g., membrane, black-plate) and audio processing pipelines, a voice command recorded by different VA devices should exhibit high similarity. In the experiment, we place the smartphones and the laptop 1m away from a subject and set the sampling rates to $8kHz$. Then we use the three devices to simultaneously record 5 trials of each of the three spoken voice commands: S1-"What's on my calendar for tomorrow?"; S2-"What is my password?"; S3-"Delete all my reminders". To quantify the similarity between voice commands, we derive a spectrogram of each recorded voice signal and then calculate the normalized 2-D cross-correlation between voice commands recorded with each pair of devices. As shown in Figure 15 (a) and Figure 15 (b), high correlation scores between the same recordings on the two pairs of different devices can be observed. These results validate that WearID can be easily extended to various VA devices with different audio recording capabilities.

8 DISCUSSION

Deployment Feasibility. WearID requires a minimum sampling rate of 100Hz for the accelerometer to capture aerial speech vibrations. This sampling rate is commonly available in the mainstream wearable devices, such as Samsung Gear Series and Fitbit. A user could pair/enroll a wearable device to an account of a VA system (e.g., Google or Alexa account), allowing the user to use critical

commands on any VA devices linked to his account. For wearable devices without WiFi/cellular modules (e.g., some activity trackers), WearID will still work by using Bluetooth to bridge the wearable devices to the VA's clouds using the paired smartphones. A user can use WearID in typical room environments without requiring the wearable and the VA device (e.g., Google Home, Amazon Echo) being close to each other. The user's voice could easily reach the wearable worn by the user and the VA device within an effective range of approximately 7 meters. WearID is especially useful in the scenarios where multiple users share the VA devices (e.g., business office and home). With the cross-domain authentication, WearID could detect unauthorized critical commands and alert the corresponding user.

Energy Consumption and Delay. WearID offloads the computationally expensive tasks (i.e., Cross-domain Voice Comparison) to the cloud, avoiding the heavy computation/energy consumption on the wearable device. Therefore, the most power-consuming task on the wearable device is data acquisition, which uses the built-in accelerometer to capture users' voice commands. We find that voice commands usually last less than 10 seconds, and the corresponding power consumption of recording the voice commands by using the built-in accelerometer on a wearable device is lower than $0.21J$. We also notice that since traditional VA systems still need to send recorded voice data to the cloud for data processing, the wearable device could send its data to the cloud at the same time. Thus, the delay of WearID is close to the response time of traditional VA systems (e.g., 1.93 seconds on average for Alexa [7]).

Replay Attack in Vibration Domain. Considering WearID exploits vibration signals for cross-domain authentication, an adversary may attack WearID via replaying well-designed audio that generates vibration signals replicating the replayed audio. Particularly, to design such an audio signal, the adversary can study time-frequency response of the same model of wearable device used by the legitimate user. However, due to the unique manufacture imperfections, each wearable device exhibits distinctive frequency-selective patterns, even for the same type of device, making it difficult to replicate the vibration signals. Additionally, the adversary needs to get very close to the wearable device (i.e., less than 30cm) to generate the vibration signals, which will be noticed by the user. We leave the study on exploring the frequency-selective pattern to defend against replay attacks in the vibration domain to our future work.

9 CONCLUSION

In this paper, we presented WearID, a wearable-assisted low-effort user authentication system that assists existing Voice Assistant (VA) systems with enhanced security, especially the critical voice commands (e.g., big purchases, critical calls). WearID authenticates the user via examining the cross-domain similarity between the unique voice characteristics captured by the accelerometers of the wearable device and the microphone of the VA system, respectively. The cross-domain comparison enables WearID to achieve training-free and privacy-preserving voice authentication. We developed the spectrogram-based conversion and frequency/amplitude selection algorithms, which model the unique and complex relationships between the voice commands across two domains under a huge sampling rate gap. By utilizing the cross-domain similarity along with the motion sensor's short response distance to voice, WearID can shield the VA system from various acoustic attacks (e.g., impersonation, replay, hidden command, and ultrasound attacks). Extensive experiments with two commodity smartwatches and 1000 voice commands showed that WearID can authenticate users' voice commands with 99.8% accuracy in the normal situation and detect 97.2% fake voice commands under audible/inaudible attacks.

10 ACKNOWLEDGMENT

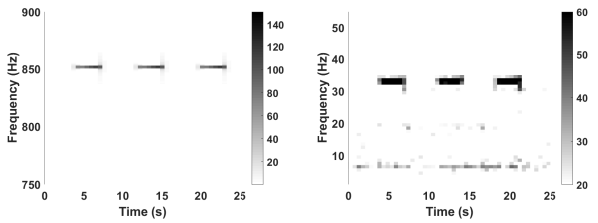
This work was partially supported by the National Science Foundation Grants CCF-1909963, CCF-2000480, CCF-2028876, CNS1526524, CNS1547350, CNS1714807, CNS1814590 and Army the Research Office Grant W911NF-18-1-0221.

REFERENCES

[1] 2015. Wearable ID: Is it a fit for your campus? https://www.cr80news.com/news-item/wearable-id-is-it-a-fit-for-your-campus/.
[2] 2016. Hidden Voice Commands Example. http://www.hiddenvoicecommands.com/white-box.
[3] Amazon. 2020. Alexa Uses Voice Profiles to Recognize Your Voice and Personalize Your Experience. https://www.amazon.com/gp/help/customer/display.html?nodeId=202199440.
[4] S Abhishek Anand and Nitesh Saxena. 2011. Speechless: Analyzing the Threat to Speech Privacy from Smartphone Motion Sensors. (2011).
[5] S Abhishek Anand, Chen Wang, Jian Liu, Nitesh Saxena, and Yingying Chen. 2019. Spearphone: A speech privacy exploit via accelerometer-sensed reverberations from smartphone loudspeakers. *arXiv preprint arXiv:1907.05972* (2019).
[6] Les Atlas and Shihab A Shamma. 2003. Joint acoustic and modulation frequency. *EURASIP Journal on Applied Signal Processing* 2003 (2003), 668–675.
[7] Anna Attkisson. 2016. Siri vs. Alexa: Why Amazon Won Our 300-Question Showdown. https://www.tomsguide.com/us/siri-vs-alexa,review-3681.html.
[8] Android Authority. 2020. Google Home and Assistant commands – here's the ones you need to know. https://www.androidauthority.com/google-assistant-commands-727911/.
[9] JenniferE Bellemare. 2018. Consumers Need Answers to Amazon Echo Privacy Concerns. https://www.identityforce.com/blog/amazon-echo-privacy-concerns.
[10] Logan Blue, Hadi Abdullah, Luis Vargas, and Patrick Traynor. 2018. 2MA: Verifying Voice Commands via Two Microphone Authentication. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*. ACM, 89–100.
[11] Joseph P Campbell. 1997. Speaker recognition: A tutorial. *Proc. IEEE* 85, 9 (1997), 1437–1462.
[12] 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.
[13] Nicholas Carlini and David Wagner. 2018. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, 1–7.
[14] Si Chen, Kui Ren, Sixu Piao, Cong Wang, Qian Wang, Jian Weng, Lu Su, and Aziz Mohaisen. 2017. You can hear but you cannot steal: Defending against voice impersonation attacks on smartphones. In *Distributed Computing Systems (ICDCS), 2017 IEEE 37th International Conference on*. IEEE, 183–195.
[15] Geumhwan Cho, Jusop Choi, Hyoungshick Kim, Sangwon Hyun, and Jungwoo Ryoo. 2018. Threat modeling and analysis of voice assistant applications. In *International Workshop on Information Security Applications*. Springer, 197–209.
[16] Kirsten Crager, Anindya Maiti, Murtuza Jadliwala, and Jibo He. 2017. Information leakage through mobile motion sensors: User awareness and concerns. In *Proceedings of the European Workshop on Usable Security (EuroUSEC)*.
[17] Phillip L De Leon, Michael Pucher, and Junichi Yamagishi. 2012. Evaluation of the vulnerability of speaker verification to synthetic speech. *IEEE Transactions on Audio, Speech, and Language Processing* 20 (2012), 2280 – 2290.
[18] Pyramid Electronics. 2018. Pyramid Car Audio, 300 Watt Aluminum Bullet Horn in Enclosure with Swivel Housing. http://www.pyramidcaraudio.com/sku/TW28/300-Watt-Aluminum-Bullet-Horn-in-Enclosure-wSwivel-Housing.
[19] Adrienne Porter Felt, Elizabeth Ha, Serge Egelman, Ariel Haney, Erika Chin, and David Wagner. 2012. Android permissions: User attention, comprehension, and behavior. In *Proceedings of the eighth symposium on usable privacy and security*. ACM, 3.
[20] 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.

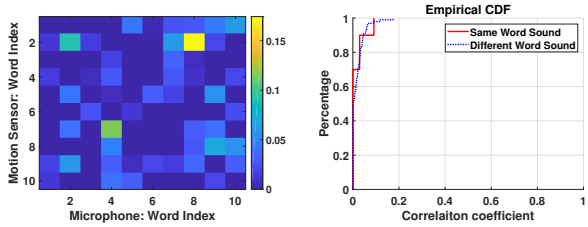
[21] Google. 2019. How you sign in with 2-Step Verification. https://support.google.com/accounts/answer/1085463?hl=en.
[22] Google. 2020. Voice Match and media on Google Nest and Google Home speakers and displays. https://support.google.com/googlenest/answer/7342711?hl=en.
[23] Tzipora Halevi, Di Ma, Nitesh Saxena, and Tuo Xiang. 2012. Secure proximity detection for NFC devices based on ambient sensor data. In *European Symposium on Research in Computer Security*. 379–396.
[24] Jun Han, Albert Jin Chung, and Patrick Tague. 2017. Pitchln: eavesdropping via intelligible speech reconstruction using non-acoustic sensor fusion. In *Proceedings of the 16th ACM/IEEE International Conference on Information Processing in Sensor Networks*. 181–192.
[25] Matthieu Hébert. 2008. Text-dependent speaker recognition. In *Springer handbook of speech processing*. Springer, 743–762.
[26] Apple IOS. 2019. Siri. https://www.apple.com/ios/siri/.
[27] Dean M Karantonis, Michael R Narayanan, Merryn Mathie, Nigel H Lovell, and Branko G Celler. 2006. Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring. *IEEE transactions on information technology in biomedicine* 10, 1 (2006), 156–167.
[28] Tomi Kinnunen, Bingjun Zhang, Jia Zhu, and Ye Wang. 2007. Speaker verification with adaptive spectral subband centroids. In *International Conference on Biometrics*. Springer, 58–66.
[29] John Krumm and Ken Hinckley. 2004. The nearme wireless proximity server. In *International Conference on Ubiquitous Computing*. 283–300.
[30] Johan Lindberg and Mats Blomberg. 1999. Vulnerability in speaker verification-a study of technical impostor techniques. In *Sixth European Conference on Speech Communication and Technology*.
[31] Logitech. 2018. Logitech S120 speaker. https://www.logitech.com/en-us/product/s120-stereo-speakers.
[32] Yan Michalevsky, Dan Boneh, and Gabi Nakibly. 2014. Gyrophone: Recognizing Speech from Gyroscope Signals. In *USENIX Security Symposium*. 1053–1067.
[33] K Sri Rama Murty and Bayya Yegnanarayana. 2006. Combining evidence from residual phase and MFCC features for speaker recognition. *IEEE signal processing letters* 13, 1 (2006), 52–55.
[34] Murray Newlands. 2017. THE TOP WEARABLE PAYMENT TECHNOLOGY. https://due.com/blog/wearable-payment-technology/.
[35] Yao Qin, Nicholas Carlini, Ian Goodfellow, Garrison Cottrell, and Colin Raffel. 2019. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. *arXiv preprint arXiv:1903.10346* (2019).
[36] Douglas A Reynolds and Richard C Rose. 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE transactions on speech and audio processing* 3, 1 (1995), 72–83.
[37] Duo Security. 2019. Secure Authentication With the Duo Mobile App. https://duo.com/product/multi-factor-authentication-mfa/duo-mobile-app.
[38] Dave Singelee and Bart Preneel. 2005. Location verification using secure distance bounding protocols. In *IEEE International Conference on Mobile Adhoc and Sensor Systems Conference*. 7–pp.
[39] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur. 2017. Deep Neural Network Embeddings for Text-Independent Speaker Verification. In *Interspeech*. 999–1003.
[40] Statista. 2018. Number of connected wearable devices worldwide from 2016 to 2021. https://www.statista.com/statistics/487291/global-connected-wearable-devices/.
[41] Keysight Technologies. 2018. Keysight Technologies 33509B. https://www.alliedelec.com/keysight-technologies-33509b.
[42] Roberto Togneri and Daniel Pullella. 2011. An overview of speaker identification: Accuracy and robustness issues. *IEEE circuits and systems magazine* 11, 2 (2011), 23–61.
[43] Niall Twomey, Tom Diethe, Xenofon Fafoutis, Atis Elsts, Ryan McConville, Peter Flach, and Ian Craddock. 2018. A comprehensive study of activity recognition using accelerometers. In *Informatics*, Vol. 5. Multidisciplinary Digital Publishing Institute, 27.
[44] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez. 2014. Deep neural networks for small footprint text-dependent speaker verification. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4052–4056.
[45] Kathryn Vasel. 2015. How your voice can protect you from credit card fraud. https://money.cnn.com/2015/11/02/pf/voice-biometrics-customer-fraud/index.html.
[46] Chen Wang, S Abhishek Anand, Jian Liu, Payton Walker, Yingying Chen, and Nitesh Saxena. 2019. Defeating hidden audio channel attacks on voice assistants via audio-induced surface vibrations. In *Proceedings of the 35th Annual Computer Security Applications Conference*. 42–56.
[47] Xiaohui Wang, Yanjing Wu, and Wenyuan Xu. 2016. WindCompass: Determine Wind Direction Using Smartphones. In *Sensing, Communication, and Networking (SECON), 2016 13th Annual IEEE International Conference on*. IEEE, 1–9.
[48] WeChat. 2017. Voiceprint. https://thenextweb.com/apps/2015/03/25/wechat-on-ios-now-lets-you-log-in-using-just-your-voice/.
[49] Xuejing Yuan, Yuxuan Chen, Yue Zhao, Yunhui Long, Xiaokang Liu, Kai Chen, Shengzhi Zhang, Heqing Huang, XiaoFeng Wang, and Carl A Gunter. 2018. Commandersong: A systematic approach for practical adversarial voice recognition. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*. 49–64.
[50] Hossein Zeinali, Lukáš Burget, Jan Černockỳ, et al. 2019. A Multi Purpose and Large Scale Speech Corpus in Persian and English for Speaker and Speech Recognition: the DeepMine Database. *arXiv preprint arXiv:1912.03627* (2019).
[51] 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.
[52] Li Zhang, Parth H Pathak, Muchen Wu, Yixin Zhao, and Prasant Mohapatra. 2015. Accelword: Energy efficient hotword detection through accelerometer. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 301–315.
[53] Linghan Zhang, Sheng Tan, and Jie Yang. 2017. Hearing Your Voice is Not Enough: An Articulatory Gesture Based Liveness Detection for Voice Authentication. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 57–71.
[54] Linghan Zhang, Sheng Tan, Jie Yang, and Yingying Chen. 2016. Voicelive: A phoneme localization based liveness detection for voice authentication on smartphones. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 1080–1091.

A APPENDIX



(a) Spectrogram of microphone (b) Spectrogram of wearable accelerometer

Figure 1: Spectrogram of the single frequency signal (850Hz) on microphone and wearable device (i.e., Huawei watch 2 sport).



(a) Correlation matrix (b) CDF of the correlation

Figure 2: The time-domain correlation between the microphone data and motion sensor, which are resampled to the same sampling rate level (Illustrated with 10 words on Amazon Echo and Huawei watch 2).

A.1 Difficulty of Comparing Microphone Data with Motion Sensor Data

Figure 2 illustrates the difficulty of comparing the microphone data with the motion sensor data, where a participant speaks ten

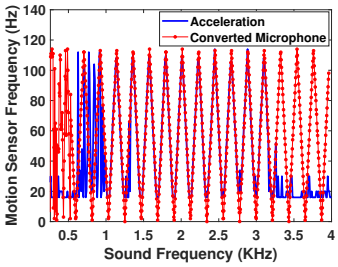


Figure 3: Directly converting the microphone data of a frequency chirp ($0 \sim 4KHz$) into the low frequency data.

words to both a microphone and a accelerometer, and both data are re-sampled to the same sampling rate for similarity comparison. Particularly, Figure 2 (a) shows the time-domain correlation coefficient between the microphone recorded sound (i.e., X axis) and motion sensor data (i.e., Y axis) by cross-comparing ten words. We observe that the correlations at the diagonal (i.e., same word sound) and non-diagonal (i.e., different word sounds) are indistinguishable. The results indicate that the re-sampling technique and the time-domain analysis are insufficient to address the similarity comparison of the two different sensing modalities. Figure 2(b), CDF of the correlation coefficients, further depicts the challenge of matching the sound across the two domains, where the sound of the same word and those of different words all show low correlation values (i.e., less than 0.1). Thus, we need to investigate the inherent unique relationship between the two sensing modalities to facilitate their similarity comparison.

A.2 Examples of the Voice Commands

We evaluate WearID with 20 representative voice commands that involve highly sensitive information or functionalities. The voice commands could be used by an adversary to access sensitive information or functionalities. Particularly, the adversary could acquire private information (e.g., schedule, password, email, contact list) of users. With these voice commands, they may also conduct unauthorized purchases or manipulate smart home devices. Note that we use voice commands of different lengths since to examine the generality of WeaID.

Table B: Representative words in voice commands.

ID	Word	ID	Word	ID	Word	ID	Word
1	Tomorrow	6	Good morning	11	Events	16	Team information
2	Answer	7	Request	12	Remember	17	Shopping list
3	Weather	8	Country music	13	Password	18	Living room camera
4	Instrument	9	Spotify	14	Flight	19	Weekend forecast
5	Information	10	Next Appointment	15	New York	20	Flash Briefing

Table C: Example of privacy leakages from voice assistant systems.

Security issues	Category	Voice Command Examples	Sentence Length
Potential privacy leakage	**Event schedule**	"What's on my calendar for tomorrow"	6
		"Where is my next appointment"	5
		"List all events for January 1st"	6
		"How much is a round-trip flight to New York"	9
	Reminder	"Remember that my password is 'money'"	6
		"What is my password"	4
		"Add 'go to the grocery store' to my to-do list"	10
	Shopping account information	"What's on my shopping list"	5
		"Track my order"	3
	Contact	"Read me my email"	4
		"Call my mother"	3
Unauthorized operation	**Neighborhood location**	"Find me a Italian near my home"	7
		"What is the traffic to my home"	7
	Unauthorized purchase	"Add paper towels to my cart"	6
		"Order all items in my cart"	6
	Voice assistant	"Answer the call"	3
		"Delete all my reminders"	4
		"Play my favorite music on Spotify"	6
	Access smart home devices	"Show the living room camera"	5
		"Clear all Bluetooth devices"	4