Low-Effort Handheld Device User Authentication Using Musical Sounds

Long Huang, *Student Member, IEEE*, and Chen Wang, *Member, IEEE*

***Abstract*—This work proposes a low-effort user authentication system for handheld devices based on active acoustic sensing. Rather than using dedicated acoustic signals, we find common media sounds like music can serve as a sensing signal to verify the phone user's hand. Specifically, when a notification comes, the smartphone can unobtrusively verify who is holding the device and then decide whether to hide or display the sensitive notification content. Since sound and vibration co-exist, we capture two novel responses via the device's microphone and accelerometer to describe how the individual's contacting palm interferes with the two-domain signals, which are then described as time–frequency images and fed into a convolutional neural network-based algorithm for user authentication. Moreover, we develop a cross-domain method to validate the hard-to-forge physical relationships among the smartphone's microphone, speaker, and accelerometer, which are embedded on the same motherboard. This prevents external sounds from cheating the system. Additionally, we consider vibration alerts as a special type of musical sound and extend our method to work with the smartphone's silent mode. Extensive experiments with ten musical sounds and five phone models show that our method verifies users with 94.5% accuracy and effectively prevents acoustic replay attacks and physical hand forgeries.**

***Index Terms*—Gripping hand, musical sounds, user authentication.**

I. Introduction

HANDHELD devices, especially smartphones, have now become people's most intimate devices. While providing a variety of services to make life easier, they also access and store the user's private information. Though different authentication methods (e.g., PINs, lock patterns, fingerprints, and face recognition) have been deployed for access control, they are either vulnerable to eavesdropping and replay attacks or require the user's participation to provide authentication inputs. For example, an adversary can easily obtain the user's PINs or lock patterns via shoulder surfing [1]. A more advanced adversary can cast the user's fingerprint from the latent fingermarks left on a surface [2], or make a 3-D mask of the user's face leveraging 3-D reconstruction and printing technologies [3]. The forged fingerprint and face can then be used to unlock the user's device. Furthermore, the smartphone user's private information is also easily leaked when the user is not able to provide the authentication input in time. In particular, smartphone notifications often bypass the authentication process and display sensitive content on the screen automatically even when the phone is locked. The inappropriately displayed notification may expose the user's privacy in front of others and cause embarrassing moments, anxieties, financial losses, and distrust [4], [5], [6], [7], especially when the device is held by others or left unattended.

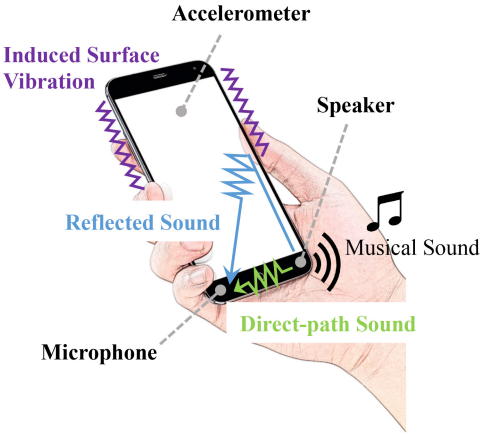


Fig. 1. Musical sound interacting with the gripping hand in the acoustic and vibration domains.

This article introduces a low-effort user authentication method for handheld devices to cover the authentication scenarios when the other authentication methods could not work. We find that musical sounds (e.g., the melodies of notification tones) have sensing capabilities and thus we propose using music to unobtrusively recognize who is gripping the device, which eliminates the need to design specialized sensing signals or modify the devices' existing audio system. Specifically, the musical sound traveling on the smartphone surface would interact with the user's gripping hand before being captured by the phone's microphone as shown in Fig. 1. Since people's hand geometries and gripping strengths are unique, they exert different effects on absorbing and reflecting the original signal. The resulting sound thus carries the user's hand-grip biometric. Furthermore, we note that sound and vibrations co-exist, which enables analyzing the gripping hand's impact in both of the acoustic and vibration domains. It is also important to note that only the musical sounds in the audible band

Received 3 April 2025; revised 22 May 2025; accepted 27 May 2025. Date of publication 29 May 2025; date of current version 8 August 2025. This work was supported in part by NSF under Grant CNS-2450046 and Grant CNS-2440238. This article was presented in part at the 27th Annual International Conference On Mobile Computing and Networking (ACM MOBICOM) [DOI: 10.1145/3447993.3483277]. *(Corresponding author: Chen Wang.)*

The authors are with the Department of Computer Science, Southern Methodist University, Dallas, TX 75205 USA (e-mail: huangl@smu.edu; cwang6@smu.edu).

Digital Object Identifier 10.1109/JIOT.2025.3575010

(not available for ultrasounds) are strong enough to generate observable vibration signals. This novel cross-domain attribute provides enhanced robustness and security, making our system immutable to both acoustic noises and attacks [8], [9].

We develop a convolutional neural network (CNN)-based algorithm to verify the handheld device user's gripping hands using musical sounds. The device's microphone data and accelerometer data are taken as the input, and we derive their time–frequency representations (e.g., spectrogram and scalogram) to capture the user's hand-grip biometrics in both acoustic and vibration domains. Based on that, a CNN model built with five convolutional layers further distinguishes users, and a cluster-based method is developed to integrate the two domains' results to achieve robust user authentication even under high noise levels. Furthermore, we examine the hard-to-forge physical relationship among the smartphone's speaker, microphones, and accelerometer based on their signal-to-noise ratios (SNRs), which measures the authentication validity and prevents external malicious sounds from cheating the systems. In addition, we extend our system with only a few parameter changes to work in the device's silent mode by considering the vibration alert as a special type of musical sound, which also has sensing capabilities to recognize the user's hand.

Our Major Contributions Are Summarized as Follows.

1) We propose a low-effort user authentication method for handheld devices using musical sounds, which closes the security gap left by current smartphone authentication methods. The smartphone can wisely hide or display sensitive notification contents on the screen by sensing whether the user is holding the device.
2) We demonstrate that common media sounds like notification tones are readily available to be used for extracting the user's hand-grip biometric. We further capture the hand's acoustic and vibration responses to the musical sound for across-domain authentication, which significantly improves system robustness and security.
3) We derive four types of time–frequency images from the acoustic and vibration responses to analyze the users' hand-grip biometrics and develop a CNN-based algorithm to verify the user. To address replay attacks, a common challenge for acoustic systems, we develop a cross-domain method to validate the hard-to-forge physical relationship among the phone's mic, speaker, and accelerometer. We further extend the capability of replay attackers to replay both vibration and sound to evaluate our system.
4) We show that smartphone vibrations also result in unique acoustic and vibration responses for each user and thus extend our system to cover the device's silent mode, considering the vibration alert as a special type of music.
5) We address a critical challenge that hinders the practical deployment of learning-based acoustic sensing, when testing and training data have different noise profiles. In particular, our cross-domain design uses the vibration data to correct the errors of its nonlinearly related acoustic data under high noise levels. Furthermore, we use the latent diffusion-based audio generation model for data augmentation, which synthesizes the user's acoustic responses under different noise levels for training. The method not only improves authentication performance in noisy scenarios, but also saves the user training efforts.

II. Related Work

To protect sensitive information in handheld devices, such as smartphones, two types of authentication methods are widely deployed. In particular, knowledge-based methods verify the user's PINs, passwords, lock patterns [10], and graphical secrets [11]. In contrast, the biometric-based methods verify the user's body traits or behavioral characteristics, including fingerprints [12], faceIDs [13], iris [14], hand geometries [15], [16], and voices [17]. However, all these methods are one-time authentications performed only at the device login phase and require active user participation. They are not suitable to handle notifications, which have an unpredictable nature and could be displayed on either locked or unlocked screens.

Continuous authentication is an emerging technique to supplement one-time authentications, which aims to verify the user's identity at any time. For example, behavioral characteristics, such as gaits [18], keystroke dynamics [19], and touch behaviors [20] are extracted during the user's daily activities for authentication. However, these methods still heavily rely on the user's inputs, and thus they could hardly cope with the many notification scenarios when the user's activity is not available. Some methods based on continuously verifying the user's vital signs, such as the breathing patterns [21], [22], heartbeat biometrics [23], [24], and brain waves [25], [26] achieve a constant unobtrusive authentication. However, these methods require a long observation window (i.e., several to tens of seconds) to acquire sufficient vital sign data. Thus, they suffer from long delays and may not work in a timely manner to respond to sudden events. Additionally, continuous authentications drain batteries fast and are not energy-friendly for mobile devices.

Current notification privacy protections mainly rely on the manual management of privacy off-line [27], [28] or on specific situations [29], [30]. However, both types of methods have limited capability to improve the security while degrading the notification usability. For example, the off-line on/off configurations sacrifice benign notification features, such as browsing and managing all app messages in the notification center. The current in-situ methods require the user to perform complicated on-screen operations to switch the phone mode back and forth accordingly to the current situation, which is obtrusive and may already arouse suspicions. To lessen the cumbersome operations, PrivacyShield enables a user to quickly shift the smartphone to a guest mode by a secret finger gesture (e.g., writing an "a") on the locked screen [7]. But similar to other in-situ methods, PrivacyShield still imposes a responsibility on the user to take an action for privacy protection, which is inconvenient and may not work well in complicated practical scenarios.

Different from prior work, we address notification privacy issues by introducing an unobtrusive authentication method.

The hand-grip biometric arouses our interest because if a user wants to read smartphone notification, he or she has to first grab this device by hand. There has been some authentication work on verifying the user's gripping behaviors [31], [32]. These methods extract a user's gripping hand signature, in the form of a unique contacting area and pressure distribution, by using an array of pressure sensors attached to the smartphone's enclosure. But they all require the installation of dedicated hardware (i.e., piezoelectric sensor array), and thus they are hard to be widely deployed. Two recent works show the potential of using dedicated acoustic signals to extract hand-related biometrics [33], [34], such as the touch gesture of a hand over the screen when it enters a PIN/pattern and the phone-holding gestures. But the dedicated sounds used in these methods are still audible and intrusive if used for notifications. Moreover, by solely relying on acoustic signals, these systems are still vulnerable to acoustic noises and replay attacks (e.g., at least 10% FAR under the record-and-replay attack as shown in Table III). In comparison, our method utilizes notification tones to sense the gripping hand and derives the responses across the acoustic and vibration domains to achieve robustness and replay-resistance.

III. Intuition of Using Musical Sounds to Verify Gripping Hands

A. Hand-Grip Biometric

According to existing works, the hand-grip biometric indicates how uniquely a user grips a handheld device, which is mainly determined by the user's hand geometries and gripping behaviors [31], [35], [36]. Based on an array of pressure sensors attached to the handheld device surface (e.g., piezoelectric materials), the hand-grip biometric is captured as the unique shape of the gripping hand (i.e., contact area) and the detailed hand pressure distribution on the device surface. However, obtaining such a biometric with smartphones requires a high-installation overhead. Our work shows that a user's hand-grip biometric can be captured by common media sounds, like music played by the handheld device. When a user holds the phone, the manner in which they grip it exerts a substantial influence on the propagation of sound through the phone's medium.

To describe the hand-grip biometric, we extract two unique responses of a hand to musical sounds by leveraging the fact that sound and vibration co-exist. Prior studies have shown that the sensor data obtained from the motor-accelerometer pair or the speaker-mic pair contain a hardware signature of the smartphone [37], [38]. Our work utilizes the smartphone speaker to transmit musical sounds and its microphone and accelerometer as receivers, and the obtained two responses for analyzing the hand-grip biometric characteristics are also device-dependent.

B. Acoustic Response

When a musical sound is emitted by the smartphone's built-in speaker, it travels along the device surface and actively interacts with the user's gripping hand as illustrated in Fig. 1. In particular, the sound is damped and reflected by the gripping hand, and the resulting signals at the smartphone microphone include the direct-path signal and the reflected-path signals. Since each person's gripping hand has unique physiological traits (e.g., hand geometry, finger size, and body fat ratio) and behavioral characteristics (e.g., gripping strength and finger position), these signals are modified distinctively before reaching the microphone. We define the microphone-received musical sound, including the direct-path and the reflected signals, as an *acoustic response*, which describes the user's hand-grip biometric in the audio domain. We now introduce its two properties.

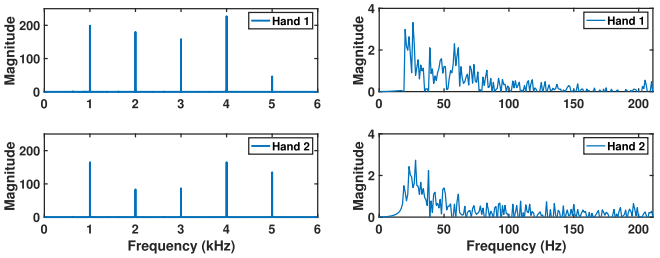


Fig. 2. Acoustic and vibration responses of two gripping hands. (a) Acoustic responses. (b) Vibration responses.

Frequency-Selective: The smartphone sound at a different frequency is uniquely affected by the gripping hand. The reasons are twofold. First, the contacting palm dampens different frequencies with different scales. Second, for each traveling route, the signals of different frequencies show different phases, and thus the multipath signals at each frequency are combined uniquely at the microphone. Therefore, each frequency signal carries one aspect of the hand-grip biometric. To show the feasibility, we play a long beep sound consisting of five single frequency sinusoidal waves using a smartphone, and during this period, the user grabs the phone from a table and places it back. The result is shown in Fig. 2, where the frequency responses of the above five-frequency long beep sounds are presented for two users' hands. The signals exhibit distinct patterns between the two users in the frequency domain, where the amplitudes of some frequencies are strengthened while the others are suppressed. The frequency selectivity indicates that we can utilize the rich frequencies of a musical sound to obtain the comprehensive acoustic representation of the user's hand-grip biometric.

Internal Versus External: Acoustic noises and replay attacks are typical issues for all acoustic systems. We notice that these sounds are mainly from external sources and thus investigate the acoustic responses to built-in and external speaker sounds. We find that external sounds attenuate heavily when propagating over long distances in the air. In comparison, the surface-borne built-in speaker sound dominates the microphone readings with higher SNRs. We conduct an experiment for illustration, in which we play the "iPhone Message" tone with different volumes using a phone's built-in speaker and an external speaker, respectively, in a typical office scenario with around 40dB noises. The external speaker is placed 20 cm away from the smartphone microphone. The Decibel X App [39] is installed on a third device to measure

the sound pressure levels near the microphone. As shown in Fig. 3(a), for the same loudness, the built-in speaker generates microphone readings with 10–35 dB greater SNRs than the external speaker. For example, the built-in speaker at 70 dB volume achieves 60 dB SNR at the microphone, which is 15 dB higher than the external speaker at the same loudness. Though the results show that the smartphone microphone responds better to the built-in speaker sounds, only relying on acoustic responses is not enough to prevent acoustic noises and attacks, which can be louder to overwhelm the built-in speaker sound.

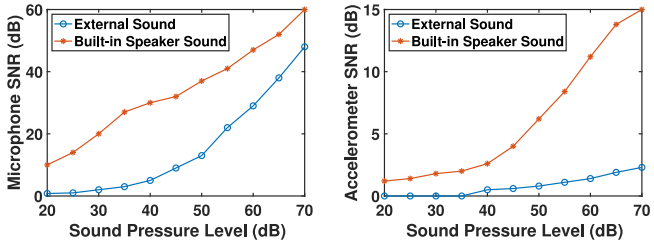


Fig. 3. Comparison of the acoustic and vibration responses under built-in and external speaker sounds. (a) Acoustic responses. (b) Vibration responses.

C. Vibration Response

We find that the musical sound also induces the device surface to vibrate at the same frequencies. The observation is consistent with a prior work that uses human speech audios [40]. The gripping hand suppresses the surface vibrations at the palm contact area determined by the hand geometry. Moreover, greater gripping strength causes higher resistance to vibrations. Thus, the same musical sound results in distinctive surface vibrations for different gripping hands, which can be captured by an accelerometer. We define the accelerometer-captured musical sound as a *vibration response* to describe the user's hand-grip biometric in the vibration domain. It is important to note that the accelerometer could not capture the phone's high-frequency sounds or ultrasounds [41]. Thus, to obtain vibration responses, low-frequency audible sounds need to be used, among which, musical sounds are unobtrusive.

Nonlinear Response: Though induced by a musical sound, the vibration-level representation of the hand-grip biometric is very different. Specifically, the accelerometer data have a nonlinear relationship with the real surface vibrations due to the low sampling rate as described by

$$f_a = |f_v - Nf_s|, N \in Z \tag{1}$$

where, f_s, f_v, and f_a denote the accelerometer sampling rate, the surface vibration frequency and the resulting accelerometer reading frequency, respectively. N can be any integer. This equation also indicates the signal aliasing effect [42], where the surface vibrations at different frequencies could be mapped to a single frequency point at the accelerometer. A minute change in the musical sound's frequency might lead to a completely different accelerometer signal. Such a nonlinear relationship enables the vibration response to further discern the minute hand-grip differences that are hard to recognize by the acoustic response. Moreover, the vibration response is also frequency-selective. Fig. 2(b) illustrates two hands' vibration responses to the above beep sound in the frequency domain. Though the beep sound only has five frequency points, the vibration signals present more frequencies due to the nonlinear relationship, and the amplitudes of these frequencies exhibit distinctive patterns for the two users.

Isolated From External Sounds: The vibration response is also relatively isolated from external sounds, including acoustic noises and various replay attacks. To demonstrate this attribute, we conduct an experiment using the same setup as in Section III-B. Fig. 3(b) compares the SNRs of the vibration responses to the built-in and external speaker sounds. We find that the accelerometer responds to the built-in speaker sound with much higher amplitudes than to the external sound, which only starts to generate 0.4 dB accelerometer readings after 40 dB. Moreover, the SNR gap between the two speakers increases exponentially with louder sound. For example, a 70 dB built-in speaker sound leaves 15 dB SNR signals at the accelerometer while an external sound of the same loudness only leads to 1.4 dB SNR, which is close to the noise level. Being isolated well from external sounds (audible and inaudible), the vibration response is naturally immutable to various acoustic noises and replay attacks. Thus, our system leverages the vibration response to address the practical noise impacts and acoustic replay attacks by checking its relationship with the acoustic response, which provides additional security gains.

IV. System and Attack Models

A. System Flow

We design a low-effort user authentication system for handheld devices using musical sounds, which serves as an external way of access control and can also be leveraged to protect the user's notification privacy while keeping full notification features. The users always set notification tones or vibration alerts on their handheld devices, as without notification signals it is hard to attend immediately to notifications [43]. The basic idea of our system is to capture the unique acoustic and vibration responses to the musical sound of a notification tone to verify the user's hand-grip biometric. Fig. 4 shows the system architecture, which takes the microphone and the accelerometer readings as the input. We perform *Data Preprocessing* to denoise, synchronize, and segment the two modalities' data. The core of our system consists of two components, the *Two-domain User Verification* and the *Cross-domain Validation*, which not only distinguish the users' gripping hands but also examine the validity of each authentication.

It is important to note that musical sounds are more complicated than dedicated sensing signals. The latter often have regular and easy-to-recognize patterns, such as one spectral point at each time index. The complex frequencies of musical sounds make themselves euphonious, but they are more difficult to analyze when used for sensing. To address this challenge, we develop the *Two-domain User Verification*

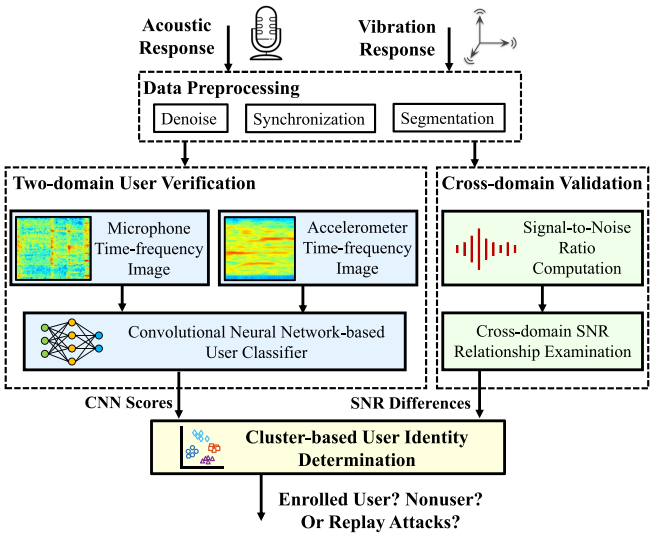


Fig. 4. Architecture of our system.

method, which verifies the user's hand-grip signatures in both the acoustic and the vibration domains. We derive the microphone and accelerometer time–frequency images to describe the detailed time–frequency characteristics of the musical sounds, and resort to using deep learning to recognize the minute differences caused by the users' gripping hands. Specifically, our CNN-based user classifier verifies the two domain time–frequency images, respectively, and the CNN scores for the enrolled user classes and for a nonuser class are output.

The *Cross-domain Validation* examines the authentication input based on the physical relationship between the two modalities' data. The SNRs of the recorded audio and vibration are computed. The cross-domain SNR relationship examination further calculates the SNR differences between the signals in two domains to determine the authentication validity. Only the authentication input resulting from the smartphone's built-in speaker passes the validity check. The sounds from external sources, such as the various replay attacks are detected as not valid and rejected. Based on the obtained CNN scores and the SNR differences, the *Cluster-based User Identity Determination* module calculates the Euclidean distances to the enrolled users' cluster centers and determines whether the authentication request is from an enrolled user, a nonuser, or a replay attack.

B. Attack Models

The adversaries considered in this work include both unintentional observers and notification snoopers whose goal is to reveal the target user's onscreen privacy or SMS verification codes. It is important to note that screen shoulder surfing (when the user holds the device) has been well addressed by prior work, such as the privacy screen film [44], customized screen overlays [45], [46] and HideScreen [47]. Thus, this work focuses on the device-borrowing and the notification-snooping scenarios, when an adversary has gained physical access to the phone and shoulder surfing prevention methods could not work. In particular, we study the following attacks.

Zero-Effort Attack: This attacking scenario includes both the inadvertent privacy leakage and the inexpert notification snooping, and the phone is placed on a table or grabbed by the nonuser in his/her own style when notifications come.

Knowledgeable Impersonation: In this scenario, we consider a more skilled adversary, who has the chance to observe how the user grips the phone and tries to fool our system by imitating a similar hand-grip in-person when notifications come.

External Speaker Replay: This attack mainly happens in the SMS verification code snooping scenario, when the adversary can trigger the message and predict the right time to play the replay sound with an external loud speaker. The replay sounds can be prerecorded stealthily in the user's proximity, when notifications come.

Built-in Speaker Replay: We take one step further to consider a new attack by assuming a much stronger adversary, who could compromise the smartphone's built-in speaker to launch replay attacks. In particular, the adversary could capture the system notification signal and replace it with the replay sound just in time. Moreover, by using the built-in speaker, the adversary forges the unique SNR relationships between the microphone and accelerometer readings to pass our external speaker defense.

3-D Fake Hand Replay: We consider a type of physical hand replay attack, which exploits the current 3-D scanning and printing techniques to physically replicate the user's gripping hand. The reproduced fake hand thus presents a similar hand shape, palm size, finger widths/lengths, and gripping pose as the user hand, which may pass the authentication process of the proposed system.

V. Approach Design

A. Data Preprocessing

1) Noise Removal: The two sensing modalities suffer from different types of noises. The microphone is impacted by the acoustic noises from external sources, while the accelerometer is mainly affected by the user's hand vibrations. These noises must be reduced, so that the minute signal differences caused by the gripping hand can be captured.

Smartphone microphones record up to 24 kHz sounds, but musical sounds are only in a small audible frequency range. For example, the "Samsung whistle" spans the frequencies from 800 to 3000 Hz, and the "iPhone Message" tone has the major frequency components within 10 kHz. We thus design a minimum-order infinite impulse response (IIR) bandpass filter to only focus on the musical sound's frequencies and remove the noises outside of its range, including the low-frequency mechanical sounds and the high-frequency noises. In particular, for each musical sound, we derive its frequency response and utilize a threshold to determine its major frequency range. Based on that, we set the lower and the upper cutoff frequencies of the bandpass filter. Moreover, we utilize the two microphones available on most smartphones to increase the acoustic response dimension and suppress the

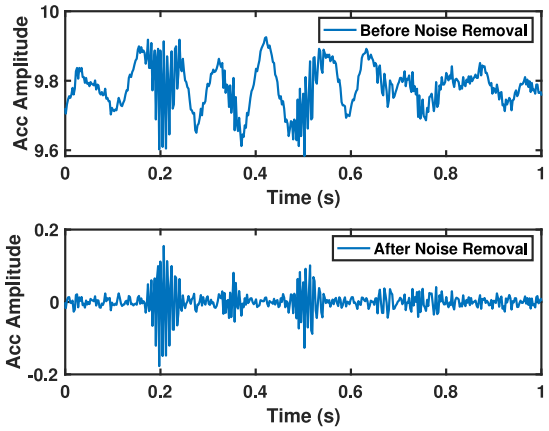


Fig. 5. Removing human body movement noises.

acoustic noises remaining in the passband. By integrating the two microphones' data (as in Section V-C), the acoustic noises that are not correlated at the two microphones are removed.

Smartphone accelerometers record up to 250 Hz vibrations, and they are relatively isolated from external sounds. But when a user holds the phone, the body movements (e.g., hand vibrations) cause the accelerometer readings to be very noisy. Because human body movements are mainly in the low-frequency range, we use a high-pass filter with the 40 Hz cuff-off frequency to remove them. Fig. 5 shows the accelerometer readings before and after passing the filter, where the hand vibration noises are filtered out and the resulting accelerometer readings reflect the vibration response of the hand.

2) Synchronization and Segmentation: The microphones are turned on just before the musical sound to record the complete acoustic response, which does not start at the first microphone sample. To find the starting sample of the acoustic response, we synchronize the microphone data by comparing it to the original signal. The synchronization further facilitates localizing the critical music events based on the original musical signal. Moreover, the calculated time delay verifies the validity of the acoustic response. If the delay exceeds a predefined short period, the acoustic response is not believed to be correctly collected and is rejected directly. In particular, we iteratively shift the received sound S^* by m samples and compute its correlation with the original musical signal S. The maximum correlation indicates the time delay for synchronizing the two signals as

$$\text{delay} = \underset{m}{\text{argmax}} \sum_{n=0}^{N-m-1} S^*(n+m)S(n). \tag{2}$$

By referring to the beginning sample of the first critical music event in the original musical signal, we determine the starting point of the acoustic response and further obtain a T-ms acoustic response segment.

The accelerometer data are also logged before the musical sound begins. Its synchronization process is similar to the above. The difference is that we use the down-sampled original musical signal as the reference, which has the same sampling rate as the accelerometer. Moreover, the synchronization delay is also compared to a threshold to examine the vibration response validity. If it is within a trustable period, we further obtain a T-ms vibration response segment.

B. Time–Frequency Image Derivation

We derive time–frequency images of the acoustic and vibration responses to describe the user's hand-grip biometric in two domains. As there are different ways to analyze a sensing signal regarding both time and frequency, we derive four types of time–frequency representations of the user's hand biometric from the microphone and the accelerometer data, respectively. In particular, we compute the spectrogram based on the short-time Fourier transform (STFT), the scalogram based on the wavelet transform, the persistence spectrum, and the Hilbert spectrum using the Hilbert transform. These are alternative ways to decompose a signal waveform into more detailed 2-D images, which are then fed into our CNN model to learn the gripping-hand biometric. Our system uses the spectrogram by default, and we compare the performance of the four different time–frequency representations in Section VI-C.

Spectrogram: The spectrogram is a time–frequency representation of the data sequence and presents the signal's temporal dynamics at every frequency point. The spectrogram of a sensor data sequence $s(t)$ is computed based on the discrete-time STFT (DT-STFT) as expressed by (3), where $w(t)$ is a window function with length T, and t and f are the time and the frequency index. Each pixel at the spectrogram position (t, f) is computed by (4)

$$\text{DTSTFT}(t,f) = \sum_{\tau=t}^{t+T-1} s(\tau)w(\tau - t)e^{-j2\pi f\tau} \tag{3}$$

$$\text{spectrogram}(t,f) = |\text{DTSTFT}(t,f)|^2. \tag{4}$$

Scalogram: The scalogram is also a function of time and frequency. Different from the spectrogram that uses a window of constant length, the scalogram fragments the signal with a wavelet that is scaled and shifted in time. This enables a better time resolution for the high-frequency components and a better frequency resolution for the low-frequency components when representing the signal. The scalogram of a sensor data sequence $s(t)$ is computed based on the continuous wavelet transform (CWT) as expressed by (5), where a is the scale parameter, b is the location of the wavelet, and ψ is the wavelet function. The scalogram is then computed as the absolute value of the CWT by (6)

$$\text{CWT}(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} s(t)\psi^* \frac{(t-b)}{a} dt \tag{5}$$

$$\text{scalogram}(a,b) = |\text{CWT}(a,b)|. \tag{6}$$

Persistence Spectrum: The persistence spectrum is a histogram of the signal's frequency components, which describes the percentage of the time that a frequency presents in the signal. To compute the persistence spectrum, the signal's spectrogram is first partitioned into 2-D bins. Then for each time value, a bivariate histogram is computed for the power spectrum. The histograms for all the time values are summed up to generate the persistence spectrum, which is plotted over the power and the frequency.

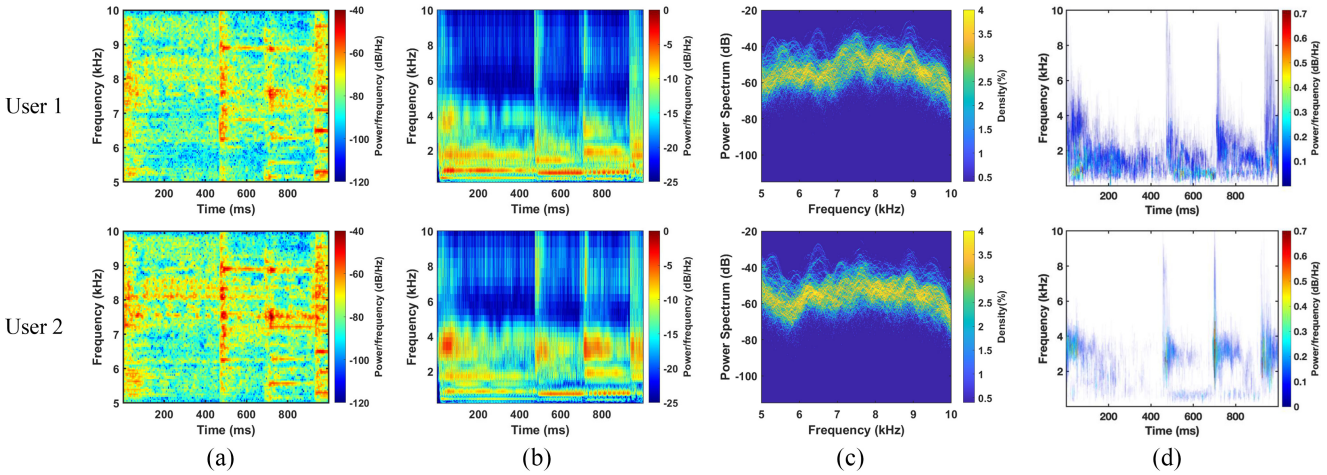


Fig. 6. Distinguishing users via microphone time–frequency representations (illustrated with the iPhone Message tone). (a) Spectrogram. (b) Scalogram. (c) Persistence spectrum. (d) Hilbert spectrum.

Hilbert Spectrum: The Hilbert spectrum can help to distinguish a mixture of moving signals. In this article, we derive the Hilbert spectrum using the Hilbert–Huang transform (HHT), which can be used for time–frequency analysis of nonstationary and nonlinear signals. The HHT first decomposes the signal into a finite number of intrinsic mode functions (IMF), as expressed by (7), where $r_N(t)$ is the residual. For each IMF, the instantaneous amplitude and the instantaneous frequency are computed by the Hilbert transform. The original signal can then be represented by (8), where $a_i(t)$ is the instantaneous amplitude and $\omega_i(t)$ is the instantaneous frequency. The instantaneous energy is computed as $|a_i(t)|^2$, which is plotted as a function of time and frequency

$$s(t) = \sum_{i=1}^{N} \mathrm{IMF}_i(t) + r_N(t) \tag{7}$$

$$s(t) = \mathrm{Real} \sum_{i=1}^{N} a_i(t) e^{j \int \omega_i(t) dt}. \tag{8}$$

1) Microphone Time–Frequency Representation: We first derive the microphone time–frequency representations to describe the user's gripping hand. As shown in Fig. 6, different types of time–frequency representations have a unique way to analyze the user's hand biometric, where "iPhone Message" sound is used for illustration. For each type of time–frequency image, the same musical sound results in distinctive patterns for two users' gripping hands. For the spectrograms in Fig. 6(a), the differences between users are especially obvious in the frequency range of 7 k–9 kHz, which is a critical music event (e.g., a note) of the played musical sound. For the scalograms, Fig. 6(b) shows that the two users' hands can be differentiated from different frequency ranges. The patterns shown by the persistence spectrum in Fig. 6(c) and the Hilbert spectrum in Fig. 6(d) can also be used to distinguish the two users. For authentication, we first analyze the original musical signal to determine a frequency span that covers all of its critical music events and then focus on the microphone time–frequency representations within this frequency span.

2) Accelerometer Time–Frequency Representation: Similarly, the accelerometer time–frequency representations are also derived. But different from the microphone time–frequency representations, which mainly carry the hand-grip biometric information by the critical music events, the accelerometer time–frequency images contain useful information at all of its frequencies. The reason is that the surface vibrations resulted from the critical music events are nonlinearly mapped to any frequency of the accelerometer. As shown in Fig. 7, the accelerometer time–frequency representations are distinctive between two users at many frequencies. All the four time–frequency representations have their unique patterns to distinguish the users. We thus use the entire accelerometer time–frequency images for authentication.

C. CNN-Based User Identification

We use a CNN to learn people's hand characteristics from complex musical sounds. Though CNN is known to be able to learn the spatial patterns in images, the derivation of the time–frequency images allows CNN to learn not only spatial (i.e., frequency in this article) but also temporal characteristics from audio signals. Furthermore, we leverage CNN's strong multiclass classification capability to address the user's behavioral inconsistency (e.g., varied hand-grip styles or grip pose) in binary classification scenarios.

Five-Convolutional-Layer CNN Model: We develop a CNN model with five convolutional layers. For each convolutional layer, we use the rectified linear unit (ReLU) as the activation function and add a max-pooling layer at the output to down-sample the feature maps in both time and frequency domains. We also add a final max-pooling layer to pool the input feature map globally over time, which allows the network to make classifications independent of the temporal positions of the acoustic signal or vibration signal. Batch normalization layers are added to speed up the training process, as well as to reduce the sensitivity to network initialization. In addition, we add a dropout layer to randomly drop 30% of the input features,

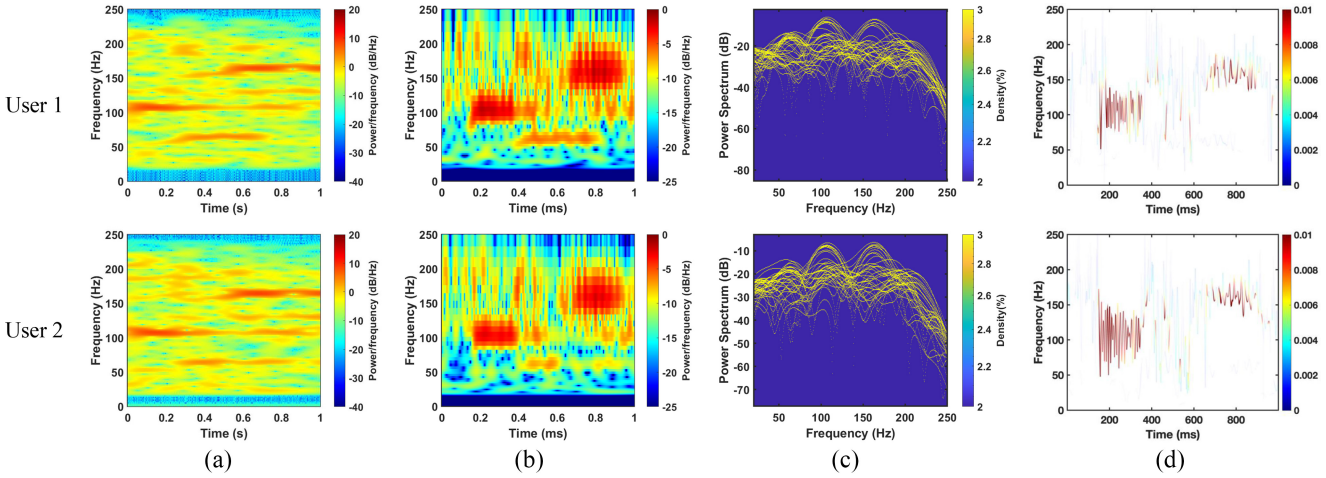


Fig. 7. Distinguishing users via accelerometer time–frequency representations (illustrated with the iPhone Message tone). (a) Spectrogram. (b) Scalogram. (c) Persistence spectrum. (d) Hilbert spectrum.

TABLE I
ARCHITECTURE OF THE CNN MODEL

Layer	Kernel Size	Output Size	# Parameters
Input: Time-frequency Images	-	(40,98,1)	0
Conv2D + RecLineU	(3,3,1,24)	(40,98,24)	240
Max Pooling	(3,3)	(20,49,24)	0
Batch Normalization	-	(20,49,24)	48
Conv2D + RecLineU	(3,3,24,24)	(20,49,24)	5208
Max Pooling	(3,3)	(10,25,24)	0
Batch Normalization	-	(10,25,24)	48
Conv2D + RecLineU	(3,3,24,48)	(10,25,48)	10416
Max Pooling	(3,3)	(5,13,48)	0
Batch Normalization	-	(5,13,48)	96
Conv2D + RecLineU	(3,3,48,48)	(5,13,48)	20784
Conv2D + RecLineU	(3,3,48,48)	(5,13,48)	20784
Max Pooling	(1,13)	(5,1,48)	0
Batch Normalization	-	(5,1,48)	96
Dropout	-	(5,1,48)	0
Fully Connected + Softmax	(240,2)	(2)	482
Output: Probability Distribution	-	(1)	0

which prevents the network from memorizing specific features of the training data and reduces overfitting. The result is finally passed through a fully connected layer and a softmax layer. The CNN scores (i.e., probabilities) are output for all the enrolled user classes and a nonuser class.

Table I shows the detailed architecture of the CNN model. In particular, the microphone time–frequency images and the accelerometer time–frequency images are interpolated and normalized into 2-D matrices of size 40 by 98, respectively. Each matrix passes through a convolutional kernel of size 3 with a stride of 1 and a 3-by-3 max-pooling layer with a stride of 2, which is iteratively repeated 5 times. The number of filters used in each convolutional layer ranges from 24 to 48. After the dropout layer, the fully connected layer performs the classification based on the flattened high-level features, and the softmax layer maps the results to each class, which is shown as a probability distribution. Cross-entropy is used as the loss function and the Adam optimizer is used for training. Since a smartphone has two acoustic channels (i.e., the top and the bottom microphones) and three vibration channels (i.e., a 3-axis accelerometer), the CNN model outputs $5(h+1)$ CNN scores for each authentication input, which present the confidence levels for the h enrolled user classes and 1 nonuser class. The CNN scores are then integrated to determine the user's identity based on a cluster-based method as introduced in Section V-E. The trained model has 58202 parameters and a size less than $0.3MB$. The time and space complexity are $21.4M$ FLOPs and $1MB$, respectively, which are suitable to deploy on most mobile devices [48].

D. Cross-Domain Validation Check

Extensions of Machine-Speaker-Based Attacks: We note that an adversary may launch acoustic attacks using a machine speaker, which shows much higher success rates to fool an acoustic-based authentication system than the in-person impersonations [8], [9], [49], [50], [51]. The typical attacking scenario considered in prior work is to place an external speaker at a distance from the target device, which is then attacked by the airborne malicious sounds. But because airborne external sounds can rarely generate sufficiently strong vibration responses as shown in Section III-C, such attacks could hardly succeed against our system. We thus take one step further and consider the possible extensions of such attacks that could generate stronger surface vibrations at the target device. In particular, we find three categories of machine speaker-based attacks: 1) when the external speaker shares a common solid surface (e.g., a table) with the target device; 2) when there is no shared surface between the two devices; and 3) when the target device is placed right on the external speaker. The intuition is that the external speaker induced vibrations can be transmitted to the target device via the physical contact.

Cross-Domain SNR Relationships: To prevent these replay attacks, we propose to examine the cross-domain physical relationships uniquely presented by each smartphone to verify the authentication validity. Because the external speaker is outside of the target device, it is hard to build up the physical relationships among the smartphone speaker, microphone, and accelerometer, which are on the same motherboard enclosed by the smartphone case. We conduct a fundamental study

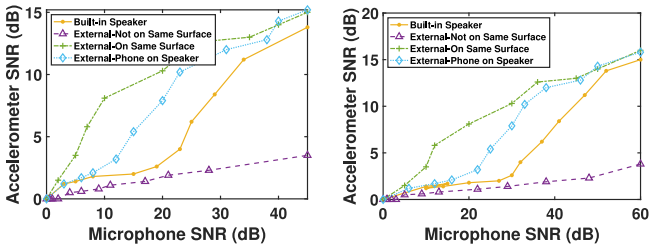


Fig. 8. Cross-domain SNR relationships to prevent external speaker sounds (illustrated with S8 phone). (a) Mic 1 versus Accel. (b) Mic 2 versus Accel.

to investigate such device-dependent physical relationships in four scenarios, when the sound is played with the sound pressure levels from 20 to 70 dB by the built-in speaker and an external speaker. Figs. 8(a) and (b) illustrates the SNR relationships between the accelerometer and each of the two microphones (Mic1 and Mic2) of an S8 phone, respectively. We observe that both cross-domain SNR curves of the built-in speaker have only a few overlaps with that of the three replay attacks. The results demonstrate that the external speaker could not mimic the cross-domain SNR relationships exerted by the built-in speaker. Specifically, the traditional not-on-same-surface curves are far apart from that of the built-in speaker, which are in-between and separated from the on-same-surface and on-speaker scenario curves. When the sound is greater than the built-in speaker's 30% volume (i.e., Mic1 at 15 dB and Mic2 at 26 dB), the built-in speaker curves are farther apart from the thee attacking scenarios. When the sound pressure level equals the built-in speaker's 100% volume (i.e., Mic1 at 44 dB and Mic2 at 60 dB), they have the closest SNR distances. But even at this point, the built-in speaker's cross-domain SNR relationships are still separated by at least 1 to 2.5 dB distances from other curves. We thus, calculate the SNR differences between the acoustic and the vibration responses to determine the authentication validity.

E. Cluster-Based Classification Decision

The obtained CNN scores in two domains and the cross-domain SNR differences are integrated by our cluster-based method to determine whether the authentication input is from an enrolled user, a nonuser, or an invalid source (e.g., replay attacks). In particular, the user's cluster is obtained during the enrollment, which is differentiated from the clusters of the nonusers and the invalid responses. During the authentication, we calculate the Euclidean distances between the authentication input and the enrolled users' cluster centers. If the Euclidean distances to each user cluster are beyond the preset thresholds, the authentication is rejected. Otherwise, the user identity is determined based on the closest user cluster. Fig. 9 illustrates our clustering results of the authentication responses from five different classes, including the enrolled user, the nonusers and three types of replay attacks. The user cluster is clearly separated from the nonuser and the attack classes. All three types of replay attacks achieve high microphone CNN scores. Thus, if only relying on the microphone for authentication, the system can be easily cheated. The replay

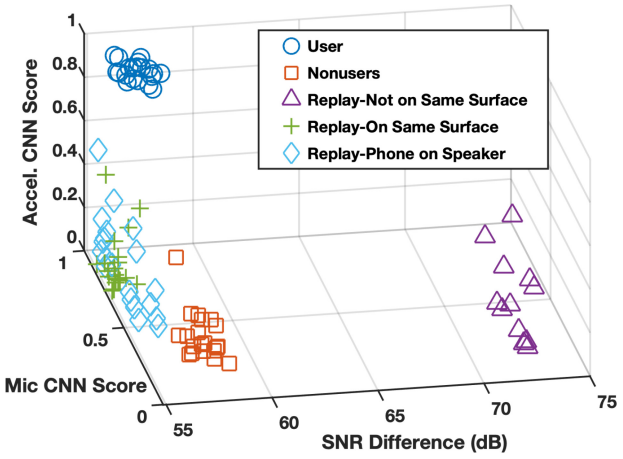


Fig. 9. Clustering results of the user, nonuser and four types of machine speaker-based impersonations.

attack is a typical issue for most acoustic-based authentication systems. Our system is different, because we further use the accelerometer's CNN score and the SNR difference to distinguish valid authentication inputs from the external replay attacks. This is only available when using smartphone musical sounds. In comparison, the nonuser class shows similar SNR differences as the user class, because they both use the built-in speaker sounds, but both its microphone and accelerometer CNN scores are low. Moreover, it is hard to forge the two domain features simultaneously, which have a nonlinear relationship.

VI. Performance Evaluation

A. Experimental Setup

Devices: We evaluate our system on five mobile device models (i.e., Samsung Galaxy S8, Samsung Galaxy S20, Google Pixel2, LG K50, and Motorola G8). All these devices are equipped with two microphones and one accelerometer. The devices run Android 9.0/10.0 and the microphone sampling rate is set to 48 kHz. The accelerometer sampling rate is set to the maximum for each device, which is 423Hz, 403Hz, 405Hz, 257Hz, and 395Hz for S8, S20, Pixel2, K50, and G8, respectively.

Experimental Platform: We develop an experimental platform based on Android, which plays musical sounds through the device's built-in speaker and records the microphone and the accelerometer readings simultaneously. In particular, the platform launches three threads to collect the authentication data. When a notification comes, the main thread first launches one thread to record the stereo sound using *android.media.AudioRecord* and one thread to play the musical sound using *android.media.MediaPlayer*. The accelerometer readings are logged in the main thread. The authentication data are processed offline.

Data Collection: We recruit 40 participants (18 males and 22 females) aged from 25 to 40 to conduct experiments for 24 months. The participants include graduate students, undergraduate students and university faculties. The work has been approved by the Institutional Review Board (IRB) of

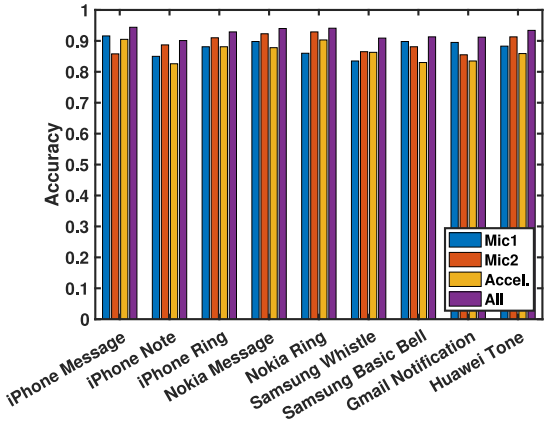


Fig. 10. User authentication with nine musical sounds.

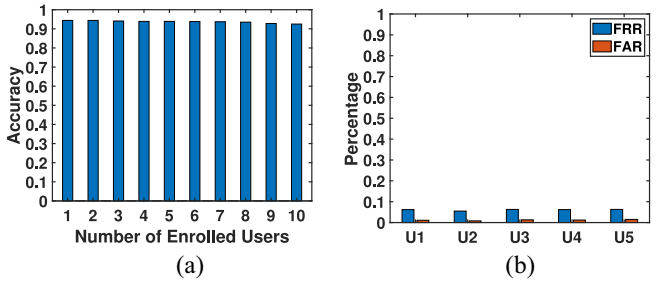


Fig. 11. Performance for the phone enrolled with multiple users. (a) Number of enrolled users. (b) Five enrolled users.

LSU and SMU. The data are collected in two sessions for each participant, with the first session only used for training and the second only for testing. The interval between each participant's two sessions is from two weeks to up to four months. In the first session, the participants are asked to be familiar with the shape and size of each given device by grabbing it and feeling it for 5 min before data collection. In the second session, no such practice process is required. The participant needs to regrab the device 20 times in the first session and 40 times in the second to involve the behavioral inconsistency that could be caused by every grabbing action. For each regrab, nine musical sounds and one vibration alert are consecutively played with 2-s silent intervals, and only 1-s sound is used for each tone. The nine musical sounds include six short message tones and three ringtones, which are "iPhone Message," "iPhone Note," "iPhone Ring," "Nokia Message," "Nokia Ring," "Samsung Whistle," "Samsung Basic Bell," "Gmail Notification," and "Huawei Tone." After the ten musical sounds, the "iPhone Message" is further played with 50%–100% volumes for the volume study. The major experiments are conducted in the typical indoor scenario with 40dB noises. We further use a loudspeaker to simulate the background noises from 40 to 80 dB while playing the "iPhone Message" on S8 to study the noise impact. A part of the system design and results has been reported in the conference version [56].

Attack Simulations: To simulate the knowledgeable impersonation attacks, the authors and four participants act as the attackers. The attackers watch each target participant's data collection process and imitate their gripping hands to attack the authentication system later. For the external speaker replay attacks, we use an external speaker to play the replay sounds in three scenarios, *not-on-same-surface*, *on-same-surface* and *on-speaker*. We directly use the target participant's microphone data as the replay sounds rather than the side-channel recordings, because side-channel recordings already suffer from degraded fidelity. We did not use ultrasounds, because ultrasounds have difficulty generating vibrations detected by the accelerometer and are thus very easy to detect. We did not use adversarial examples either, because no adversarial learning method has yet been developed to forge both acoustic and vibration signals simultaneously. For the built-in speaker replay attacks, the replay sounds are directly played by the smartphone's built-in speaker. For the 3-D fake hand attacks, we use the fake hand to hold the device during the authentication process.

B. Verifying Users via Musical Sounds

Overall Performance: As smartphones are personal devices, they usually have a single user for most cases. We thus start the system evaluation with one enrolled user and test nine popular musical sounds (notification tones). In particular, when one of the 40 participants is selected as the legitimate user, the other 39 participants are then treated as the nonusers. This process is repeated for all the 40 participants. The averaged user verification accuracies are presented in Fig. 10. We observe that our system achieves a high accuracy of authenticating the users with all of the nine musical sounds when integrating the responses in two domains. In particular, "iPhone Message," "Nokia Message," and "Nokia Ring" perform the best with around 94.5% accuracy. Though different tones have different frequency ranges, signal powers, and frequency occupation ratios [52], they all present good capabilities to distinguish people's hands. Their accuracies are all over 90.1%. Furthermore, using the two-domain responses to verify users achieves better performance than using a single domain. For example, the "iPhone Message" tone achieves 94.4% accuracy by integrating both responses. It is higher than using the microphones or the accelerometer alone, which are 92.3% and 90.5%, respectively. The results indicate that the two novel responses extracted from the audible sounds help verify users with high performance, and our system generally works for different musical sounds.

Multiple Enrolled Users: We next evaluate our system when there are more than one enrolled users (e.g., h). This is a classification problem with $h+1$ classes, where all nonusers are included in one class. Fig. 11(a) presents the classification accuracy, when there are 1 to 10 enrolled users, respectively. We observe that the classification accuracy is high for all these cases. In particular, when the enrollment number is less than 6, our system achieves above 94% accuracy. When the number of enrolled users increases, the accuracy slightly decreases. When there are 10 enrolled users, our system achieves 92.5% accuracy. We future present the false rejection rate (FRR) and false acceptance rate (FAR) for each enrolled user when the enrollment number is five. Fig. 11(b) shows that our system achieves both a low FRR and a low FAR for each user. Specifically, the FRRs of the users are all below 6.3%, and

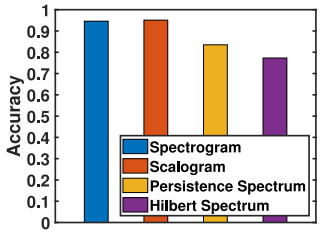


Fig. 12. Using different time–frequency analytical methods for authentication.

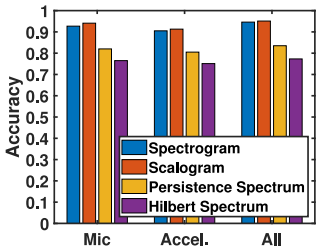


Fig. 13. Time–frequency analysis performance in separate acoustic and vibration domains.

the FARs are around 1%. The nonusers have 5.9% FAR to be misclassified as any one of the enrolled users. The results indicate that our system can accurately verify users and prevent nonusers in the multiuser scenarios.

C. Different Time–Frequency Analytical Methods

Fig. 12 shows our system performance using different time–frequency representations derived in Section V-B, when "iPhone Message" is played on an S8 phone. We find that the scalogram performs the best with a 95.1% accuracy, which is slightly higher than the 94.6% accuracy achieved by the spectrogram. The reason is that compared to the STFT, the wavelet transform's time–frequency resolution is not fixed but can adjust to the signal's different frequencies, where lower frequency components are represented with finer frequency resolution and coarser time resolution, while higher frequency components are represented with coarser frequency resolution and finer time resolution. Differently, the persistence spectrum and the HHT both achieve lower user authentication performance. Specifically, the persistence spectrum achieves an 83.5% accuracy and the Hilbert transform achieves a 77.3% accuracy. The reason is that compared to the spectrogram and the scalogram, the persistence spectrum and the Hilbert spectrum describe the signal's time–frequency features with coarser resolutions.

Fig. 13 further shows the system performance using the time–frequency representations derived separately from the acoustic and the vibration domains, when "iPhone Message" is played on an S8 phone. In particular, for the microphone time–frequency representations, the scalogram performs the best with a 94.1% accuracy, which is slightly higher than the 92.7% accuracy achieved by the spectrogram. The persistence spectrum and the Hilbert spectrum achieve an 82% and a 76.5% accuracy, respectively. For the accelerometer time–frequency representations, our system achieves a 91.3%, 90.5%, 80.5%, and 75.1% accuracy when using the scalogram, the spectrogram, the persistence spectrum, and the Hilbert spectrum, respectively, with the scalogram performing the best. We thus, recommend using the scalogram or the spectrogram to analyze the two responses.

D. Fusion of Two-Domain Responses

We now discuss the different fusion methods to leverage the two-domain sensing information for making the final authentication decision. Our proposed authentication system, by default, fuses the acoustic response and the vibration response at the decision level, which is achieved by integrating their CNN scores. We also develop an alternative feature-fusion method, which integrates the two-domain responses at the feature level. Specifically, we tune the structure of our CNN-based user verification model by fusing the two domains' high-level features before the fully connected layer, whose output is then processed by the softmax function to predict the user's identity. In the tuned CNN model, the microphone time–frequency representations and the accelerometer time–frequency representations are first separately processed by an identical structure formed with a series of convolutional layers, max-pooling layers, and dropout layers. The two responses' high-level features output by the dropout layers are then concatenated to examine the user's biometrics in two domains. The performance comparison of the feature-level fusion and the decision-level fusion methods is conducted using "iPhone Message" and Samsung Galaxy S8. The result shows that the feature-level fusion of the two-domain responses performs slightly better than the decision-level fusion, whose verification accuracies are 95% and 94.4%. There is still room to further explore an optimal method to fuse the two sensing domains for authentication, which is left to future work. For example, rather than concatenating the two domain sensing's results at the decision level, a new learning architecture to integrate the two domain sensor data at the algorithm level is worth further exploring.

E. Impact of Other Factors

Device Models: Since different devices have different microphone and accelerometer configurations, shapes, sizes and surface materials, we evaluate our system on five different device models. Fig. 14(a) shows the user verification accuracy for each device, when "iPhone Message" is used. We find that all five device models perform well in verifying the users, and achieve accuracies between 92% and 95% based on two domains. In particular, Samsung Galaxy S8, Samsung Galaxy S20, and LG K50 achieve around 94.5% accuracy, and perform better than Google Pixel 2 and Motorola G8. When only using the microphones for user authentication, LG K50 achieves the best performance with 92.5% accuracy. When only using the accelerometer for user authentication, Samsung Galaxy S8 performs the best with 90.5% accuracy. The receiver operating characteristic (ROC) curves in Fig. 14(b) further confirm the high performance of the five models, which present high true acceptance rates (TARs) and low FARs. The result indicates our system generally works for different device models.

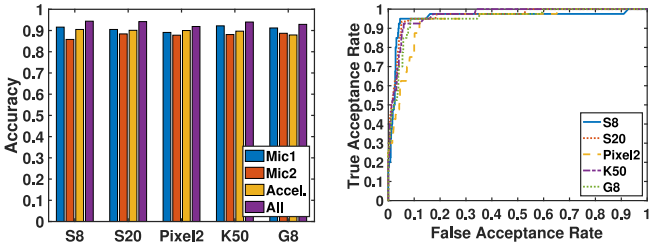


Fig. 14. Verification on different device models. (a) Accuracy. (b) ROC curves.

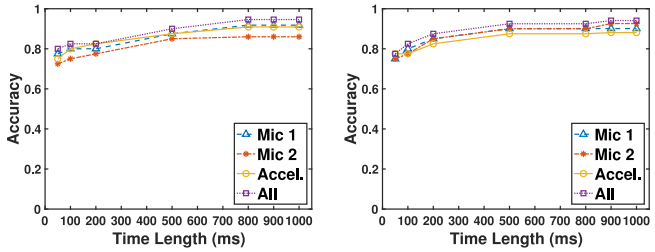


Fig. 15. Musical-sound time-length study. (a) iPhone message. (b) Nokia message.

TABLE II
IMPACT OF THE SMARTPHONE SPEAKER VOLUMES

Volume	50%	60%	70%	80%	90%	100%
Mic1+2	0.865	0.878	0.893	0.904	0.910	0.924
Accel.	0.782	0.831	0.869	0.891	0.901	0.909
Mic + Accel.	0.881	0.896	0.909	0.926	0.934	0.946

Sound Lengths: A longer duration of the musical sound means more microphone and accelerometer samples to describe a higher resolution of the user's hand-grip biometric. However, the cost is a longer authentication time, which may influence the user experience. Therefore, we study the impact of the musical sound lengths on the system performance and try to find a proper duration to collect the authentication input. Fig. 15 shows the time length study for two musical sounds (i.e., "iPhone Message" and "Nokia Message") on S8. We find that, for both sounds, a longer time length slowly leads to a higher accuracy. Such an increasing trend can be observed no matter whether the microphone and accelerometer are separately used or are integrated together. When integrating the two domain responses, both musical sounds achieve over 90% accuracy at 500 ms. The result indicates that our system can verify a user effectively by only using a 0.5–1-s part of the musical sound.

Speaker Volume Impact: The volume to play the tone will also influence the user verification's performance. Table II presents such impacts when the "iPhone Message" is played on S8 with six volume levels from 100% to 50%. We find that the user authentication accuracy slowly drops from 94.6% to 92.6% when the built-in speaker volume is set from 100% to 80%. This is because the SNRs of the acoustic and vibration responses are decreased due to the reduced tone sound and surface vibrations. But even when the volume is set to 60%, our system still achieves around 90% accuracy. The result indicates that our system is not limited to high speaker volumes and can also work well with low volumes.

Nonhand Scenarios: Our system not only verifies who is holding the device but also recognizes the various contexts when the smartphone is not in a hand (unattended, lost, or under zero-effort attack). We include one *nonhand* class in our model to cover the scenarios when the smartphone is placed on a table, bed, or sofa. The "iPhone Massage" sound is played by the five phone models on the three surfaces. For each surface, 80 instances are collected in two sessions, where the regrabbing and repositioning are performed per instance. Fig. 16 shows the accuracy of our system in identifying

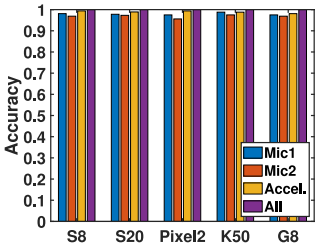


Fig. 16. Recognizing nonhand scenarios (table, bed, and sofa).

whether the smartphone is in a hand or not (i.e., placed on a table, bed, or sofa). We find our system achieves a high performance of recognizing the nonhand contexts with all five phone models, which all achieve 100% accuracy when both acoustic and vibration responses are used. When only the microphone is used, they achieve 97.5% to 98.8% accuracy. When only the accelerometer is used, their accuracy is 98.1% to 99.4%. The result confirms the effectiveness of our system in protecting unattended phones and defending against zero-effort attacks.

Phone Case Impact: We further consider the scenario when people use phone cases to protect their devices, as the phone cases may block or absorb the stimulus signals, thus influencing our system's performance. In particular, we attach a rubber phone case to Samsung Galaxy S8 and use it to collect data from the participants. Fig. 17 compares the user verification performance of Samsung Galaxy S8 with and without a phone case, when "iPhone Message" is played. We find our system achieves similar performance for these two scenarios. When integrating the acoustic and vibration responses, the two scenarios achieve 94% and 94.4% accuracy, respectively. When only microphone data are used, they achieve 92.1% and 92.3% accuracy, respectively. When only the accelerometer data are used, the scenario with the phone case turns out to have a slightly lower performance than the scenario without the phone case, achieving an accuracy of 89.7% and 90.5%, respectively. The reason is that the rubber phone case will absorb and attenuate part of the induced vibrations, resulting in weaker accelerometer readings. The results indicate that the proposed system can work well on the devices covered by cases.

F. Performance Under Attacks

Knowledgeable Impersonations: The performance of our system in defending against the knowledgeable impersonation

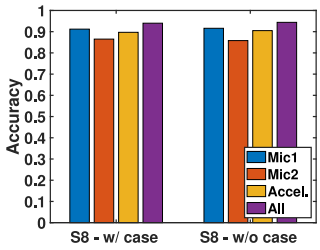


Fig. 17. Adding a phone case.

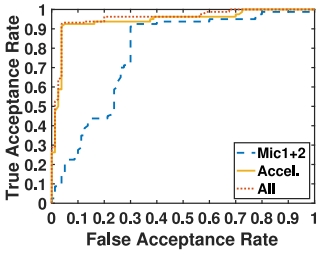


Fig. 18. Defending against built-in speaker replay attacks.

TABLE III
UNDER KNOWLEDGEABLE IMPERSONATIONS AND EXTERNAL SPEAKER REPLAYS

Scenario	Knowledgeable Impersonation	External Speaker Replay		
		Not on Same Surface	On Same Surface	On Speaker
FAR - Mic Only	0.069	0.100	0.118	0.127
FAR	0.061	0	0	0
FRR	0.025	0	0	0
Accuracy	**0.943**	**1**	**1**	**1**

attack is shown in Table III using FAR and FRR. Our system prevents the knowledgeable impersonation attacks with 6.1% FAR, while the enrolled user has 2.5% FRR to be rejected. The EER is 5.6%. The overall accuracy is 94.3%, which is only slightly lower than that in the above normal situations or zero-effort attacks. The result shows that it is hard for an adversary to imitate the user's hand-grip for passing the authentication, which is an implicit biometric compared to the traditional physiological and behavioral biometrics. The handgrip biometrics is not only determined by the gripping pose but is also influenced by other factors, such as gripping strength, hand shape, and body fat ratio, which could be hard for attackers to imitate.

External Speaker Replays: Table III also presents the FARs and FRRs of our system in the context of preventing three types of external-speaker replay attacks. We find that our system successfully prevents all the acoustic attacks from the external speakers, and both FARs and FRRs are 0% when the two domains' CNN scores and the cross-domain SNR relationships are used. The accuracy is 100%. For comparison, we also present the FARs when only the microphone data is used, which are all over 10%. The results further confirm that only relying on the audio domain is vulnerable to replay attacks. In comparison, it is hard for an adversary to simultaneously forge the two nonlinearly related responses and the cross-domain physical relationships.

Built-In Speaker Replays: We assume a more challenging scenario when the attacker could compromise the device's built-in speaker and launch attacks at the right time, because such attacks could easily escape our cross-domain validity check, which is mainly designed to defend against the external sounds. We leave the attack implementations to the future work. Table IV presents the FAR and FRR of the system under such attacks. We find that the microphones do not show good performances to defend against the built-in speaker replays, but the accelerometer still achieves a high accuracy. In particular, Mic 1 and Mic 2 obtain 37.3% and 42.1% FAR to accept the built-in speaker replays. In comparison, the accelerometer alone achieves 4.9% FAR and 7.4% FRR. When combining the microphones and the accelerometer, our system achieves a 5.7% equal error rate. The ROC curves of the microphones, the accelerometer, and their combination are presented in Fig. 18, to compare their differences. The result shows that microphones failed to differentiate well the replay sounds from the live gripping hand. But the surface vibrations affected by the gripping hand are still unique and hard to be forged by the replay sounds. Based on the two domains, our system thus builds up enhanced security.

TABLE IV
UNDER BUILT-IN SPEAKER REPLAYS

Sensor	Mic1	Mic2	Mic1+2	Accel.	**Mics + Accel.**
FAR	0.373	0.421	0.313	0.049	**0.045**
FRR	0.124	0.097	0.088	0.074	**0.068**

3-D Fake Hand: We use a commodity 3-D scanner (i.e., Revopoint POP 2) to scan the user's hand model. Based on the reconstructed 3-D hand, we use a commodity 3-D printer (i.e., Creality CR-10S) and a soft material (i.e., thermoplastic polyurethane) to print the fake hand. Specifically, we scan and print two users' gripping hands and use them to attack our user authentication system. We find our system effectively defends against these fake hands with 100% accuracy. The results indicate that although the fake hands can reproduce the user's hand shape and hand geometry, they can hardly forge other features of the user's hand-grip biometric, such as the body-fat ratio, bone structure, gripping strength, etc.

VII. ADDRESSING PRACTICAL NOISE IMPACT

Since our system uses audible musical sounds to sense the user's gripping hand, it may be interfered by high-level ambient noises. More particularly, when using learning-based acoustic systems in practical scenarios, the testing data showing different noise levels from the training data may severely degrade the system's performance, because the system has not seen the sensing signals under the influence of such noises before. This is a critical challenge to prevent many acoustic sensing systems from being deployed practically. This section first illustrates this noise-incurred issue and demonstrates how we address this issue using cross-domain sensing. We further propose two training data augmentation methods to improve the system performance. To study the system's performance under noise, we use five different types of daily noise resources from YouTube, including office, air conditioning, conversation, vehicle, and busy traffic, and play these audios using a loudspeaker to jam the audible

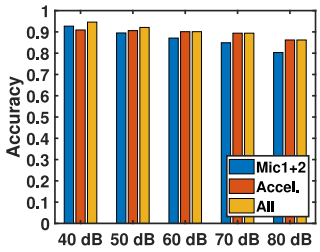


Fig. 19. Performance under different levels of noise.

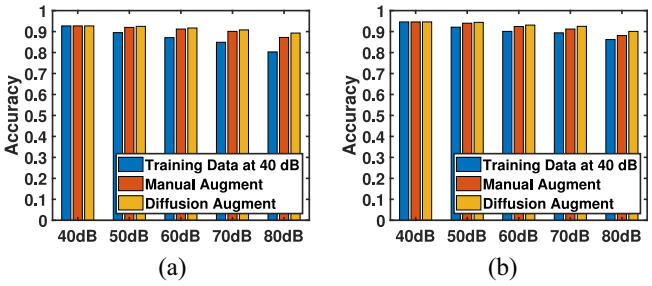


Fig. 20. Improving under-noise performance via extra training data collection and training data augmentation. (a) Mic. (b) Mic + Accel.

frequencies. We use the Decibel X App to make sure that they are played at their referenced noise levels from 40 to 80 dB [53] and then collect one user's data by playing the "iPhone Message" on S8 phone. For each noise level, 20 instances are collected in the first session, and 40 instances are collected in the second session. Testing only uses the second session data.

When Testing Data Has Different Noise Levels: Typically, a user enrolls the authentication system in relatively quiet indoor environments, such as office or home, with a noise level of around 40dB. We thus use the user data collected under 40dB noise for training (only the first session data). Because authentication may happen under any noise level, we use the second session user data collected from 40 to 80 dB for testing. Fig. 19 shows that the performance when using the microphone data alone decreases rapidly when the noise level increases. Specifically, when the acoustic noise increases from 40 to 70 dB and 80 dB, the verification accuracy using microphones drops from 92.7% to 84.9% and 80.3%. In comparison, the accelerometer data is robust to these high-level noises. When the acoustic noise increases from 40 dB to 60dB, the accelerometer still achieves over 90% accuracy. The reason is that the accelerometer is relatively isolated from the external sounds. When integrating the two domain data, our system achieves 95%, 89%, and 86% accuracy under 40, 70, and 80 dB noises, respectively. Please note that noises over 70 dB begin to damage hearing. Further improving the system performance under high-level noises requires solving the noise profile inconsistency challenge for learning-based acoustic sensing. We propose two training data augmentation methods. One relies on the user's effort to manually collect training data with different noises. The other utilizes a latent diffusion model to synthesize the training data with diverse noises.

Manual Data Augmentation: We collect the user data under different levels of noise to augment the training data set. Specifically, the first-session user data under 40, 50, 60, 70, and 80 dB noises are used for training. The updated model is then tested on the second-session data under these noises. Fig. 20(a) shows that with the extra training data, the performance achieved by using the microphone data alone is significantly improved. The system now achieves 92%, 91.2%, 90.1%, and 87.2% accuracy using the microphone data under 50, 60, 70, and 80 dB noises. Fig. 20(b) shows that when integrating the microphone and accelerometer data, our system achieves 94%, 92.4%, 91.2%, and 88.1% accuracy under the 40, 60, 70, and 80 dB noises, respectively. Though promising, the manual method requires high user efforts to collect training data under different noise levels.

Diffusion Model-Based Data Augmentation: We propose to leverage the state-of-the-art latent diffusion model to synthesize the training under different levels of noise. We develop the data augmentation method based on AudioLDM [54], which is a text-to-audio system built on the latent diffusion model and trained using the contrastive language-audio pretraining (CLAP) [55] embeddings. The original model is capable of generating audio for a given text prompt and transferring the style of a given audio. To use AudioLDM to synthesize the user data under different levels of noise, we first download different types of daily noises (with captions), including the above-mentioned five types of noise (e.g., from YouTube) and use them to fine-tune the pretrained AudioLDM model. In particular, the downloaded noise audios last around 30 min in total, which are used to fine-tune the AudioLDM model for 200 epochs, taking around 1 h on four A100 GPUs. Then, given the first-session user data with 40 dB noise (e.g., office), we use the fine-tuned AudioLDM model to transfer the real user data to generate synthetic user data with 50, 60, 70, and 80 dB noise levels, respectively. This is achieved by tuning the style transfer parameters of the AudioLDM model. The synthesized user data with different noise levels are then used as the augmented training data for our CNN-based system to recognize the user's hand in the unseen noisy environments.

The updated CNN model is then tested on the second-session data of different noise levels. Fig. 20(a) shows that with the diffusion-model augmented training data, the system's performance using the microphone data alone is further improved. Compared to the manual training data augmentation, the diffusion model improves the authentication accuracy to 92.5%, 91.7%, 90.8%, and 89.3% under 50, 60, 70, and 80 dB noises. As shown in Fig. 20(b), the verification accuracy by integrating the two domains is then increased to 94.4%, 93.1%, 92.5%, and 90.1% under 50, 60, 70, and 80 dB noises, respectively. The result indicates that our diffusion-based data augmentation solves the noise profile inconsistency challenge to improve acoustic sensing-based authentication without requiring additional user efforts.

VIII. Silent Mode Authentication

We further investigate the notification scenarios in circumstances when using musical tones may disturb people, such

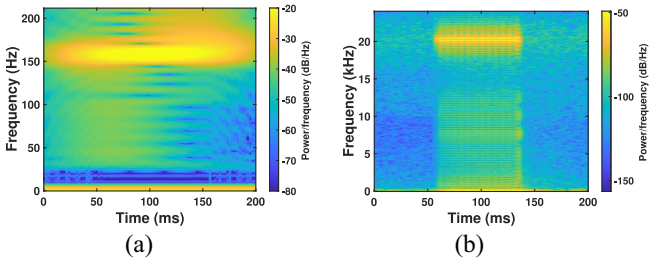


Fig. 21. Two responses of a vibration alert. (a) On accelerometer. (b) On microphone.

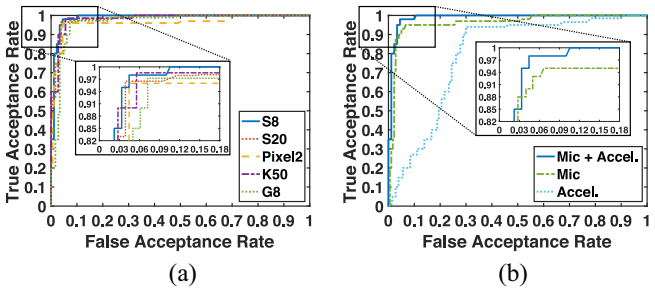


Fig. 22. User verification using vibration alerts. (a) Five device models. (b) Mic versus Accel. on S8.

as in meetings and museums. We note that in order to still be attentive to notifications, users choose vibrating alerts as notification signals [43]. By considering the vibration alert as a special musical sound, we extend our method to work in this silent/vibrate mode with only some parameter changes. The intuition is that sound and vibration co-exist, and we observe that smartphone microphones can capture the motor vibration sounds. We then derive the acoustic and vibration time–frequency images (e.g., spectrograms) of the vibration alert for authentication as shown in Fig. 21 (illustrated with the Samsung S8 phone's vibrations). We note that while the accelerometer mainly captures the phone's motor frequency at 157 Hz, the microphone captures the vibration sounds of up to 20 kHz, which contain the harmonics of the motor and the smartphone's surface vibrations. Thus, the microphone provides much richer information than the accelerometer to capture the user's hand responses to vibrations.

We directly apply our method to vibration alerts and distinguish 40 participants on five phone models, which are equipped with two types of motors. Specifically, Pixel 2 uses the linear resonant actuator (LRA) motor, while the other four use the eccentric rotating mass (ERM) motor. The motor frequencies are 157, 157, 155, 198, and 208 Hz for S8, S20, Pixel 2, K50 and G8, respectively. Only 150-ms vibration signals are used. Fig. 22(a) presents the user verification performances of the five devices when integrating the two domain responses. We find that all five phone models achieve over 95% accuracy. In particular, S8 performs the best with 96.8% accuracy. S20, Pixel 2, K50, and G8 achieve 96.3%, 96%, 96.5%, and 95% accuracy, respectively. The results also indicate that our method works well for both linear and rotation motors. Fig. 22(b) further studies the two responses based on S8. We find that the microphone performs better than the accelerometer with around 15% performance enhancement. The result confirms that the microphone is better than accelerometers in the vibration sensing, though the microphone has seldomly been used to capture vibrations in prior work. Additionally, we find vibration notifications perform better than musical tones. This is because musical sounds are more complex to analyze than single-frequency motor vibrations.

IX. Music Tone Analysis

To further study the capabilities of different music tones, we compute the frequency range, SNR, and the frequency occupation ratio for each tone signal to quantize its frequency richness and critical music events. The frequency range is the frequency span that the music tone signal covers. The SNR is the average tone signal power over the ambient noise power calculated within the tone's duration, which measures how strong the music signal is. The frequency occupation ratio is used to measure the frequency richness of a signal. We compute this ratio based on the spectrogram of the music tones as shown in Fig. 23, and a magnitude threshold above the noise level is selected. Then, within a sliding time window, the ratio is computed as the number of spectral points whose magnitudes are above the threshold over the total number of spectral points within the window. The final frequency occupation ratio is averaged through the entire tone period. Equation (9) expresses the computation of this ratio, where N is the number of sliding windows, and P is the total number of spectral points within the time window. G is the number of points whose magnitudes are greater than the threshold in the current time window

$$FreOR = \frac{1}{N}\sum_{T=1}^{N}\frac{G_T}{P_T}. \tag{9}$$

This frequency occupation ratio thus reveals the energy distribution of the music tone spectrogram.

We now present the analytical results of the nine music tones and their performances in Table V. We find that the music tones with higher frequency occupation ratios and SNRs tend to achieve higher user identification accuracies. The frequency range span of the music tone does not have an obvious impact on the performance. For example, "iPhone Messag," "iPhone Rin," "Nokia Messag," "Nokia Ring," and "Huawei Tone" achieve an accuracy of around 94%. They have frequency occupation ratios between 0.52 and 0.87 and the signal SNRs between 33.6 and 37.8 dB. The "iPhone Note" has a low frequency occupation ratio of 0.28 and a low signal SNR of 26.9 dB. It shows the lowest accuracy 90.4%.

While we do not find notable impact by the tones' frequency ranges, we take a step further to divide the speaker's frequency range into three intervals, which describe the low-frequency components (0–1 kHz), the med-frequency components (1–10 kHz), and the high-frequency components (above 10 kHz). The nine musical sounds are then analyzed regarding these frequency intervals. As we can observe in Table V, nearly

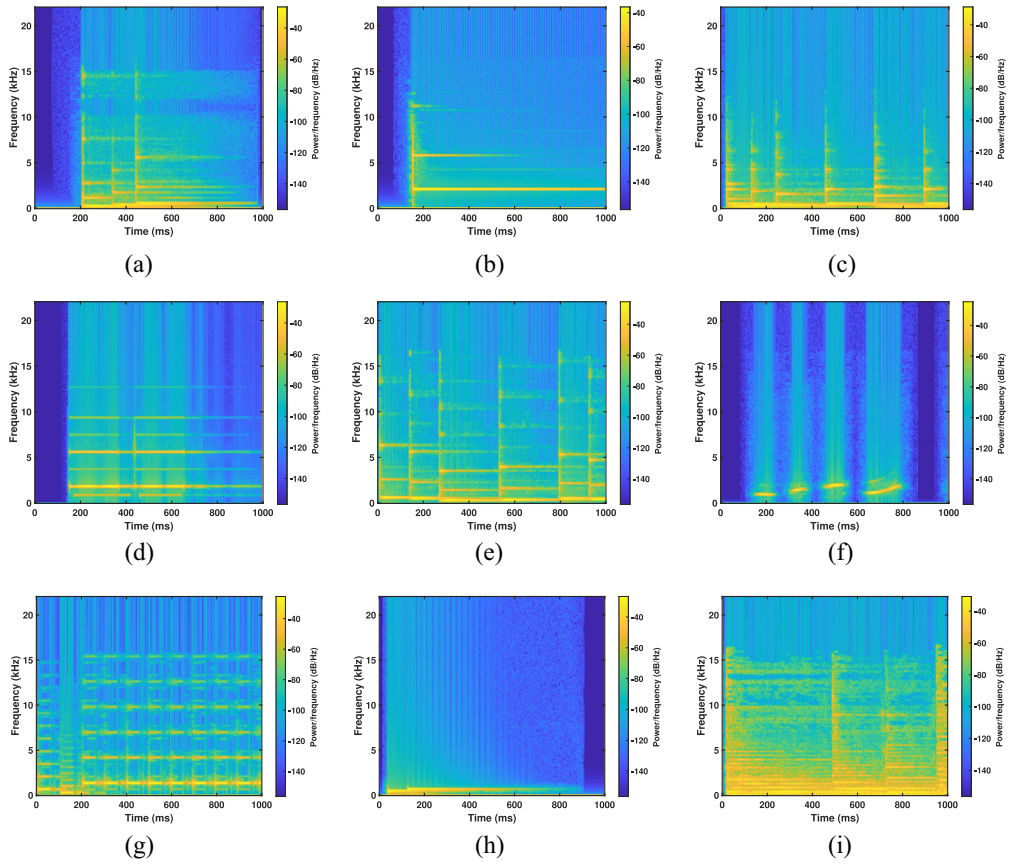


Fig. 23. Spectrogram of the music tones. (a) iPhone message. (b) iPhone note. (c) iPhone ring. (d) Nokia message. (e) Nokia ring. (f) Samsung whistle. (g) Samsung basic bell. (h) Gmail notification. (i) Huawei tone.

TABLE V
ANALYSIS OF THE MUSICAL TONES

Musical Tone	**Freq. Range** (kHz)	**Freq. Occu. Ratio**	**SNR** (dB)	**Accuracy**	**Low-Freq. Comp. (0-1kHz)**	**Medium-Freq. Comp. (1kHz-10kHz)**	**High-Freq. Comp. (above 10kHz)**
iPhone Message	0-15	0.87	34.7	0.946	✓	✓	✓
iPhone Note	2-10	0.28	26.9	0.904		✓	
iPhone Ring	0-10	0.52	33.6	0.932	✓	✓	
Nokia Message	0.5-9	0.77	34.3	0.941	✓	✓	
Nokia Ring	0.25-17	0.68	37.8	0.943	✓	✓	✓
Samsung Whistle	1-3	0.35	31.3	0.912		✓	
Samsung Basic Bell	1.2-16	0.55	34.3	0.917		✓	✓
Gmail Notification	0.15-1	0.72	32.3	0.915	✓		
Huawei Tone	0-17	0.84	35.0	0.938	✓	✓	✓

all the tones, except the "Gmail Notification," contain mid-frequency components. In particular, the "iPhone Note" and "Samsung Whistle" only have mid-frequency components. They achieve the lowest performances. The "Gmail Notification" tone only has the low-frequency components, which performs slightly better. None of the nine tones use high-frequency components alone. The reason is that the sounds of lower frequencies have stronger signal power and generate stronger vibrations, which present greater sensing capabilities. Furthermore, the tones that can achieve over 93% accuracy all contain two frequency components, which are the low-frequency and the mid-frequency components. The best-performed tones, "iPhone Message," "Nokia Ring" and "Huawei Tone" all have three frequency components. The results indicate that the music tones with rich frequency features and high signal powers (at least sufficiently high at relatively low frequencies) are good for sensing.

X. DISCUSSION

A. Key Takeaways From Experimental Evaluations

This work successfully demonstrates the improved robustness of hand-based smartphone authentication brought by cross-domain sensing. In particular, the deeply mixed vibrations and sounds resulting from the common media sounds of smartphones capture the user's hand biometrics in two different sensing domains and are hard to be replicated by an adversary. Experiments show that our cross-domain sensing system achieves a high-accuracy in verifying smartphone users, and nine different notification sounds achieve 90%–95%

accuracy with two domain's sensing information. If looking at each sensing domain alone, some sounds generating strong vibrations are better at vibration sensing, while some sounds are better at acoustic sensing. Additionally, the comparison of four different time–frequency analysis methods show that the scalogram and the spectrogram perform well in describing cross-domain sensing, while the scalogram performs the best. When working with multiple accounts or enrolled users, the system has a 1% and 2% accuracy loss compared to the single-user scenario.

Although using common media sounds is subject to many ambient noises, our results show that the cross-domain sensing and the inclusion of the sensors, microphones and speakers within the phone shell have the capability to address the ambient sound impact. To further suppress acoustic noises, we successfully use a diffusion model to simulate the sensing under different types/levels of acoustic noises and augment the training data to improve the system robustness, showing only around 0.5% accuracy loss to 50–80 dB ambient noises, compared to the typical office scenario.

The proposed system also achieves high performance with different mobile device models, the varying speaker volumes from 50% to 100%, and with different phone cases. We further adjust the system to accommodate users' silent-mode demands. By using the vibration alert as a special tone, our cross-domain sensing performs with up to 2% accuracy increase compared to media sounds, owing to its reduced susceptibility to ambient noise.

At last, it is important to recognize the security gain achieved by cross-domain sensing, which successfully mitigates 100% external acoustic replays, which are threats to all acoustic systems (e.g., 10% FAR when using the acoustic domain of our system alone), and physical replay attacks. The experiments demonstrate the difficulty in launching replay attacks to simultaneously attack both microphones and inertial sensors, since sounds and vibrations co-exist.

B. Limitations and Future Work

As the first musical sound-based user authentication work, our current method has several limitations that need further explorations. The first issue is the tone-dependent verification. Changing notification tones would require additional efforts to retrain the model. But for the vibrating alerts in the silent mode, there is no such concern, as the motor frequency is fixed on current smartphones. Additionally, if the user likes to switch hands to use the smartphone, both hands need to be trained, though this is not a problem for the CNN model. For future improvement, we consider two possible solutions to minimize the training efforts for using more tones and hands. One is transfer learning, and the other is the training data augmentation by adding specially designed noises. We leave this to future work.

This work shows the potential of using musical sounds for sensing. There is still room to further improve the musical sound-based sensing by developing more advanced deep learning algorithms, which may better tolerate the users' behavioral inconsistency, background noises, and replay attacks. Moreover, we have only exploited using notification sounds for sensing. It is worth studying whether more general media sounds, such as human voices can be also used for sensing or not. In addition, the vibration alert is viewed as a special type of notification sound. Future work will include combining both speaker sound and motor vibration as the stimulus signal to provide more robust authentication. Furthermore, potential threats to our system need to be investigated, including new adversarial learning algorithms to forge the acoustic and vibration responses and their physical relationships simultaneously, how to compromise built-in speakers to forge the cross-domain relationships, and how to ensure the replay sound is played at the right time. In addition, a larger-scale study with more participants is needed to further improve our method in practical scenarios.

We consider more notification scenarios in a user's daily life and discuss the capability or limitation of the proposed method to cope with them.

1) When the phone is on a shared table, our method does protect privacy by hiding notification previews as the phone would be detected to be in the nonhand scenario as shown in Section VI-E.
2) If the user holds the phone to share the screen with others, the effectiveness of our method depends on how the user grips the phone. If the other people are on the opposite side, the user's hand-grip should not be the same, and our method works. If they are on the same side as the user, our method may not work well unless the user grips the phone differently, consciously or unconsciously.
3) If the user wears a glove, he or she needs to include this scenario in the training data to use our method, which requires additional training efforts.
4) When the user is moving while using her/his phone, additional noises are generated in both the acoustic and vibration domains. We find that the footstep sounds have little impact on the acoustic responses because they are external sounds and their SNRs are low. The body movements and hand vibration noises can be removed from the vibration responses by a high-pass filter. But the contact relationship between the palm and the phone may change slightly depending on how firmly the user grips the phone while walking. The incurred behavioral inconsistency needs to be further studied and addressed to enable the use of our method for pedestrians.
5) Other impact factors, including moisture, lotions, and moods, also need to be further studied. We believe that our method has a stronger capability to cope with these scenarios as it leverages the sensing information across the acoustic and vibration domains.

XI. Conclusion

This work proposes a low-effort user authentication method for handheld devices using musical sounds, which serves as an external way of access control and can be used to protect the user's notification privacy while keeping full notification features. The proposed method protects the user's notification

privacy in both silent and nonsilent device modes by directly using the notification signals (such as musical tones and vibrating alerts) as the stimulus signal. We show that the handheld devices' musical sounds, though more complicated than dedicated signals, can be directly used for sensing and verifying the user's gripping hand. Moreover, we find that both musical tones and vibrating alerts generate strong acoustic and vibration responses, which can be used to address the acoustic noises and attacks that threaten all acoustic systems. In particular, we derive time–frequency images to describe the people's gripping hand biometrics in two domains and develop a CNN-based algorithm for user authentication. We further derive the unique cross-domain physical relationships among the smartphone's microphone, speaker, and accelerometer, which are all embedded on the same motherboard, to prevent external sounds. We show that an adversary is hard to cheat our system across two domains even by taking control of the target device's built-in speaker or reproducing the user's 3-D hand. In addition, we improve the system's performance under different noise levels, even when the testing data have different noise profiles. We utilize a latent diffusion-based audio generation model for training data augmentation, which synthesizes the sensing data in different ambient noises. Extensive experiments show that our system verifies the user with around 94.5% accuracy and effectively defends against acoustic replays and physical hand forgeries.

REFERENCES

[1] M. Eiband, M. Khamis, E. Von Zezschwitz, H. Hussmann, and F. Alt, "Understanding shoulder surfing in the wild: Stories from users and observers," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2017, pp. 4254–4265.

[2] M. Espinoza, C. Champod, and P. Margot, "Vulnerabilities of fingerprint reader to fake fingerprints attacks," *Forensic Sci. Int.*, vol. 204, nos. 1–3, pp. 41–49, 2011.

[3] N. Erdogmus and S. Marcel, "Spoofing face recognition with 3D masks," *IEEE Trans. Inf. Forensics Security*, vol. 9, pp. 1084–1097, 2014.

[4] C. Hoffman. "How to hide sensitive notifications from your iPhone's lock screen." 2017. [Online]. Available: https://www.howtogeek.com/252483/how-to-hide-sensitive-notifications-from-your-iphones-lock-screen/

[5] A. Bouchard. "Improve your iPhone's notification privacy with blurification." 2019. [Online]. Available: https://www.idownloadblog.com/2019/05/07/blurification/.

[6] L. Mannering (New York Times, New York, NY, USA). *How to Not Ruin Your Life (or Just Die of Embarrassment) With a Screen Share*. 2019. [Online]. Available: https://www.nytimes.com/2019/03/21/style/screen-share-privacy-tips.html.

[7] S. Pushp, Y. Liu, M. Xu, C. Koh, and J. Song, "PrivacyShield: A mobile system for supporting subtle just-in-time privacy provisioning through off-screen-based touch gestures," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 2, pp. 1–38, 2018.

[8] J. Lindberg and M. Blomberg, "Vulnerability in speaker verification-a study of technical impostor techniques," in *Proc. 6th Eur. Conf. Speech Commun. Technol.*, 1999, pp. 1–4.

[9] 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.

[10] S. Uellenbeck, M. Dürmuth, C. Wolf, and T. Holz, "Quantifying the security of graphical passwords: The case of android unlock patterns," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, 2013, pp. 161–172.

[11] R. Dhamija and A. Perrig, "Déjà Vu: A user study: Using images for authentication," in *Proc. 9th USENIX Security Symp. (USENIX Security)*, 2000, pp. 1–4.

[12] C. Tsikos, "Capacitive fingerprint sensor," U.S. Patent 4 353 056, Oct. 5, 1982.

[13] J. Chamary. "How face ID works on iPhone X." 2019, https://www.forbes.com/sites/jvchamary/2017/09/16/how-face-id-works-apple-iphone-x/.

[14] J. L. Cambier and J. E. Siedlarz, "Portable authentication device and method using iris patterns," U.S. Patent 6 532 298, Nov. 2003.

[15] A. de Santos Sierra, J. G. Casanova, C. S. Avila, and V. J. Vera, "Silhouette-based hand recognition on mobile devices," in *Proc. 43rd Annu. Int. Carnahan Conf. Security Technol.*, 2009, pp. 160–166.

[16] M. Choraś and R. Kozik, "Contactless palmprint and knuckle biometrics for mobile devices," *Pattern Anal. Appl.*, vol. 15, no. 1, pp. 73–85, 2012.

[17] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[18] Y. Ren, Y. Chen, M. C. Chuah, and J. Yang, "User verification leveraging gait recognition for smartphone enabled mobile healthcare systems," *IEEE Trans. Mobile Comput.*, vol. 14, no. 9, pp. 1961–1974, Sep. 2014.

[19] N. Zheng, K. Bai, H. Huang, and H. Wang, "You are how you touch: User verification on Smartphones via tapping Behaviors," in *Proc. ICNP*, 2014, pp. 221–232.

[20] M. Frank, R. Biedert, E. Ma, I. Martinovic, and D. Song, "Touchalytics: On the applicability of touchscreen input as a behavioral biometric for continuous authentication," *IEEE Trans. Inf. Forensics Security*, vol. 8, pp. 136–148, 2013.

[21] J. Chauhan, Y. Hu, S. Seneviratne, A. Misra, A. Seneviratne, and Y. Lee, "BreathPrint: Breathing acoustics-based user authentication," in *Proc. 15th Annu. Int. Conf. Mobile Syst., Appl., Services*, 2017, pp. 278–291.

[22] J. Liu, Y. Chen, Y. Dong, Y. Wang, T. Zhao, and Y.-D. Yao, "Continuous user verification via respiratory biometrics," in *Proc. INFOCOM IEEE Conf. Comput. Commun.*, 2020, pp. 1–10.

[23] J. da Silva Dias, I. Traore, V. G. Ferreira, and J. David, "Exploratory use of PPG signal in continuous authentication," in *Proc. Brazilian Symp. Inf. Comput. Syst. Security*, 2015, pp. 142–155.

[24] W. Louis, M. Komeili, and D. Hatzinakos, "Continuous authentication using one-dimensional multi-resolution local binary patterns (1DMRLBP) in ECG biometrics," *IEEE Trans. Inf. Forensics Security*, vol. 11, pp. 2818–2832, 2016.

[25] I. Nakanishi, S. Baba, and C. Miyamoto, "EEG based biometric authentication using new spectral features," in *Proc. Int. Symp. Intell. Signal Process. Commun. Syst. (ISPACS)*, 2009, pp. 651–654.

[26] W. Khalifa, A. Salem, M. Roushdy, and K. Revett, "A survey of EEG based user authentication schemes," in *Proc. 8th Int. Conf. Inform. Syst. (INFOS)*, 2012, pp. BIO–55.

[27] A. Mehrotra, R. Hendley, and M. Musolesi, "PrefMiner: Mining user's preferences for intelligent mobile notification management," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2016, pp. 1223–1234.

[28] S. Lee, S. Pushp, C. Min, and J. Song, "Exploring relationship-aware dynamic message screening for mobile messengers," in *Proc. ACM Int. Joint Conf. Int. Symp. Pervasive Ubiquitous Comput. Wearable Comput.*, 2018, pp. 134–137.

[29] J. Evans. "How to use guided access to secure your iPad or iPhone." 2017, https://www.computerworld.com/article/3162738/how-to-use-guided-access-to-secure-your-ipad-or-iphone.html.

[30] Google. "Supporting multiple users." 2020. [Online]. Available: https://source.android.com/devices/tech/admin/multi-user.

[31] A. S. Weksler, N. J. Peterson, and R. S. VanBlon, "Grip signature authentication of user of device," U.S. Patent 14 098 180, Nov. 2015.

[32] M. Ota, Y. Morinaga, M. Tsukamoto, and T. Higuchi, "Portable terminal and gripping-feature learning method," U.S. Patent 13 881 386, May 2013.

[33] H. Chen, F. Li, W. Du, S. Yang, M. Conn, and Y. Wang, "Listen to your fingers: User authentication based on geometry biometrics of touch gesture," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 4, no. 3, pp. 1–23, 2020.

[34] Y. Yang, C. Wang, Y. Chen, and Y. Wang, "EchoLock: Towards low effort mobile user identification," 2020, *arXiv:2003.09061*.

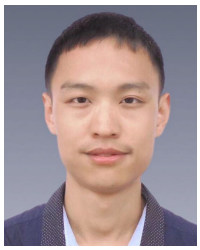
[35] Z. Chen and M. Recce, "Handgrip recognition," *J. Eng., Comput. Archit.*, vol. 1, no. 2, pp. 1–11, 2007.

[36] C. J. Migos and D. H. Sloo, "Personalization using a hand-pressure signature," U.S. Patent 8 172 675, Aug. 2012.

[37] S. Dey, N. Roy, W. Xu, R. R. Choudhury, and S. Nelakuditi, "AccelPrint: Imperfections of accelerometers make smartphones trackable," in *Proc. Netw. Distrib. Syst. Security Symp. (USENIX NDSS)*, 2014.

[38] A. Das, N. Borisov, and M. Caesar, "Do you hear what i hear? Fingerprinting smart devices through embedded acoustic components," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, 2014, pp. 441–452.

[39] SkyPaw. "Decibel X." 2020. [Online]. Available: http://skypaw.com/decibelx.html
[40] C. Wang, S. A. Anand, J. Liu, P. Walker, Y. Chen, and N. Saxena, "Defeating hidden audio channel attacks on voice assistants via audio-induced surface vibrations," in *Proc. 35th Annu. Comput. Security Appl. Conf.*, 2019, pp. 42–56.
[41] S. A. Anand, C. Wang, J. Liu, N. Saxena, and Y. Chen, "SpearPhone: A speech privacy exploit via accelerometer-sensed reverberations from smartphone loudspeakers," 2019, *arXiv:1907.05972*.
[42] Y. Michalevsky, D. Boneh, and G. Nakibly, "Gyrophone: Recognizing speech from gyroscope signals," in *Proc. 23rd USENIX Security Symp. (USENIX Security)*, 2014, pp. 1053–1067.
[43] Y.-J. Chang and J. C. Tang, "Investigating mobile users' ringer mode usage and attentiveness and responsiveness to communication," in *Proc. 17th Int. Conf. Human-Comput. Interact. Mobile Devices Services*, 2015, pp. 6–15.
[44] S. W. MacMASTER, "Privacy screen for a display," U.S. Patent 7 052 746, May 2006.
[45] mortisApps. "Screen guard privacy screen." 2013. https://play.google.com/store/apps/details?id=com.screen.guard.screenfilter.hidescreen&hl=en_US.
[46] HueySoft. "Privacy screen filter." 2019. [Online]. Available: https://play.google.com/store/apps/details?id=com.hueysl.privacyscreen_paid&hl=en_US
[47] C.-Y. Chen, B.-Y. Lin, J. Wang, and K. G. Shin, "Keep others from peeking at your mobile device screen!" in *Proc. 25th Annu. Int. Conf. Mobile Comput. Netw.*, 2019, pp. 1–16.
[48] M. Xu, J. Liu, Y. Liu, F. X. Lin, Y. Liu, and X. Liu, "A first look at deep learning apps on smartphones," in *Proc. World Wide Web Conf.*, 2019, pp. 2125–2136.
[49] N. Carlini et al., "Hidden voice commands," in *Proc. 25th USENIX Security Symp. (USENIX Security)*, 2016, pp. 513–530.
[50] T. Vaidya, Y. Zhang, M. Sherr, and C. Shields, "Cocaine noodles: Exploiting the gap between human and machine speech recognition," in *Proc. 9th USENIX Workshop Offensive Technol. (WOOT)*, 2015, pp. 1–16.
[51] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *Proc. IEEE Security Privacy Workshops (SPW)*, 2018, pp. 1–7.
[52] *Spectrum Occupancy Measurements and Evaluation*, ITU-Rec. SM.2256-1, Int. Telecommun. Union, Geneva, Switzerland, 2017.
[53] Engineering Toolbox. "Sound pressure." 2004, https://www.engineeringtoolbox.com/sound-pressure-d_711.html
[54] H. Liu et al., "AudioLDM: Text-to-audio generation with latent diffusion models," 2023, *arXiv:2301.12503*.
[55] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2023, pp. 1–5.
[56] L. Huang and C. Wang, "Notification privacy protection via unobtrusive gripping hand verification using media sounds," in *Proc. 27th Annu. Int. Conf. Mobile Comput. Netw.*, 2021, pp. 491–504.



Long Huang (Student Member, IEEE) received the B.Eng. degree in electrical and electronic engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2015, and the M.Sc. degree in electrical and electronic engineering from Stevens Institute of Technology, Hoboken, NJ, USA, in 2019. He is currently pursuing the Ph.D. degree with Southern Methodist University, Dallas, TX, USA.

His research interests include mobile computing, signal processing, and Internet of Things.

Mr. Huang received the Best Paper Award from the 14th International Workshop on Wireless Sensor, Robot and UAV Networks in 2021.

Chen Wang (Member, IEEE) received the Ph.D. degree from Rutgers University, New Brunswick, NJ, USA, in 2019.

He is an Associate Professor of Computer Science with SMU Lyle, Dallas, TX, USA, where he leads the Mobile and Internet Security Laboratory. He has published a number of papers at high-impact conferences, including IEEE S&P, ACM CCS, ACM Mobicom, and IEEE Infocom. His research interests include cyber security and privacy, sensing, mixed reality, robotics security, and smart healthcare.

Dr. Wang has received five the Best Paper Awards from IEEE CNS 2018, IEEE CNS 2014, ASIACCS 2016, EAI HealthyIoT 2019, and IEEE INFOCOM WKSHPS 2021. He is the recipient of the NSF CAREER award. From 2014 to 2023, his research studies have been reported by over 170 media outlets, including *IEEE Spectrum*, *NSF Science 360*, CBS TV, BBC News, NBC, *IEEE Engineering 360*, *Fortune*, ABC News, and *MIT Technology Review*.