Authentication of Voice Commands by Leveraging Vibrations in Wearables

Cong Shi | Rutgers University
Yan Wang | Temple University
Yingying (Jennifer) Chen | Rutgers University
Nitesh Saxena | University of Alabama at Birmingham

Voice assistant systems are convenient, but attackers can mimic users' voices to access them and steal private information. We develop an authentication system that defends against acoustic attacks by using unique characteristics captured by the accelerometers in wearable devices.

Advanced speech recognition technologies have enabled intelligent voice assistant (VA) systems (e.g., Google Home and Amazon Alexa) in our lives. Users can speak naturally to VAs and have them perform simple tasks, such as playing music, managing calendar events, shopping online, and controlling smart appliances. More people are beginning to use VAs to complete sensitive tasks, such as unlocking the front door of a house or calling a bank to conduct transactions. While enjoying the convenience, users may not realize that their conversations with VAs often contain critical personal information (e.g., credit card numbers, passwords, and payments), which has driven adversaries to attack the systems, threatening people's privacy and property. For example, an adversary could get a user's credentials to access personal devices[1] by asking Google's VA, "OK, Google, what is my password?" A bad actor could also use a VA to make a significant purchase[2] by telling Amazon's system "Alexa, order a MacBook from Prime Now." Recently, adversaries learned to hack low-cost smart appliances (e.g., smart TVs) and use them to give voice commands to security-critical VAs[1] to, for instance, disarm smart locks.

Moreover, VAs have acquired abilities that are suitable for the workplace. However, their open nature makes sensitive business information vulnerable to hackers, who could obtain, e.g., meeting schedules and employee contact information by asking for it. This is more dangerous when VAs are deployed in high-security environments (e.g., nuclear power stations, stock exchanges, and data centers), where all voice commands are critical and must be authenticated around the clock.

Vulnerability of Voice Assistance

To ensure the successful large-scale deployment of VA systems, we need to address their inherent vulnerabilities and make them trustworthy to users. In this work, we consider an adversary to be a malicious user who aims to obtain sensitive information or undertake unpermitted actions through voice commands to VAs. We assume that an adversary cannot physically break a VA, take control of the system's cloud service, or get possession of a

Digital Object Identifier 10.1109/MSEC.2021.3077205
Date of current version: 24 May 2021

user's wearable device. Depending on whether a malicious party launches attacks in users' presence, we summarize potential attacks according to two categories.

Attacks in Users' Absence

When a user is away from a VA, an attacker can approach the device to launch the following:

1. *Random attacks*: In these, an attacker does not know what a user's voice sounds like. Therefore, he or she can try only to fool a VA with his or her own voice, giving commands to provide sensitive information, such as credentials and personal schedules. While this sounds naive, it succeeds around 3.5% of the time, due to the imperfect voice verification mechanisms in current VAs.[3]
2. *Impersonation attacks*: When an attacker has heard a user's voice, he or she can try to mimic it, e.g., in the pitch and tone. A bad actor can also use speech synthesis techniques to deliver commands in what sounds like a user's voice.
3. *Replay attacks*: An attacker inconspicuously records voice commands when a user interacts with a VA, later replaying them to trick the system. These attacks are easy to launch since smartphones can record audio without attracting any attention. Replay attacks have drawn great public awareness because they can spoof most current voice verification mechanisms with high success rates.

Attacks in Users' Presence

When a user is close to a VA, an attacker can take no overt action. However, a bad actor can attack the system by using imperceptible and inaudible commands, as in the following:

1. *Hidden voice command attacks*:[4] A malicious party can attack a VA by using obfuscated voice commands (e.g., voice commands that are similar to ambient noises), which the system can understand but the user cannot. These are referred to as *hidden voice commands*. They are usually realized by converting voice commands into sound signals that are meaningful to the speech recognition models used by VA systems. An attacker can even embed voice commands into background music and video stream audio channels to remotely target VAs through home and auto media systems. Research[5] has demonstrated that VAs are vulnerable to these assaults. Users' private information could leak (e.g., posting locations on social media), people could experience denials of service (e.g., by hackers activating airplane mode), and attackers could perform unauthorized operations (e.g., visiting malicious websites).
2. *Ultrasound attacks*:[6] Bad actors can launch completely inaudible attacks by mapping users' voice commands into ultrasound frequency bands that human ears cannot hear (i.e., higher than 20 kHz). Due to their nonlinearity, VA microphones can render commands in ultrasound frequency bands into normal frequency bands, enabling hackers to fool systems without being noticed. Since it is hard to detect ultrasound attacks, they are perfect for spying (e.g., making a VA initiate outgoing phone calls) and injecting false information (e.g., publishing fake online posts).

Because all these attacks pose severe security and privacy concerns, a practical scheme is highly desired for verifying voice commands' authenticity and enabling users to freely access VAs. Traditional voice authentication methods mainly rely on acoustic features (e.g., voice timbre and vocal tract resonances) extracted from microphone data to identify users.[7–8] Given their dependence on microphones, they, too, are vulnerable to acoustic attacks. Recent research has explored liveness detection techniques, such as detecting dynamic acoustic characteristics of human voices, to defend against acoustic attacks.[9–10] However, these solutions are designed for smartphones, whose microphone must be placed close to a user's mouth. Thus, they are not applicable to VAs that take commands from a distance. To add another layer of defense, some VAs exploit a second factor to secure voice commands,[11] such as asking challenge questions via text messages, phone calls, and virtual buttons on a mobile device.[12] This requires significant effort since users must confirm every voice command they give. Second factors could be undermined by users who are careless about their confirmations and might accept attempted attacks without paying attention.

Our Approach

In this work, we design and develop a user authentication framework for VA systems. In particular, we exploit a user's wearable device to capture unique human voice characteristics in the vibration domain and harness them to verify commands when a VA is triggered. The major advantage of our approach is that it does not require extra user effort (e.g., answering challenge questions and replying to messages/calls) and additional training with privacy-sensitive voice samples. In addition, our solution works on commercial off-the-shelf wearable devices that have been widely accepted, making it scalable under real-world scenarios without requiring specialized infrastructure.

To address the security vulnerabilities of VA systems, we propose a voice authentication system, WearID, that leverages speech similarity between the vibration domain and the audio domain to provide enhanced security to the ever-growing deployment of voice command devices. The insight is that when a user gives a command to a VA, his or her voice creates similar characteristics in both aerial speech vibration and audio. By leveraging wearable devices as a personal identity token, our solution captures users' voice characteristics in the aerial speech vibration through a wearable's accelerometer and compares them with the voice characteristics in the audio speech captured by a VA device's microphone. When a legitimate user gives a command, the similarity between the voice characteristics obtained from the vibration domain and the audio domain should have high similarity. Otherwise, the command is from an adversary.

The idea of WearID is illustrated in Figure 1. Upon detecting a wake word, WearID exploits a wearable's accelerometer and a VA device's microphone to simultaneously capture commands in the vibration domain and the audio domain, respectively. The accelerometer data can be sent to the VA system's cloud along with the recorded voice commands for user authentication. To realize the 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 and audio domains to verify a voice command. Our system involves minimum hardware modification. To enable WearID in practice, a customized plug-in of a VA system's app, such as "skills" on Amazon Alexa, helps to establish an agreement between a wearable device and a VA system. By accepting the agreement in the VA's app, a user associates his wearable device with the VA system. After the initial association, the user's wearable and the VA are connected to a cloud server. They simultaneously record a voice command upon detecting a wake word and perform user authentication in the cloud.

Wearable-Assisted User Authentication

Advantages of Using Wearables

Using wearable devices in our solution has natural causes. The number of wearable device users has reached half a billion worldwide,[13] and the total continues to quickly grow. Since wearable devices are rarely left unattended, it is natural for us to leverage them as trusted tokens to facilitate voice command authentication. In fact, wearable devices have been widely used as personal tokens in many applications. For example, smartwatches and smart wristbands have been explored as replacements for student ID cards because they are hard to forget to carry.[14] As another example, smartwatches have been accepted as a convenient and valid security token for contactless payment, such as through Google Pay and Apple Wallet. Furthermore, most commercial wearable devices have integrated various types of sensors (e.g., accelerometers, gyroscopes, and microphones) that can facilitate many mobile sensing applications. By

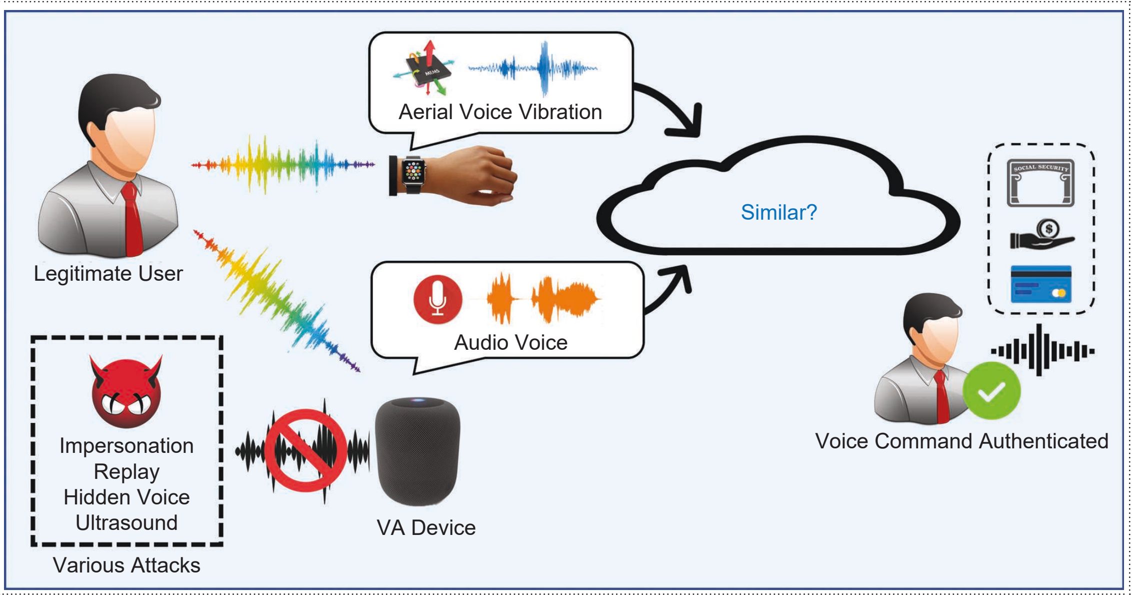


Figure 1. The use of human voice vibration domain representation to defend against audio-based VA attacks.

using wearable devices, our solution avoids additional user expense.

In particular, we propose to utilize accelerometers in wearable devices to capture users' voice commands for authentication. Instead of employing a second microphone, we choose to use the accelerometer, as it is sensitive only to sounds within a short distance. Such a short sensing distance can ensure that the accelerometer captures only the voice of the wearable user. It also shields VAs from various acoustic attacks, which are normally launched from afar to avoid being noticed. Furthermore, compared to a second microphone, our design reveals more inherent acoustic characteristics of a voice command, making speech in the vibration domain difficult to forge. As a result, our solution is invulnerable to acoustic attacks, including audible and inaudible attacks.

System Challenges

Recent research studies have shown that it is possible to use accelerometers on smartphones to capture speech signals.[12] However, harnessing weak aerial speech vibrations captured with wearables' accelerometers to authenticate users' voice commands remains challenging. We summarize the major challenges to implement our system as follows:

- *Weak response to human speech*: Because wearables' accelerometers are designed to measure movement, they have low sensitivity to aerial vibration caused by speech. In addition, people's movements during daily activities, such as walking and raising their arms, can create unpredictable accelerations, which are considered to be noise. Noisy readings make it difficult to detect and segment speech in accelerometer data.
- *Complex relationship between audio and vibration domains*: Due to different wearables' microphone and accelerometer designs, a voice recorded by the microphone and the aerial vibration captured by the accelerometer present distinct patterns. Therefore, it is impractical to directly compare commands recorded by the accelerometer and the microphone. Moreover, the microphones' minimum sampling rate is usually 8,000 Hz, while the accelerometers' maximum sampling rate is generally 200 Hz. Such a huge difference makes any direct comparison between vibration signals and audio signals impossible.
- *Unsynchronized sensing devices*: Since we use a microphone and an accelerometer from separate hardware (i.e., a VA and a wearable) to capture a voice command, it is hard to guarantee that the devices can trigger the data collection at the same time. Without an appropriate signal alignment, the acoustic and vibration data of the same voice command cannot have a one-to-one comparison, leading to false authentication. While it is possible to use a local area network to synchronize the data collection by triggering the devices with one message, doing so can introduce unpredictable delays, as local area networks are prone to network congestion.

System Overview

We realize our wearable-assisted user authentication system with three components: voice data synchronization, acoustic and vibration data processing, and voice characteristic comparison. The flow of our system is provided in Figure 2. Upon detecting a wake word, WearID performs synchronization to simultaneously trigger the data collection processes on a wearable device and a VA. We synchronize the data collection processes with two alternative approaches based on network conditions. When there is little delay and the network is stable, we exploit Wi-Fi to send a message to the wearable device to begin data collection. When there is considerable delay or the network is unstable, WearID detects a wake word via the accelerometer, triggering data collection in parallel with a VA. The voice command following the wake word is recorded by both devices for user authentication.

Next, WearID extracts meaningful features from the accelerometer and microphone data. It first denoises the accelerometer data by removing the impacts of human motions and determines the data segments corresponding to speech. We derive the time–frequency representations of the accelerometer data and use them as vibration domain features. WearID also denoises the microphone data by removing the effects of environmental noise. To fill the large sampling rate gap and enable direct comparison between features in the vibration and acoustic domains, we develop a method to convert the time–frequency representations of the microphone data to the low-frequency band through signal aliasing. The converted time–frequency representations are considered acoustic domain features.

Finally, WearID examines the similarity between the vibration domain features and acoustic domain features to determine whether the received voice command is from a user or an attacker. To further accommodate the residual synchronization errors in the sensing data, we derive several similarity scores by comparing the vibration domain features and acoustic domain features within a sliding window. A threshold-based method is applied to the maximum similarity score to determine the authenticity of the voice command. WearID can also be deployed on a shared VA device and verify voice commands from multiple users, such as colleagues and family members. In such scenarios, each user needs to associate his or her account and wearable device with the VA system.

Capturing Voice Commands Through Wearables' Accelerometers

Similarities and Differences of Microphones and Accelerometers

Microphones and accelerometers are microelectronic sensors, but their design is significantly different. Microphones are widely used in various sound recording devices, including smartphones and VAs. They have a membrane and a complementary perforated black plate, which are used to sense sound waves. When the sound of a voice passes through the holes in the black plate and hits the membrane, a microphone captures the sound waves by recording analog capacity change signals. The analog signals are amplified and fed to a low-pass filter, which removes audio signals that are beyond half the sampling rate. An analog-to-digital converter (ADC) converts the analog signals into digital ones. The ADC sampling rate determines the maximum frequency of the recorded sounds, although the analog signals may capture sounds with higher frequencies. The accelerometers used in most mobile and wearable devices measure sound as subtle inertial mass movements caused by changing sound wave pressures. An accelerometer does not contain a low-pass filter between its amplifier and ADC, and thus it can capture vibrations approaching its sensing limit (e.g., up to 4 kHz), making it able to sense human voices, which mainly reside at lower frequencies.

Aliasing Effects in the Vibration Domain

Although accelerometers can capture high-frequency vibration signals, wearables' operating systems limit accelerometers' sampling rate to a much lower frequency of around 100 Hz to reduce power consumption. When we use a wearable's accelerometer to capture the aerial speech vibrations of voice commands, the accelerometer data shows distinctive aliased voice patterns when compared to the original sound. Signal aliasing is an effect that makes different frequency signals indistinguishable when they are undersampled. According to the Nyquist theorem, any frequency signal sampled at a frequency that is less than half its frequency is indistinguishable from a lower-frequency signal within the half-sampling frequency. For illustration, in Figure 3, we show the energy of a chirp signal (i.e., a signal sweeping through 500~1,000 Hz) sampled at a frequency of 100 Hz. We can find a "zigzag" curve in the energy heat map, where the frequency signal at 720 and 780 Hz is mapped to 20 Hz after sampling. Such an aliasing effect enables a wearable accelerometer to record vibrations beyond the ADC sampling frequency.

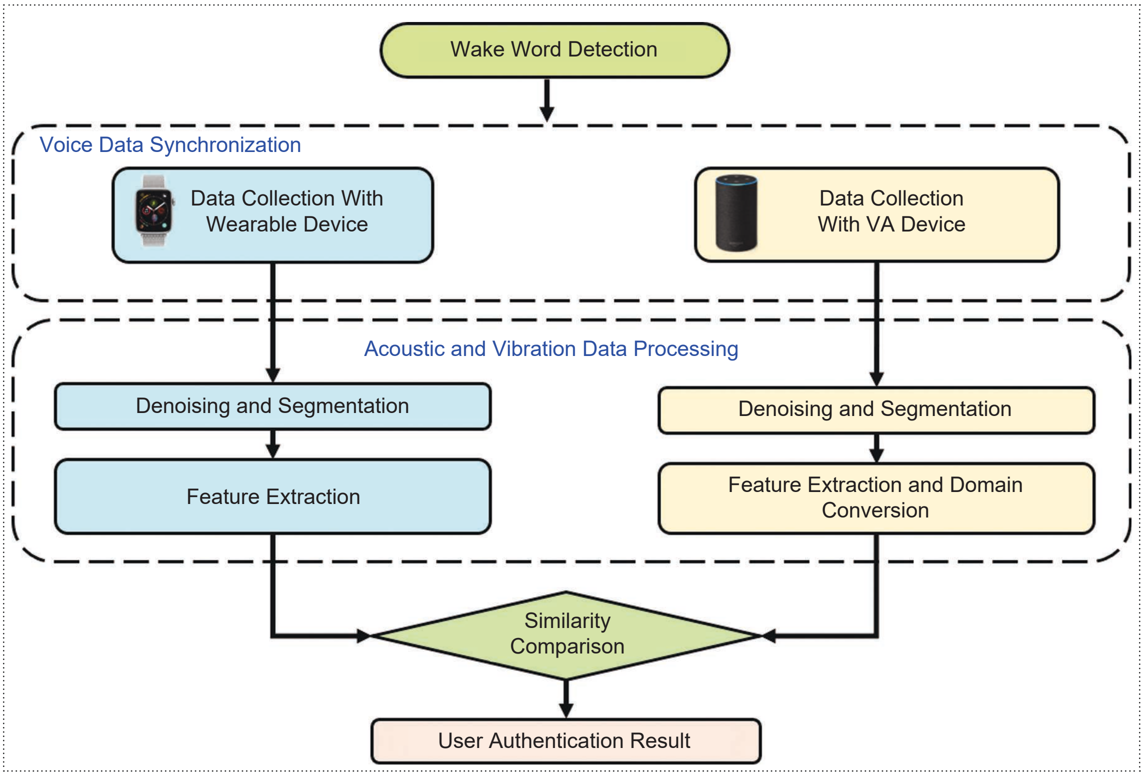

Wake Word Detection
Voice Data Synchronization
Data Collection With Wearable Device
Data Collection With VA Device
Acoustic and Vibration Data Processing
Denoising and Segmentation
Feature Extraction
Denoising and Segmentation
Feature Extraction and Domain Conversion
Similarity Comparison
User Authentication Result

Figure 2. The user authentication system's flow.

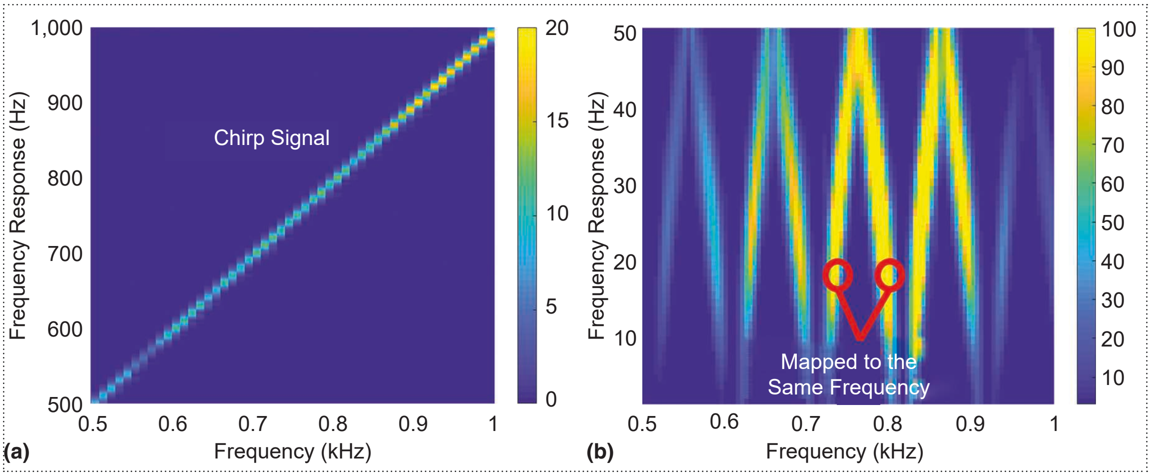


Figure 3. Frequency responses of (a) a VA microphone and (b) a wearable's accelerometer to a chirp signal of 500~1,000 Hz.

Note that aliasing effects are usually removed in microphones by the low-pass filter, which is not included in an accelerometer.

Unique Response to Aerial Speech Vibrations

Due to different hardware designs, accelerometer and microphone data show distinctive human speech patterns. Instead of directly capturing a sound wave, an accelerometer measures wearable device vibrations triggered by speech. Such a mechanism produces a unique response to speech in terms of magnitude and frequency characteristics. To understand these patterns, we examine accelerometer data from a wearable (i.e., an LG Urbane W150) and microphone data from a smartphone (i.e., a Nexus 6) when playing a chirp signal of 0~22 kHz from a loudspeaker. We find that the accelerometer shows only high magnitudes between 400 and ~3,400Hz, while the microphone readings have high magnitudes across a much wider frequency range of 80 Hz~15 kHz. By analyzing the accelerometer data time–frequency patterns, we find that the device has a distinctive energy distribution across frequencies when compared to the microphone. Such unique accelerometer characteristics make it impossible for attackers to reproduce a user's voice command. However, a malicious party may succeed at faking a voice command on microphones.

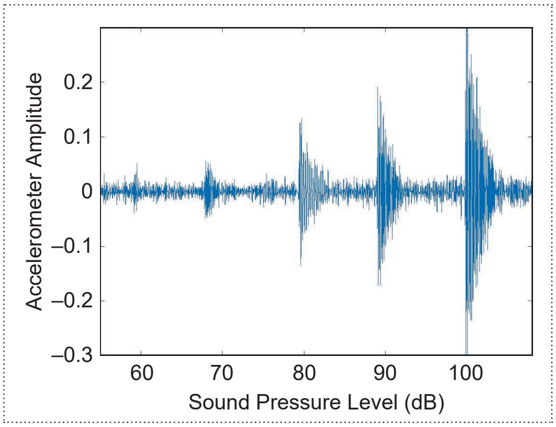


Figure 4. An accelerometer's response to live human speech at different SPLs.

Recording Live Speech Through Wearables

To show that using a wearable accelerometer to capture live human speech is possible, we conduct an experiment using a smartwatch (i.e., a Huawei Watch 2 Sport). We ask a volunteer to speak a word ("calendar") at the sound pressure levels (SPLs) of 60, 70, 80, 90, and 100 dB. The volunteer wears the smartwatch on her left hand and speaks to the watch at a distance of 10 cm. Figure 4 shows the accelerometer data. We observe that the wearable can capture speech with an SPL higher than 70 dB and that the magnitude grows with the SPL. When the SPL reaches 80 dB (presentation-level volume), the accelerometer can clearly reveal the speech. We also test the ability of the wearable's accelerometer to capture a human voice under various subject-to-wearable distances of 5–35 cm (with a 5-cm gap). The subject gives commands to the smartwatch at an average SPL of 80 dB. We find that when the distance increases to 30 cm, speech patterns can barely be observed. Such a short response distance can help WearID prevent many acoustic attacks.

Voice Characteristic Comparison for Audio and Vibration Signals

Voice Data Synchronization

To enable reliable voice characteristic comparison, WearID needs to simultaneously collect accelerometer and audio recordings from a wearable device. We develop two voice data synchronization approaches that depend on network delay. WearID uses Wi-Fi to synchronize the data collection if there is little delay. If the wearable is equipped with a Wi-Fi module, the VA, upon detecting a wake word, sends a message to the device to trigger data collection. If the wearable is not equipped with a Wi-Fi module, the VA can send the message to a smartphone paired with the device. As an alternative, when there is significant network delay, the wearable device can use its accelerometer to detect a wake word and trigger data collection with the VA. Our study shows that a machine learning approach can accurately detect wake words based on the time–frequency features of accelerator data. Given that wearables' accelerometers usually run in the background around the clock, our approach does not introduce additional energy consumption.

Data Denoising and Segmentation

An accelerometer captures noise caused by human motions along with aerial speech vibrations. The noise is unpredictable and can significantly distort speech vibration patterns in accelerometer data. Since the impacts of human motion usually reside at low frequencies, we apply a high-pass filter to remove the effects of low-frequency motions and enhance speech vibrations. After data denoising, we develop a segmentation method to extract accelerometer data containing voice commands that is based on variation. Intuitively, the accelerometer data experiences high variations in the presence of speech vibrations. Therefore, we calculate the accelerometer data variance within short frames and use a threshold on the variance to segment speech.

Time–Frequency Feature Extraction

We apply time–frequency analysis[15] to extract effective features from accelerometer and microphone data for user authentication. Time–frequency analysis has shown great success in speech and speaker recognition tasks. In particular, we exploit a short-time Fourier transform, which calculates energy distributions across frequencies of short frames sampled with a sliding window to extract time–frequency features. As discussed, we need to mitigate the impacts of signal aliasing in the accelerometer data. We develop a transformation method to convert microphone data time–frequency features to an aliasing form that is comparable with the time–frequency features of accelerometer data. Our method mimics signal aliasing effects by calculating the aliasing components based on microphone data time–frequency features. If multiple aliasing components are found to overlap at the same frequency, we accumulate their values during the transformation. Figure 5(b) presents time–frequency features converted from a microphone recording "Alexa" through the proposed method. We find that the converted features have a form that is "equivalent" to the motion sensor features, as given in Figure 5(a).

Voice Characteristics Comparison

We find that the scales of the time–frequency features of the accelerometer and microphone data may vary a lot due to different operating systems. To address this, we develop a scheme to normalize the feature values across frequencies. This process is applied to the time–frequency features in the vibration and audio domains. We also notice that while the accelerometer and microphone data are coarsely synchronized, a minor residual delay may still exist, introducing uncertainty into the authentication results. We develop an algorithm that finds the maximum correlation (i.e., referring to the similarity score) between the time–frequency features in the vibration and audio domains to address this. In particular, we fix the time–frequency features from the microphone data and shift

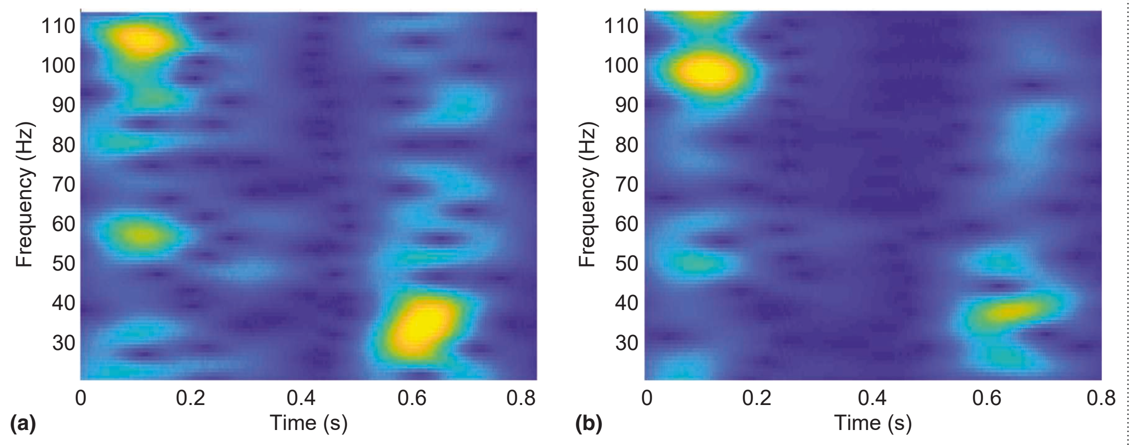


Figure 5. A comparison of (a) an accelerometer spectrogram with (b) a converted microphone spectrogram.

the time–frequency features from the accelerometer data within a small time window to calculate a set of similarity scores. Finally, a threshold-based method is applied to the maximum similarity score to authenticate a user's voice command if the score is higher than an empirical threshold.

System Performance Evaluation

We evaluate WearID by using two smartwatches, a Huawei Watch 2 Sport and an LG W150. The LG W150 is equipped with Invensense M6515 accelerometer that supports sampling frequencies within 4~4,000 Hz. The maximum acceleration that can be measured with this accelerometer is ±16 g. The Huawei Watch 2 Sport has the same measurement range but supports lower sampling frequencies up to 1,600 Hz. Although the accelerometers can capture vibrations of 1.6~4 kHz, the vendors constrain the sampling rates to lower power consumption. We investigate the performance of WearID in a typical office environment with dynamic ambient noises, such as those from air conditioners and ventilation systems, people walking, and conversations. A volunteer wears a smartwatch and gives commands to a VA that is 1 m away. He or she speaks at a presentation-level volume, which is reasonable since most users subconsciously talk louder when using VAs from a distance.

We involve 10 volunteers in a situation where there is no attacker. The study was approved by the Institutional Review Board at Rutgers University, and the reference number is Pro2018001234. The volunteers are asked to speak 20 commands. We compare the commands recorded by the VA against the wearables' accelerometer data to simulate legitimate users. In addition, we record 20 samples of ambient noise by using the smartwatch's accelerometer to test WearID in daily scenarios, where friendly users (e.g., family members and colleagues) may mistakenly trigger WearID and the wearable device records only ambient noise. As shown in Table 1, WearID shows a more than 99% true positive rate (TPR), meaning that almost all legitimate command samples are correctly authenticated. We also find that WearID has a 0% false positive rate (FPR) on both smartwatches, indicating that voice commands from friendly users will be blocked.

Table 1. The WearID performance during normal situations, random attacks, and impersonation attacks.

	Normal situations	Random attacks	Replay attacks
True positive rate	99%	94%	91%
False positive rate	0%	5%	10%

To evaluate WearID during random attacks, we consider one volunteer as a legitimate user and the remaining participants as adversaries. We compare the adversaries' commands recorded by the VA against the accelerometer data of the legitimate user. We find that WearID can verify voice commands with a more than 94% TPR, given a low FPR of 5%, for both smartwatches. For more sophisticated impersonation and replay attacks, WearID achieves higher than 91% TPRs for the two smartwatches, given an FPR of 10%. These results show that WearID is effective at defending against random attacks and impersonation/replay attacks.

We also evaluate WearID during hidden voice command attacks. We collect 100 samples of 10 hidden voice commands replayed by a loudspeaker and compute the similarity between the microphone and accelerometer recordings. The results show that the similarities are approximately zero for the hidden voice commands, meaning that the commands can be well differentiated from the similarity scores of the legitimate users (i.e., around 0.5 for the Huawei Watch 2 Sport and 0.4 for the LG W150). We test the ability of WearID to defend against ultrasound attacks by replaying a signal sweeping across 15 ~25 kHz from a tweeter speaker. In the experiment, we do not observe any sound signals in the recorded accelerometer readings, which confirms that WearID is not vulnerable to ultrasound attacks.

An interesting finding is that the accelerometer can capture unique characteristics in human voices that can be used to determine users' identity. In practical scenarios, multiple users (e.g., colleagues and family members) may use WearID on the same VA. In such cases, WearID can identify them by comparing the accelerometer data of the voice command against the profiled audio data of multiple users. The voice command is determined to belong to a user with the highest similarity. Our experiment results show that WearID can identify 10 users with a 96.5% accuracy when using the Huawei Watch 2 Sport and 91.3% when using the LG W150. This shows that WearID can correctly identify users.

WearID requires little power from a wearable device since computationally intensive tasks (i.e., feature extraction and cross-domain comparison) are performed in the VA system's cloud. A wearable needs only to collect accelerometer data and forward the information to the VA. Since accelerometers have a lower power profile and because voice commands usually last a few seconds, the WearID energy consumption is very low:

our empirical calculation shows that it is less than 0.21 J for one voice command. In addition, traditional VAs need to send voice command data to the cloud for data processing; the communication delay thus dominates the system delay. WearID does not introduce additional delays to a VA.

We presented WearID, a wearable-assisted user authentication system that provides enhanced security for VAs, especially for critical voice commands (e.g., big purchases and important phone calls). WearID authenticates users by comparing the similarity between voice commands captured by a wearable device's accelerometer and a VA microphone. This enables WearID to verify commands without building a user profile based on privacy-sensitive recordings. We developed a feature conversion method that models complex relationships between voice commands recorded with two sensors and designed a method to compare characteristics. Since an accelerometer captures voices within only a short distance (e.g., less than 30 cm), WearID can shield a VA against various acoustic attacks. ■

Acknowledgments

This work was partially supported by National Science Foundation grants CCF-1909963, CCF-2000480, CCF-2028858, CCF-2028876, CNS-1526524, CNS-1714807, CNS-1814590, CNS-2030501, and OAC-1547350 and Army Research Office grant W911NF-18-1-0221.

References

1. "Google Home and Assistant commands—here's the ones you need to know." Android Authority, 2020. https://www.androidauthority.com/google-assistant-commands-727911/ (accessed May 12, 2020).
2. "Consumers need answers to Amazon Echo privacy concerns." IdentityForce, 2018. https://www.identityforce.com/blog/amazon-echo-privacy-concerns (accessed June 10, 2019).
3. H. Zeinali, L. Burget, and J. Černocký, "A multi purpose and large scale speech corpus in Persian and English for speaker and speech recognition: The DeepMine database," in *Proc. IEEE Autom. Speech Recogn. Understanding Workshop*, 2019, pp. 397–402.
4. N. Carlini et al., "Hidden voice commands," in *Proc. USENIX Security Symp.*, 2016, pp. 513–530.
5. G. Cho, J. Choi, H. Kim, S. Hyun, and J. Ryoo, "Threat modeling and analysis of voice assistant applications," in *Proc. Int. Workshop on Inf. Security Appl.*, 2018, pp. 197–209.
6. G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu. "Dolphinattack: Inaudible voice commands," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, 2017, pp. 103–117.
7. D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, 1995. doi: 10.1109/89.365379.
8. R. Togneri, Roberto and D. Pullella, "An overview of speaker identification: Accuracy and robustness issues," *IEEE Circuits Syst. Mag.*, vol. 11, no. 2, pp. 23–61, 2011. doi: 10.1109/MCAS.2011.941079.
9. S. Chen et al., "You can hear but you cannot steal: defending against voice impersonation attacks on smartphones," in *Proc. IEEE Int. Conf. Distrib. Comput. Syst.*, 2017, pp. 183–195.
10. L. Zhang, S. Tan, J. Yang, and Y. Chen, "Voicelive: A phoneme localization based liveness detection for voice authentication on smartphones," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, 2016, pp.1080–1091.
11. D. Wang and P. Wang, "Two birds with one stone: Two-factor authentication with security beyond conventional bound," *IEEE Trans. Dependable Secure Comput.*, vol. 15, no. 4, pp. 708–722, 2016. doi: 10.1109/TDSC.2016.2605087.
12. "Secure authentication with the duo mobile app." Duo Security, 2019. https://duo.com/product/multi-factor-authentication-mfa/duo-mobile-app (accessed May 16, 2019).
13. "Number of connected wearable devices worldwide from 2016 to 2022." Statista, 2021. https://www.statista.com/statistics/487291/global-connected-wearable-devices/ (accessed Feb. 27, 2021).
14. B. A. Roston. "Apple Watch replaces student ID cards for Alabama students." Slash GEAR, 2018. https://www.slashgear.com/apple-watch-replaces-student-id-cards-for-alabama-students-18534624/ (accessed Feb. 27, 2021).
15. L. Cohen, *Time-Frequency Analysis*, vol. 778. Englewood Cliffs, NJ: Prentice Hall, 1995.

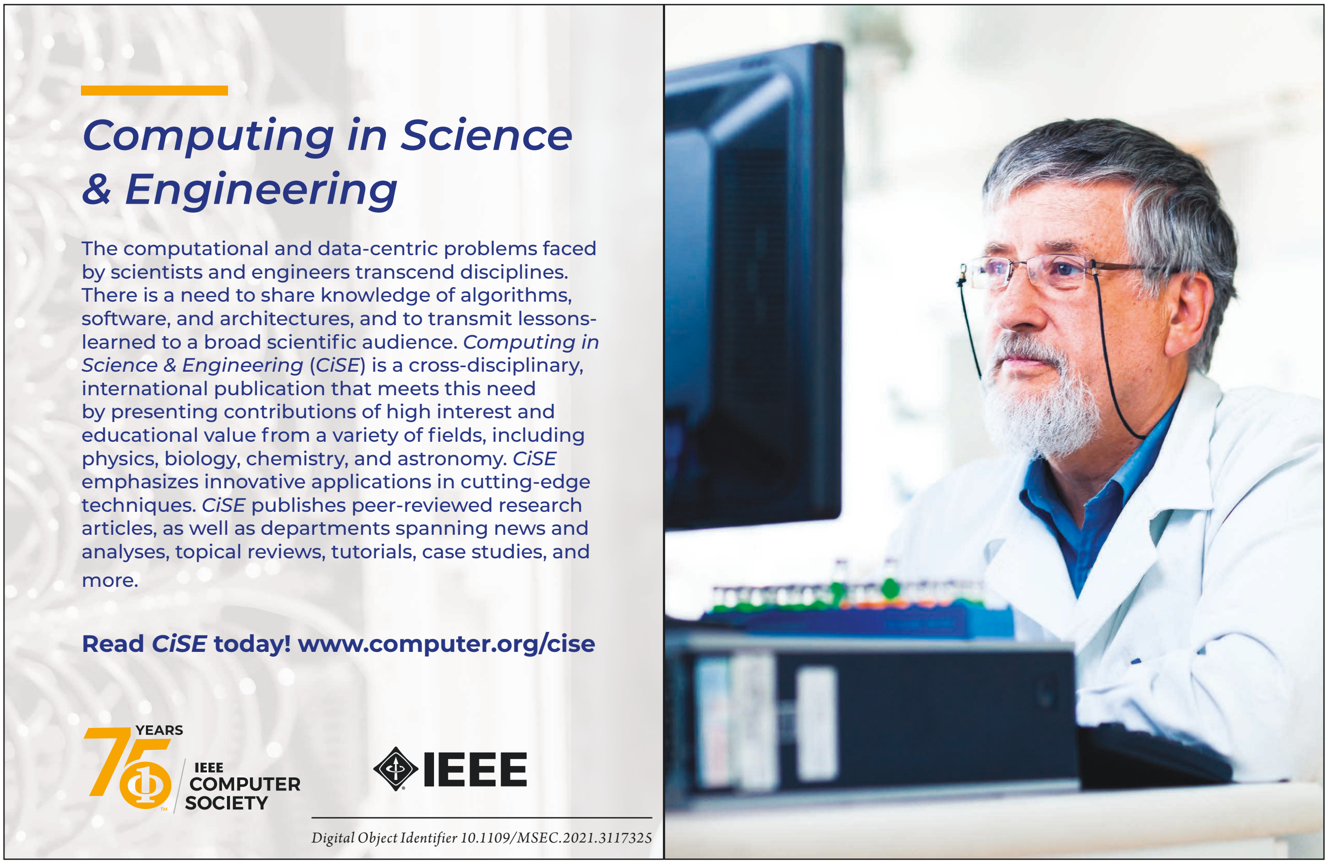
Cong Shi is a Ph.D. candidate in the Department of Electrical and Computer Engineering, Rutgers University, New Brunswick, 08854, New Jersey, USA, where he works in the Data Analysis and Information Security Lab with Prof. Yingying Chen. His research interests include cybersecurity/privacy, mobile computing/sensing, and machine learning. Cong received an M.E. from Stevens Institute of Technology. He is the recipient of a Siemens FutureMakers Fellowship. Contact him at cs1421@scarletmail.rutgers.edu.

Yan Wang is an assistant professor in the Department of Computer and Information Sciences, Temple University, Philadelphia, 19122, Pennsylvania, USA. His

research interests include mobile and pervasive computing, cybersecurity and privacy, and smart health care. Yan received a Ph.D. in electrical engineering from Stevens Institute of Technology. He is the recipient of the Best Paper Award at the 2018 IEEE Conference on Communications and Network Security; 2017 IEEE International Conference on Sensing, Communication, and Networking; and 2016 Association for Computing Machinery Asia Conference on Computer and Communications Security. Contact him at y.wang@temple.edu.

Yingying (Jennifer) Chen is a professor of electrical and computer engineering and a Peter Cherasia Endowed Faculty Scholar at Rutgers University, New Brunswick, 08854, New Jersey, USA, where she is also the associate director of the Wireless Information Network Laboratory and leads the Data Analysis and Information Security Lab. Her research interests include mobile sensing and computing, cybersecurity and privacy, the Internet of Things, and smart health care. Chen received a Ph.D. in computer science from Rutgers University. She received a National Science Foundation CAREER Award, Google Faculty Research Award, New Jersey Inventors Hall of Fame Innovator Award, and IEEE Region 1 Technological Innovation (Academic) Award. She serves or has served on the editorial boards of *IEEE Transactions on Mobile Computing, IEEE Transactions on Wireless Communications, IEEE/ACM Transactions on Networking,* and *ACM Transactions on Privacy and Security*. She is a Fellow of IEEE. Contact her at yingche@scarletmail.rutgers.edu.

Nitesh Saxena is a professor of computer and information sciences at the University of Alabama at Birmingham (UAB), Birmingham, 35294, Alabama, USA, where he is the founding director of the Security and Privacy in Emerging Systems group/lab. His research interests include computer and network security and applied cryptography, with a keen interest in wireless and mobile device security and the emerging field of usable security. Saxena received a Ph.D. in information and computer science from the University of California-Irvine. He is the recipient of two Google Faculty Research Awards, among numerous other recognitions and grants. He serves as an associate editor of *IEEE Transactions on Information Forensics and Security* and *International Journal of Information Security*. Contact him at saxena@uab.edu.



Digital Object Identifier 10.1109/MSEC.2021.3117325