RESEARCH-ARTICLE

Notification privacy protection via unobtrusive gripping hand verification using media sounds

LONG HUANG, Louisiana State University, Baton Rouge, LA, United States

CHEN WANG, Louisiana State University, Baton Rouge, LA, United States

Published: 25 October 2021

Citation in BibTeX format

ACM MobiCom '21: The 27th Annual International Conference on Mobile Computing and Networking
October 25 - 29, 2021
Louisiana, New Orleans

Conference Sponsors:
SIGMOBILE

Notification Privacy Protection via Unobtrusive Gripping Hand Verification Using Media Sounds

Long Huang
Louisiana State University
Baton Rouge, LA 70803
lhuan45@lsu.edu

Chen Wang
Louisiana State University
Baton Rouge, LA 70803
chenwang1@lsu.edu

ABSTRACT

This work proposes a media sound-based authentication method to protect smartphone notification privacy unobtrusively, which wisely hides or presents sensitive content by verifying who is holding the phone. We show that media sounds, such as the melodies of notification tones (e.g., iPhone message and Samsung whistle) can be directly used to sense and verify the user's gripping hand. Because sounds and vibrations co-exist, we capture two novel responses via the smartphone mic and accelerometer to describe how the individual's contacting palm interferes with the signals in two different domains. Based on the two responses, we develop a convolutional neural network-based algorithm to verify the user. Moreover, because the smartphone sensors are all embedded on the same motherboard, we develop a cross-domain method to validate such hard-to-forge physical relationships among the mic, speaker and accelerometer. They prevent external sounds from cheating the system. Additionally, we consider the notification vibration as a special type of media sound, which also results in two responses, and extend our method to work in the silent mode. Extensive experiments with ten notification tones and four phone models show that our system verifies users with 95% accuracy and prevents replay sounds with 100% accuracy.

CCS CONCEPTS

• **Security and privacy** → **Authentication**; **Biometrics**.

KEYWORDS

Notification Privacy, User Authentication, Gripping Hand

ACM Reference Format:
Long Huang and Chen Wang. 2022. Notification Privacy Protection via Unobtrusive Gripping Hand Verification Using Media Sounds. In *The 27th Annual International Conference on Mobile Computing and Networking (ACM MobiCom '21), January 31-February 4, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3447993.3483277

1 INTRODUCTION

Smartphones have become our most intimate devices. While providing a multitude of services anytime and anywhere, they also access the user's private data. Though a number of authentication methods (e.g., PINs, fingerprints and face IDs) have been deployed for access control, smartphone notification is an exception, which bypasses these authentications and automatically displays incoming messages on the screen (even locked), which could leak sensitive content. For example, when the phone is in others' hands or left unattended, an inappropriately displayed notification may expose the user's privacy in front of others and cause embarrassing moments, anxieties, financial losses and distrust [3, 22, 31, 41]. The reason is that notification mechanisms are designed to alert users of timely information and allow managing notifications of all Apps in the notification center without the trouble of entering credentials to unlock the phone and log into each App [23], which only considers convenience.

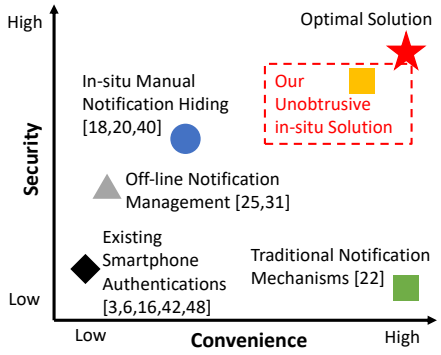


Figure 1: Comparison of notification solutions.

To address notification security, mobile operating systems allow turning off some Apps' notifications via off-line configurations [26, 33]. But this method is tedious and has a limited security improvement. To balance convenience and security, iOS introduces the *Guided Access* feature [19], which allows users to restrict the phone usage to a single App and disable all notifications before handing the device to others. Android supports setting up multiple accounts via different login passwords, so that the phone can be switched to a guest mode to hide the host data when needed [21]. However, these in-situ methods impose the responsibility on the user to decide when to display or hide notifications, and the series of on-screen operations are not convenient and even awkward in front of an observer. A study reports that 76% of smartphone users are not satisfied with current notification security [41]. This is confirmed by our recent survey in 2020 (in Section 2.2), which shows that 72.1% of smartphone users want an unobtrusive and easy-to-use solution.

This work aims to unobtrusively verify the user before displaying notifications to enable both full notification features and privacy

ACM MobiCom '21, January 31-February 4, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-8342-4/22/01…$15.00
https://doi.org/10.1145/3447993.3483277

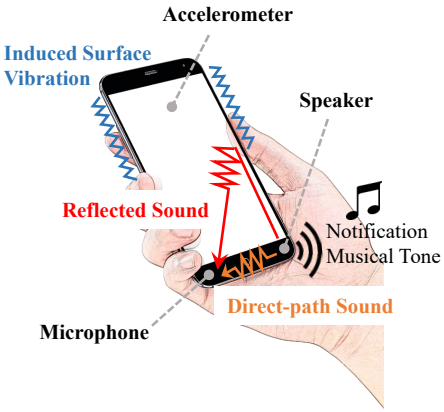


Figure 2: A notification tone interacting with the gripping hand in acoustic and vibration domains.

preservation. The comparisons to existing solutions regarding convenience and security are shown in Figure 1. We propose to directly use the notification signals (i.e., tones and vibrations), which are designed to alert the user, to recognize who is gripping the phone. We find that any media sounds (e.g., music) have the ability to sense, and thus, there is no need to design specialized signals and modify existing audio systems. Specifically, the notification sound (e.g., iPhone message and Gmail tones) traveling on the phone surface would interact with the user's gripping hand before reaching the smartphone mic as shown in Figure 2. Since people's hand geometries and gripping strengths are unique, they exert different effects on absorbing and reflecting the original signal. The resulted sound thus carries the user's hand-grip biometric. Furthermore, we note that sounds and vibrations co-exist, which enables analyzing the gripping hand's impact in both acoustic and vibration domains. Only smartphone media sounds (not available for ultrasounds) are strong enough to generate observable vibration signals. This novel cross-domain attribute provides the enhanced robustness and security, making our system immutable to both acoustic noises and attacks [27, 56].

We design a Convolutional Neural Network (CNN)-based system to verify people's gripping hands based on notification tones. The recordings of the phone's microphones and accelerometer are the input, and we derive their spectrograms to describe the user's hand-grip biometric in both acoustic and vibration domains. Then, a CNN-based algorithm with five convolutional layers distinguishes users in two domains. Their results are integrated by a cluster-based method to achieve robust user authentication even under high noise levels. Furthermore, we examine the physical relationship between microphones and accelerometer based on 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 silent mode using the notification vibration signals for sensing.

Our major contributions are summarized as:

- We address the mobile notification security issues by introducing a low-effort in-situ user authentication system based on notification tones.
- We are the first to show that daily media sounds are sufficient to sense and verify the user's hand unobtrusively, which saves the efforts of designing dedicated signals and modifying notification audio systems. Moreover, we provide a solution with two novel responses to address acoustic noises and attacks that threaten all acoustic systems.
- We derive spectrograms to describe users' hand-grip biometrics in acoustic and vibration domains and develop a CNN-based algorithm for verification. Furthermore, we design a cross-domain method to validate each acoustic input by leveraging the hard-to-forge physical relationship among the smartphone's mic, speaker and accelerometer.
- We show that vibration notifications also generate responses in two domains and thus extend our system to work for the smartphone's silent mode.

2 BACKGROUND AND TWO RESPONSES

2.1 Notification Threats

2.1.1 Sensitive Notification Types. Notification systems have been used by a broad range of mobile Apps to interact with users, including emails, text messages, calendar reminders, social media, mobile payment and healthcare apps. The average US smartphone user receives 46 notifications daily in 2019 [39]. Accordingly, a notification may contain sensitive information about the user, such as social relationships, demographics, online accounts, personal schedules, mobile transactions, health status, hobbies and classified correspondences. A survey finds that 29% of the daily notifications contain sensitive private information [36]. We use **user privacy** to refer to this type of sensitive information. As a relatively recent application, **verification code** relies on Short Message Service (SMS) to provide two-factor authentications, especially when an online account is accessed from a new device or browser [2]. However, an adversary, who obtains the user's smartphone and knows the online account passwords, can easily read the SMS verification codes from the locked screen to break the two-factor authentication. Turning off message notifications on the locked screen prevents such leakages, but the user also loses the ability to manage messages in the notification center and has to unlock the phone to check every message, which is not convenient.

2.1.2 Threat Scenarios. We investigate the scenarios when notifications are under threat and divide them into two categories. In either scenario, the phone screen could be locked or unlocked. 1) **Inadvertent leakage** refers to the scenario when the phone is given to others or left unattended, and the on-screen notifications are accidentally observed by others. User privacy is the major concern in this scenario. 2) **Notification snooping** refers to the scenario when an adversary intentionally snoops on the user's notification by stealing or temporarily obtaining the phone. Both the user privacy and verification codes are the concerns. For the verification code snooping, the adversary can further **actively trigger SMS notifications** via logging into the user's online accounts rather than passively waiting.

2.2 Our Online Survey Statistics.

We conducted an online survey to study smartphone users' privacy concerns in regards to notifications. For a nine-month period, we

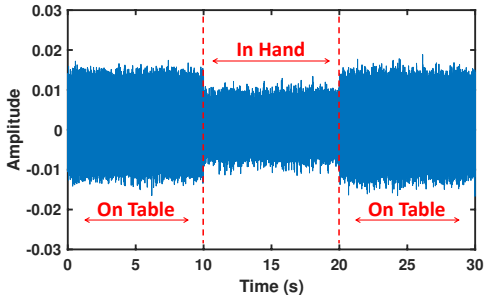
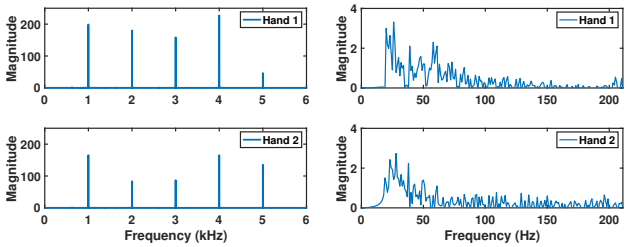


Figure 3: Impact of gripping hand on phone sounds (the data during grabbing and dropping are removed for simplicity).



(a) Acoustic responses (b) Vibration responses

Figure 4: Illustration of the acoustic and vibration responses of two gripping hands.

received 111 respondents, of which 92.8% are between ages 16 to 45 and 22.5% are female. The statistics show that 73.9% of respondents have privacy concerns for smartphone notifications, while 9.9% are not sure. Only 16.2% have no such concerns, and their three major reasons are "I never share my screen or lend my smartphone to others", "I can handle the privacy leakage myself" and "Existing notification privacy solutions are enough". Moreover, we find that 75% (51/68) of iPhone users and 65.2% (30/46) of Android users have not heard of their phones' security features to protect notification privacy or don't know how to use them. In comparison, 72.1% of respondents want a low-effort solution that could manage their notifications smartly according to situations. To summarize, our survey reflects: 1) Notification privacy has long been underestimated or neglected by mobile OS providers but not smartphone users; 2) Current solutions are tedious and intrusive, which are not adopted well by users; 3) Most users are aware of protecting notification privacy and want an easy-to-use solution.

2.3 Intuition of Using Musical Tones to Verify Gripping Hands

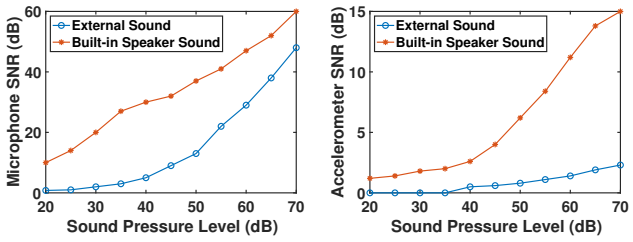
2.3.1 Hand-grip Biometric. The hand-grip biometric describes how uniquely a user grips a handheld device, which is closely associated with people's hand geometries and gripping behaviors [12, 35, 53]. Based on an array of pressure sensors attached to the handheld device surface (e.g., piezoelectric materials), 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. But obtaining such a biometric with smartphones requires high installation overhead. Our work shows that a user's hand-grip biometric can be captured by the phone's media sounds. 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 recorded sound is presented in Figure 3. We find that the gripping hand imposes a significant impact on the sounds, which shows the potential of analyzing the impact to distinguish gripping hands.

To describe the hand-grip biometric, we extract two unique responses of a hand to media 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 the hardware signature information of the smartphone [15, 17]. Our work utilizes smartphone speaker to transmit media sounds and its microphone and accelerometer as receivers, and the obtained two responses for analyzing the hand-grip biometric characteristics are also device-dependent.

2.3.2 Acoustic Response. When a notification 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 Figure 2. In particular, the sound is damped and reflected by the gripping hand, and the resulting signals at the smartphone mic 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 notification 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.

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 multi-path signals at each frequency are combined uniquely at the microphone. Therefore, each frequency signal carries one aspect of the hand-grip biometric. The feasibility is shown in Figure 4, where the frequency responses of the above five-frequency beep sounds are presented for two users' hands. The signals exhibit distinct patterns between the two users in the frequency domain, and 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 tone to obtain the comprehensive acoustic representation of the user's hand-grip biometric.

Internal vs. External. Acoustic noises and attacking sounds are typical issues for all acoustic systems. We note that these sounds are mainly from external sources and thus investigate the acoustic responses to the built-in and external speaker sounds. We find that the 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 40dB noises. The external speaker is placed 20cm away from the smartphone microphone. The Decibel X



(a) Acoustic responses (b) Vibration responses

Figure 5: Comparison of the acoustic and vibration responses under built-in and external speaker sounds.

App [45] is installed on a third device to measure the sound pressure levels near the microphone. As shown in Figure 5 (a), for the same loudness sound, the built-in speaker generates the microphone readings with 10~35dB greater SNRs than the external speaker. For example, the built-in speaker at the 70dB volume achieves 60dB SNR at the microphone, which is 15dB 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.

2.3.3 Vibration Response. We find that the notification sound also induces the device surface to vibrate at the same frequencies. The observation is consistent with the prior work that uses human speech audios [51]. The gripping hand suppresses the surface vibrations at the palm contact area determined by the hand geometry. Moreover, the greater gripping strength causes the higher resistance to vibrations. Thus, the same notification sound results in distinctive surface vibrations for different gripping hands, which can be captured by an accelerometer. We define the accelerometer-captured notification 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 [1]. Thus, to obtain vibration responses, the low-frequency audible sounds need to be used, among which musical tones are unobtrusive.

Non-linear Response. Though induced by the notification sound, the vibration-level representation of the hand-grip biometric is very different. Specifically, the accelerometer data has a non-linear relationship with the real surface vibrations due to the low sampling rate as described by

$$f_a = |f_v - N f_s|, N \in Z, \quad (1)$$

where, f_s, f_a and f_v denote the accelerometer sampling rate, the surface vibration frequency and the resulting accelerometer reading frequency. N can be any integer. This equation also indicates the signal aliasing effect [34], where the surface vibrations at different frequencies could be mapped to a single frequency point at the accelerometer. A minute change in the notification sound frequency might lead to a completely different accelerometer signal. Such a non-linear 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. Figure 4 (b) illustrates two hands' vibration

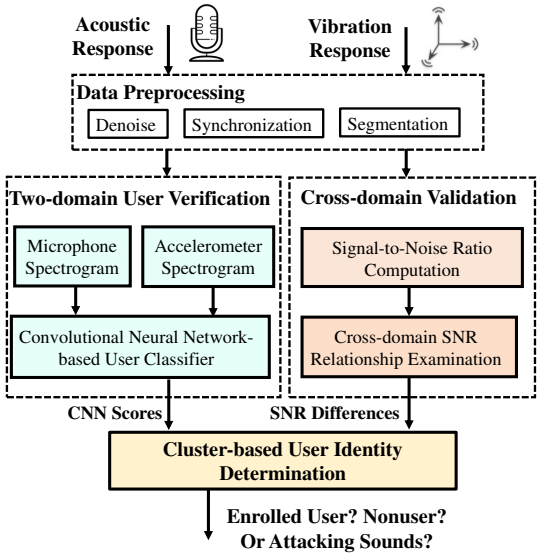


Figure 6: The architecture of our system.

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 non-linear 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 attacking sounds. To demonstrate this attribute, we conduct an experiment using the same setup as in Section 2.3.2. Figure 5 (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 the external sound, which only starts to generate 0.4dB accelerometer readings after 40dB. Moreover, the SNR gap between the two speakers increases exponentially with the louder sound. For example, a 70dB built-in speaker sound leaves 15dB SNR signals at the accelerometer while the same loudness external sound only leads to 1.4dB 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 attacking sounds. Thus, it provides additional security gains.

3 SYSTEM AND ATTACK MODELS

3.1 System Flow

We design an unobtrusive authentication system for the users who want to both protect their privacy and keep full notification features. They use notification tones or vibrations, as without notification signals it is hard to attend immediately to notifications [8]. The basic idea is to capture the unique acoustic and vibration responses of a notification tone to verify the user's hand-grip biometric. Figure 6 shows the system architecture, which takes the microphone and the accelerometer readings as the input. We perform the *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 users' gripping hands but also examine the validity of each authentication.

It is important to note that ringtones 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 notes make ringtones euphonious, but they are more difficult to analyze when used for sensing. To address this challenge, we develop the *Two-domain User Verification* method, which verifies the user's hand-grip signatures in both the acoustic and the vibration domains. We derive the microphone and accelerometer spectrograms to describe the detailed time-frequency characteristics of the ringtones 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 spectrograms respectively, and the CNN scores for the enrolled user classes and 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 resulted from the smartphone's built-in speaker passes the validity check. The sounds from external sources such as the various attacking sounds are detected as not valid and rejected. Based on the obtained CNN scores and the SNR differences, the *Cluster-based User Identity Determination* calculates the Euclidean distances to the enrolled users' cluster centers and determines whether the authentication request is from an enrolled user, a nonuser or an attacking sound.

3.2 Attack Models

The adversaries considered in this work include both the unintentional observers and the notification snoopers, whose goal is to reveal the target user's 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 the prior work, such as the privacy screen film[30], customized screen overlays [24, 37] and HideScreen [10]. 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. We assume the attacker could not compromise the phone hardware and software. 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 non-user in his/her own style when notifications come.

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

Replay Attack: 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 attacking sound. The replay sounds can be prerecorded stealthily in the user's proximity, when notifications come.

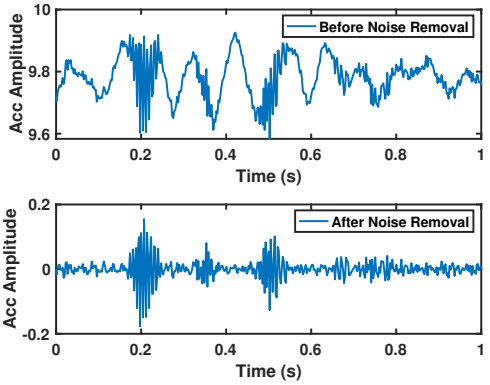


Figure 7: Removing human body movement noises.

4 APPROACH DESIGN

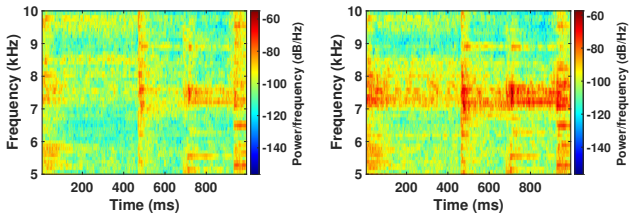
4.1 Data Pre-processing

4.1.1 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 24kHz sounds, but musical tones are only in a small audible frequency range. For example, the "Samsung whistle" spans the frequencies from 800Hz to 3000Hz, and the "iPhone Message" tone has the major frequency components within 10kHz. We thus design a minimum-order Infinite Impulse Response (IIR) bandpass filter to only focus on the tone'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 tone, 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 acoustic noises remaining in the passband. By integrating the two microphones' data (as in Section 4.3), the acoustic noises that are not correlated at the two microphones are removed.

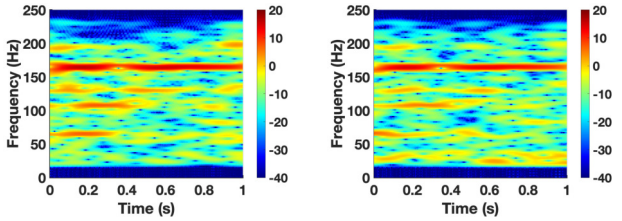
Smartphone accelerometers record up to 250Hz vibrations, and they are relatively isolated from external sounds. But when a user holds the phone, the body movements (e.g., hand vibrations) cause the accelerometer readings to be very noisy. Because human body movements are mainly in the low-frequency range, we use a high-pass filter with the 40Hz cuff-off frequency to remove them. Figure 7 shows the accelerometer readings before and after the filter, where the hand vibration noises are filtered out and the resulting accelerometer readings reflect the vibration response of the hand.

4.1.2 Synchronization and Segmentation. The microphones are turned on just before the notification tone 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



(a) User 1 (b) User 2

Figure 8: Distinguishing users via microphone spectrograms (illustrated with the iPhone Message tone).



(a) User 1 (b) User 2

Figure 9: Distinguishing users via accelerometer spectrograms (illustrated with the iPhone Message tone).

music events based on the original tone 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 believed to be not correctly collected and rejected directly. In particular, we iteratively shift the received sound S^* by m samples and compute its cross-correlation with the original tone signal S. The maximum cross-correlation indicates the time delay for synchronizing the two signals as

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

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

The accelerometer data is also logged before the notification tone begins. Its synchronization process is similar to the above. The difference is that we use the down-sampled original tone 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.

4.2 Spectrogram Derivation

We derive the spectrograms of the acoustic and vibration responses to describe the user's hand-grip biometric in two domains. Spectrogram is a time-frequency image of the data sequence, which presents the signal's temporal dynamics at every frequency point. The spectrogram of a sensor data sequence $s(\tau)$ is computed based on the Short-Time Discrete-Time Fourier Transform (ST-DTFT) as expressed by Equation 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 Equation 4. The derived microphone and vibration spectrograms are fed to our CNN model to learn gripping hand characteristics.

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

$$spectrogram(t,f) = |STDTFT(t,f)|^2 \quad (4)$$

Microphone Spectrogram. By using the above equations, we derive the microphone spectrogram. As shown in Figure 8, the same tone results in distinctive microphone spectrograms for two users' gripping hands, where "iPhone Message" tone is used for illustration. The differences are especially obvious in the frequency range of 7k~8kHz, which is a critical music event (e.g., a note) of the tone. We analyze the original tone signal to determine a frequency span that covers all of its critical music events. The microphone spectrogram in this frequency span is used for authentication.

Accelerometer Spectrogram. Similarly, the accelerometer spectrogram is derived. But different from the microphone spectrogram, which mainly carries the hand-grip biometric information by the critical music events, the accelerometer spectrogram contains useful information at all of its frequencies. The reason is that the surface vibrations resulted from the critical music events could be nonlinearly mapped to any frequency of the accelerometer. As shown in Figure 9, the accelerometer spectrograms are distinctive between two users not only at around 160Hz derived based on Equation 1, but also many other frequencies. Thus, we use the entire accelerometer spectrogram for authentication.

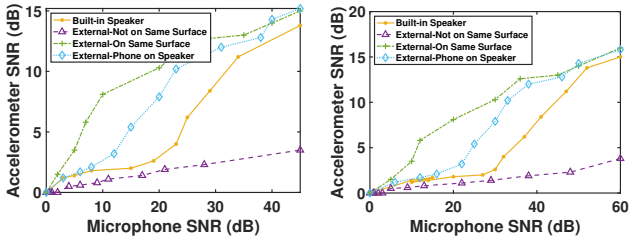
4.3 CNN-based User Identification

We use CNN to learn people's hand characteristics from complex musical tones. The CNN model is used to solve complicated image pattern recognition problems, and its strong multi-class classification capability enables us to address the user's behavioral inconsistency, such as the varied hand-grip styles resulted from each phone grabbing action.

Five-Convolutional-Layer CNN Model. We develop a CNN model with five convolutional layers. For each convolutional layer,

Table 1: Architecture of the CNN model.

Layer	Kernel Size	Output Size	# Parameters
Input: Response Spectrogram	-	(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



(a) Mic 1 vs. Accel. (b) Mic 2 vs. Accel.

Figure 10: Cross-domain SNR relationships to prevent external speaker sounds (illustrated with S8 phone).

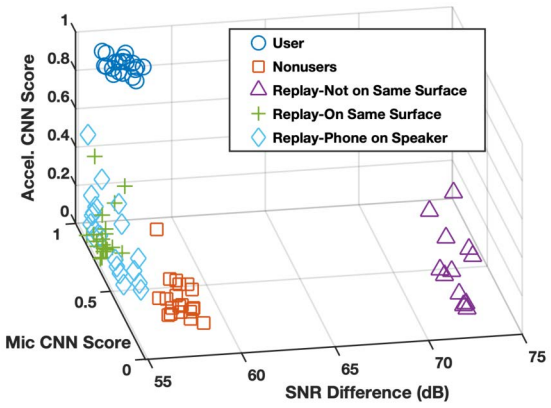


Figure 11: Clustering results of the user, nonuser and four types of machine speaker-based impersonations.

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 reducing the sensitivity to network initialization. In addition, we add a dropout layer to randomly drop 30% of the input features, 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 1 shows the detailed architecture of the CNN model. In particular, the microphone spectrograms and the accelerometer spectrograms are interpolated and normalized into 2D 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., 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 4.5. 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 [54].

4.4 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 [5, 6, 27, 50, 56]. The typical attacking scenario considered in prior work is to place an external speaker at a distance to the target device, which is then attacked by the airborne malicious sounds. But because the airborne external sounds are difficult to generate sufficiently strong vibration responses as shown in Section 2.3.3, such attacks could hardly succeed in attacking 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 attacking sounds, 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 to investigate such device-dependent physical relationships in four scenarios, when the sound is played with the sound pressure levels from 20dB to 70dB by the built-in speaker and an external speaker, respectively. Figure 10 (a) and (b) illustrate the SNR relationships between the accelerometer and two microphones (Mic1 and Mic2) of an S8 phone. We observe that both cross-domain SNR curves of the built-in speaker have no 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 15dB and Mic2 at 26dB), 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 44dB and

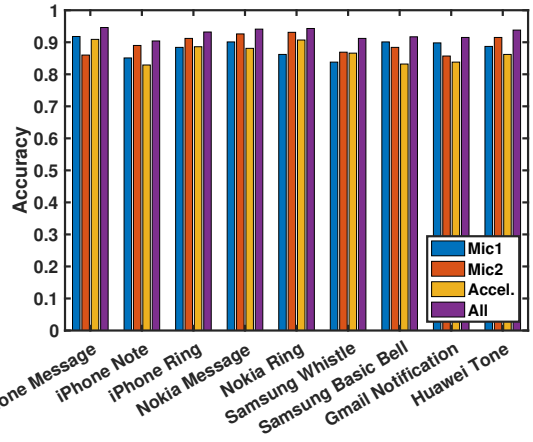


Figure 12: User identification with nine musical tones.

Mic2 at 60dB), 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.

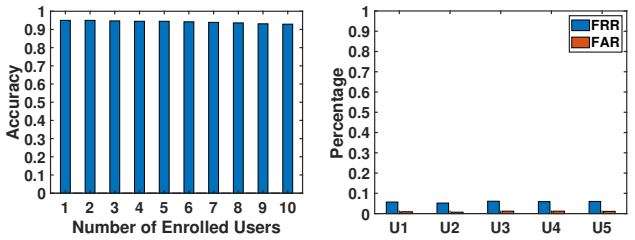
4.5 Cluster-based Classification Decision

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., attacking sounds). 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. Figure 11 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 Mic CNN scores. Thus, if only relying on the microphone for authentication, the system can be easily cheated. The replay attack is a typical issue for most acoustic-based authentication systems. Our system is different, because we further use the accelerometer CNN score and the SNR difference to distinguish valid authentication inputs from the external attacking sounds. This is only available when using smartphone media 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 their microphone and accelerometer CNN scores are low. Moreover, it is hard to forge the two domain features simultaneously, which have a non-linear relationship.

5 PERFORMANCE EVALUATION

5.1 Experimental Setup

Devices. We evaluate our system on four mobile device models (i.e., Samsung Galaxy S8, Google Pixel2, LG K50 and Motorola G8).



(a) Number of enrolled users (b) Five enrolled users

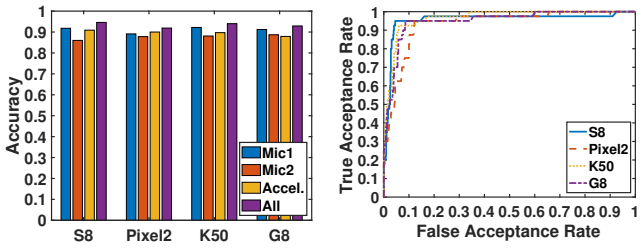
Figure 13: Performance with multiple enrolled users.

All these devices are equipped with two microphones and one accelerometer. The devices run Android 9.0 and the microphone sampling rate is set to 48kHz. The accelerometer sampling rate is set to the maximum for each device, which is 423Hz, 405Hz, 257Hz, 395Hz for S8, Pixel2, K50 and G8, respectively.

Experimental Platform. We develop an experimental platform based on Android, which plays tones 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 notification tone using *android.media.MediaPlayer*. The accelerometer readings are logged in the main thread. The authentication data is processed offline.

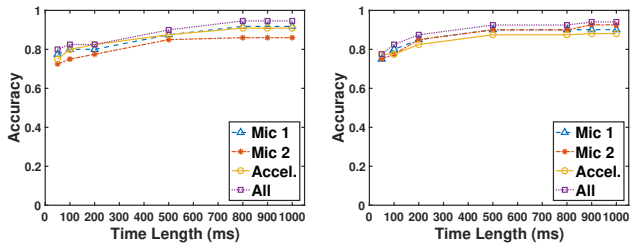
Data Collection. We recruit 30 participants (16 males and 14 females) aged from 25 to 40 to conduct experiments in 12 months. The participants include graduate students, undergraduate students and university faculties. The work has been approved by IRB. The data is collected in two sessions for each 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 minutes before data collection. In the second session, no such practice process is required. The participant needs to re-grab 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 re-grab, nine tones and one vibration alert are consecutively played with 2-second silent intervals, and only 1-second sound is used for each tone. The nine tones 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 tones, the "iPhone Message" is further played with 50% $\sim$ 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 40dB to 80dB while playing the "iPhone Message" on S8 to study the noise impact.

Attack Simulations. To simulate the human 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 replay attacks, we use an external speaker to play



(a) Accuracy (b) Roc curves

Figure 14: Performance of different device models.



(a) iPhone Message (b) Nokia Message

Figure 15: Musical tone time length study.

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 the degraded fidelity. We did not use ultrasounds, because ultrasounds have difficulty generating vibrations detected by the accelerometer and are thus very easy to be detected. We did not use adversarial examples either, because no adversarial learning method has yet been developed to forge both acoustic and vibration signals simultaneously.

5.2 Verifying Users via Musical Tones

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 notification tones. The user verification accuracies are presented in Figure 12. We observe that our system achieves a high accuracy of identifying the users with all of the nine tones when integrating the responses in two domains. In particular, "iPhone Message", "Nokia Message", and "Nokia Ring" perform the best with around 95% accuracy. Though different tones have different frequency ranges, signal powers and frequency occupation ratios [44], they all present good capabilities to distinguish people's hands. Their accuracies are all over 90.4%. Furthermore, using the two domain responses to verify users achieves better performance than using a single domain. For example, the "iPhone Message" tone achieves 94.6% accuracy by integrating both responses. It is higher than using the microphones or the accelerometer alone, which are 92.5% and 91% respectively. The results indicate that the two novel responses extracted from the audible sounds help verify users with the high performance, and our system generally works for different notification tones.

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. Figure 13(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 around 95% accuracy. When the number of enrolled users increases, the accuracy slightly decreases. When there are 10 enrolled users, our system achieves 92.9% accuracy. We future present the FRR and FAR for each enrolled user when the enrollment number is five. Figure 13(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%, and the FARs are around 1%. The nonusers have 5.7% 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 multi-user scenarios.

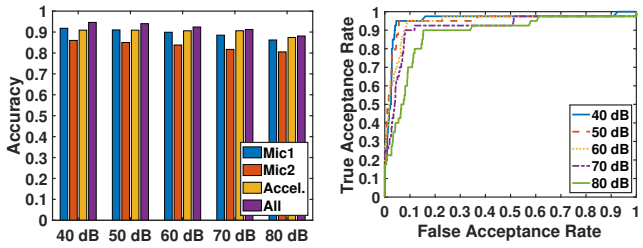
5.3 Impact of Other Factors

Device Models. Since different devices have different microphone and accelerometer configurations, shapes, sizes and surface materials, we evaluate our system on four different device models. Figure 14(a) shows the user verification accuracy for each device, when "iPhone Message" is used. We find that all the four device models perform well to identify users, which achieve accuracies between 92% and 95% based on two domains. In particular, Samsung Galaxy S8 and LG K50 achieve around 95% accuracy, which perform better than Google Pixel 2 and Motorola G8. When only using the microphones for user identification, LG K50 achieves the best performance with 92.5% accuracy. When only using the accelerometer for user identification, Samsung Galaxy S8 performs the best with 92.5% accuracy. The ROC curves in Figure 14(b) further confirm our high performance over the four models, which present high True Acceptance Rates (TARs) and low FRRs. The result indicates our system generally works for different device models.

Tone Lengths. A longer duration of the tone means more microphone and accelerometer samples to describe a higher resolution of the user's hand-grip biometric. However, the cost is the longer authentication time, which may influence the user experience. Therefore, we study the impact of the tone lengths on the system performance and try to find a proper duration to collect the authentication input. Figure 15 shows the time length study for two tones (i.e., "iPhone Message" and "Nokia Message") on S8. We find that for both tones, the longer time length slowly leads to the 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 tones achieve over 90% accuracy at 500ms. The result indicates that our system can verify a user effectively by only using a 0.5- to 1-second part of the tone.

Table 2: 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



(a) Accuracy (b) Roc curves

Figure 16: Performance under different levels of noise.

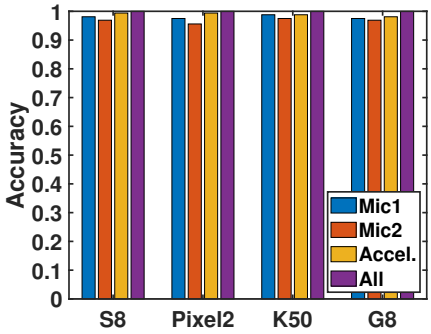


Figure 17: Performance of identifying the non-hand contexts when the smartphone is placed on a table, bed, or sofa.

Speaker Volume Impact. The volume to play the tone will also influence the user verification performance. Table 2 presents such impacts when the "iPhone Message" is played on S8 with six volume levels from 100% to 50%. We find the user identification 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.

Acoustic Noise Impact. We find 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 frequencies. We use the Decibel X App to make sure that they are played at their referenced noise levels from 40dB to 80dB [47]. The user verification performance of "iPhone Message" on S8 is shown in Figure 16. We do observe that the performance of the microphone data decreases rapidly when the noise level increases. Specifically, when the noise is increased from 40dB to 80dB, the accuracy of Mic 1 drops from 92% to 86%. However, the performance of the accelerometer data sees very little impact. When the noise is increased from 40dB to 70dB, 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 responses, our system achieves the 95%, 92% and 89% accuracy respectively under the 40dB, 70dB and 80dB noises.

Non-hand 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 *non-hand* class in our model to cover the scenarios when the smartphone is placed on a table, bed or sofa. The "iPhone Massage" tone is played by the four phone models on the three surfaces. For each surface, 80 instances are collected in two sessions, where the re-grabbing and repositioning are performed per instance. Figure 17 shows the accuracy of our system in identifying 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 non-hand contexts with all four 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 to protect the unattended phones and defend against zero-effort attacks.

5.4 Performance Under Attacks

Human Impersonations. The performance of our system to defend against the human impersonation attack is shown in Table 3 with FAR and FRR. Our system prevents the human 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.

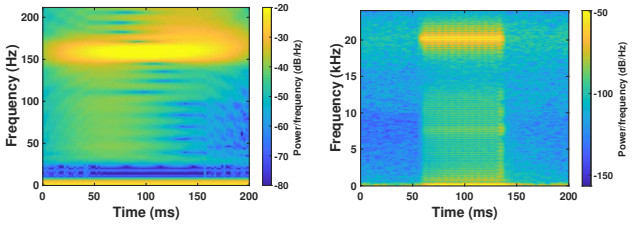
Replay Attacks. Table 3 also presents the FARs and FRRs of our system to prevent three types of replay attacks. We find that our system successfully prevents all the acoustic attacks from the external speakers, and both FARs and FRRs are 0% when the two domains' CNN scores and the cross-domain SNR relationships 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 non-linearly related responses and the cross-domain physical relationships.

6 EXTENSION TO SILENT-MODE VIBRATION ALERTS

We further investigate the notification scenarios when using musical tones may disturb people, such as in meetings and museums. We note that in order to still be attentive to notifications, users choose vibrating alerts as notification signals [8]. By considering

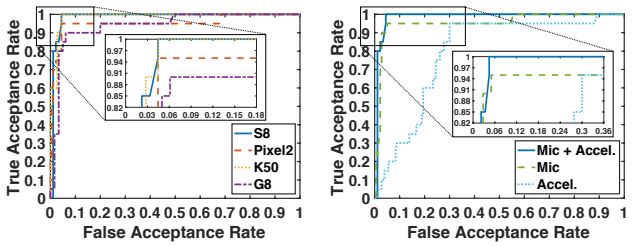
Table 3: Under impersonation and replay attacks.

Scenario	Human Impersonation	External Speaker Impersonation		
		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**



(a) On accelerometer (b) On microphone

Figure 18: Two responses of a vibration alert.



(a) Four device models (b) Mic vs. Accel. on S8

Figure 19: User verification using vibration alerts.

the vibration alert as a special tone, we extend our method to work in this silent/vibrate mode with only some parameter changes. The intuition is that sounds and vibrations co-exist, and we observe that smartphone microphones can capture the motor vibration sounds. We then derive the acoustic and vibration spectrograms of the vibration alert for authentication as shown in Figure 18 (illustrated with the Samsung S8 phone's vibrations). We note that while the accelerometer mainly captures the phone's motor frequency at 157Hz, the microphone captures the vibration sounds of up to 20kHz, 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 30 participants on four phone models, which are equipped with two types of motors. Specifically, Pixel 2 uses the Linear Resonant Actuator (LRA) motor, while the other three use the Eccentric Rotating Mass (ERM) motor. The motor frequencies are 157Hz, 155Hz, 198Hz and 208Hz for S8, Pixel 2, K50 and G8, respectively. Only 150-ms vibration signals are used. Figure 19(a) presents the user verification performances of the four devices when integrating the two domain responses. We find that all four phone models achieve over 95% accuracy. In particular, S8 performs the best with 97% accuracy. Pixel 2, K50 and G8 achieve 96%, 96.5% and 95% respectively. The results also indicate that our method works well for both linear and rotation motors. Figure 19(b) further studies the two responses based on S8. We find 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 is seldomly used to capture vibrations in prior work. Additionally, we find vibration notifications perform better than musical tones. This is because musical tones are more complex to analyze than single-frequency motor vibrations.

7 DISCUSSION AND FUTURE WORK

As the first media 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 the future work.

This work shows the potential of using media sounds for sensing. There is still room to further improve the media sound-based sensing by deriving other time-frequency images and developing more advanced deep learning algorithms, which may better tolerate users' behavioral inconsistency, background noises and attacking sounds. Moreover, the potential threats to our system need to be further 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 attaching 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 non-hand scenario as shown in Section 5.3. 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 handgrip 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 our method has a stronger capability to cope with these scenarios by leveraging the sensing information across the acoustic and vibration domains.

Face recognition might be a potential solution to address notification privacy issues. For example, the smartphone could recognize who is in front of the screen before displaying sensitive notifications. We thus compare our method with face recognition solutions. We admit that face recognition achieves higher accuracy than our method (i.e., 0.49% to 0.85% FNR [32] vs. 5% FNR of our method). But face recognition may not be a good solution to protect notification privacy due to the following weaknesses. Firstly, the smartphones' front cameras have narrow view angles (e.g., +/- 30 degrees), which requires the user to look straight at the devices during the face recognition process. Second, face recognition has higher energy consumption (i.e., around 5W [52]) compared to audio processing (i.e., 200-500mW [58]), and turning on a camera incurs a higher latency (i.e., 1.5-1.8s [46]) than opening a microphone. Moreover, face recognition may not work well in low light conditions and suffer from potential replay attacks (e.g., 3D face). It also has limited application when the user wears a mask (e.g., in the Covid-19 pandemic). In comparison, our method is not subject to these issues.

8 RELATED WORK

To protect sensitive information in smartphones, two types of authentications methods are widely deployed. The knowledge-based methods verifies the user's PINs, passwords, lock patterns [49] and graphical secrets [18]. The biometric-based methods verifies the user's body traits or behavioral characteristics, including fingerprints [48], faceIDs [7], iris [4], hand geometries [13, 16] and voices [43]. However, all these methods are one-time authentications and require active user participation. They are not suitable to handle notifications, which have an unpredictable nature and could be displayed on locked or unlocked screens.

Continuous authentication is an emerging technique to supplement the one-time authentications, which aims to verify the user's identity at any time. For example, the behavioral characteristics such as gaits [42], keystroke dynamics [57] 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 [9, 28], heartbeat biometrics [14, 29] and brain waves [25, 38] achieve the anytime 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 well in a timely manner to respond to sudden events. Additionally, the continuous authentications drain batteries fast and are not energy-friendly for mobile devices.

The current notification privacy protections mainly rely on the manual management of privacy off line [26, 33] or in situations [19, 21]. 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 the 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 situations, which is obtrusive and may already arouse others' 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 [41]. But similar to other in-situ methods, PrivacyShield still imposes the responsibility on the user to take the initiation for privacy protection, which is inconvenient and may not work well in the 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 the smartphone notification, he or she has to first grab this handheld device in hand. There has been some authentication work on verifying the user's gripping behaviors [40, 53]. 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 [11, 55], 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 attacking sounds (e.g., at least 10% FAR under the record-and-replay attack as shown in Table 3). 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.

9 CONCLUSION

This work provides an unobtrusive solution to protect the user's notification privacy while keeping full notification features. The proposed system protects the user's notification privacy in both silent and non-silent smartphone modes by directly using the notification signals such as musical tones and vibrating alerts. We show that the smartphone media sounds, though more complicated than dedicated signals, can be 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 spectrograms to describe 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 mic, speaker and accelerometer, which are embedded on the same motherboard, to prevent external sounds (e.g., noises and attacks) from obstructing the system. Extensive experiments show that our system verifies the user with 95% accuracy and prevents 100% replay sounds from external speakers.

ACKNOWLEDGMENTS

This work was partially supported by LEQSF(2020-23)-RD-A-11. We would also like to thank our anonymous shepherd and all the reviewers for helping us improve the paper.

REFERENCES

[1] S Abhishek Anand, Chen Wang, Jian Liu, Nitesh Saxena, and Yingying Chen. 2019. Spearphone: A speech privacy exploit via accelerometer-sensed reverberations from smartphone loudspeakers. *arXiv preprint arXiv:1907.05972* (2019).

[2] AppleSupport. 2020. Get a verification code and sign in with two-factor authentication. https://support.apple.com/en-us/HT204974.

[3] Anthony Bouchard. 2019. Improve your iPhone's notification privacy with Blurification. https://www.idownloadblog.com/2019/05/07/blurification/.

[4] James L Cambier and John E Siedlarz. 2003. Portable authentication device and method using iris patterns. US Patent 6,532,298.

[5] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. 2016. Hidden voice commands. In *25th {USENIX} Security Symposium ({USENIX} Security 16)*. 513–530.

[6] Nicholas Carlini and David Wagner. 2018. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, 1–7.

[7] JV Chamary. 2019. How Face ID Works On iPhone X. https://www.forbes.com/sites/jvchamary/2017/09/16/how-face-id-works-apple-iphone-x/?sh=636ddfe3624d.

[8] Yung-Ju Chang and John C Tang. 2015. Investigating mobile users' ringer mode usage and attentiveness and responsiveness to communication. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 6–15.

[9] Jagmohan Chauhan, Yining Hu, Suranga Seneviratne, Archan Misra, Aruna Seneviratne, and Youngki Lee. 2017. BreathPrint: Breathing acoustics-based user authentication. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. 278–291.

[10] Chun-Yu Chen, Bo-Yao Lin, Junding Wang, and Kang G Shin. 2019. Keep Others from Peeking at Your Mobile Device Screen!. In *The 25th Annual International Conference on Mobile Computing and Networking*. 1–16.

[11] Huijie Chen, Fan Li, Wan Du, Song Yang, Matthew Conn, and Yu Wang. 2020. Listen to Your Fingers: User Authentication Based on Geometry Biometrics of Touch Gesture. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–23.

[12] Zong Chen and Michael Recce. 2007. Handgrip recognition. *Journal of Engineering, Computing and Architecture* 1, 2 (2007).

[13] Michał Choraś and Rafał Kozik. 2012. Contactless palmprint and knuckle biometrics for mobile devices. *Pattern Analysis and Applications* 15, 1 (2012), 73–85.

[14] Júlio da Silva Dias, Issa Traore, Vitor GRA Ferreira, and Julibio David. 2015. Exploratory use of PPG signal in continuous authentication. In *The Brazilian Symposium on Information and Computational Systems Security*.

[15] Anupam Das, Nikita Borisov, and Matthew Caesar. 2014. Do you hear what i hear? fingerprinting smart devices through embedded acoustic components. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. 441–452.

[16] Alberto de Santos Sierra, Javier Guerra Casanova, Carmen Sanchez Avila, and Vicente Jara Vera. 2009. Silhouette-based hand recognition on mobile devices. In *Security Technology, 2009. 43rd Annual 2009 International Carnahan Conference on*. IEEE, 160–166.

[17] Sanorita Dey, Nirupam Roy, Wenyuan Xu, Romit Roy Choudhury, and Srihari Nelakuditi. 2014. Accelprint: Imperfections of accelerometers make smartphones trackable. In *Proceedings of the Network and Distributed System Security Symposium (USENIX NDSS)*.

[18] Rachna Dhamija, Adrian Perrig, et al. 2000. Deja Vu-A User Study: Using Images for Authentication.. In *USENIX Security Symposium*, Vol. 9. 4–4.

[19] Jonny Evans. 2017. How to use Guided Access to secure your iPad or iPhone. https://www.computerworld.com/article/3162738/how-to-use-guided-access-to-secure-your-ipad-or-iphone.html.

[20] Mario Frank, Ralf Biedert, Eugene Ma, Ivan Martinovic, and Dawn Song. 2013. Touchalytics: On the applicability of touchscreen input as a behavioral biometric for continuous authentication. *IEEE transactions on information forensics and security* 8, 1 (2013), 136–148.

[21] Google. 2020. Supporting Multiple Users. https://source.android.com/devices/tech/admin/multi-user.

[22] Chris Hoffman. 2017. How to Hide Sensitive Notifications From Your iPhone's Lock Screen. https://www.howtogeek.com/252483/how-to-hide-sensitive-notifications-from-your-iphones-lock-screen/.

[23] Mat Honan. 2014. Why Notifications Are About to Rule the Smartphone Interface. https://www.wired.com/2014/06/smartphone-notifications/.

[24] HueySoft. 2019. Privacy Screen Filter. https://play.google.com/store/apps/details?id=com.hueysl.privacyscreen&hl=en$_U$S.

[25] W Khalifa, A Salem, M Roushdy, and K Revett. 2012. A survey of EEG based user authentication schemes. In *2012 8th International Conference on Informatics and Systems (INFOS)*. IEEE, BIO–55.

[26] Seungchul Lee, Saumay Pushp, Chulhong Min, and Junehwa Song. 2018. Exploring Relationship-aware Dynamic Message Screening for Mobile Messengers. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. 134–137.

[27] Johan Lindberg and Mats Blomberg. 1999. Vulnerability in speaker verification-a study of technical impostor techniques. In *Sixth European Conference on Speech Communication and Technology*.

[28] Jian Liu, Yingying Chen, Yudi Dong, Yan Wang, Tiannming Zhao, and Yu-Dong Yao. 2020. Continuous user verification via respiratory biometrics. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 1–10.

[29] Wael Louis, Majid Komeili, and Dimitrios Hatzinakos. 2016. Continuous authentication using one-dimensional multi-resolution local binary patterns (1DMRLBP) in ECG biometrics. *IEEE Transactions on Information Forensics and Security* 11, 12 (2016), 2818–2832.

[30] STEVEN WILLIAM MacMASTER. 2006. Privacy screen for a display. US Patent 7,052,746.

[31] Lindsay Mannering. 2019. How to Not Ruin Your Life (or Just Die of Embarrassment) With a Screen Share. The New York Times. https://www.nytimes.com/2019/03/21/style/screen-share-privacy-tips.html.

[32] Michael McLaughlin and Daniel Castro. 2020. The Critics Were Wrong: NIST Data Shows the Best Facial Recognition Algorithms Are Neither Racist Nor Sexist. https://itif.org/publications/2020/01/27/critics-were-wrong-nist-data-shows-best-facial-recognition-algorithms.

[33] Abhinav Mehrotra, Robert Hendley, and Mirco Musolesi. 2016. PrefMiner: mining user's preferences for intelligent mobile notification management. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 1223–1234.

[34] Yan Michalevsky, Dan Boneh, and Gabi Nakibly. 2014. Gyrophone: Recognizing speech from gyroscope signals. In *23rd {USENIX} Security Symposium ({USENIX} Security 14)*. 1053–1067.

[35] Charles J Migos and David H Sloo. 2012. Personalization using a hand-pressure signature. US Patent 8,172,675.

[36] Chulhong Min, Saumay Pushp, Seungchul Lee, Inseok Hwang, Youngki Lee, Seungwoo Kang, and Junehwa Song. 2014. Uncovering embarrassing moments in in-situ exposure of incoming mobile messages. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. 1045–1054.

[37] mortisApps. 2013. Screen Guard privacy screen. https://play.google.com/store/apps/details?id=com.mortisapps.privacyfilter.

[38] Isao Nakanishi, Sadanao Baba, and Chisei Miyamoto. 2009. EEG based biometric authentication using new spectral features. In *2009 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*. IEEE, 651–654.

[39] Business of Apps. 2019. Push Notifications Statistics (2019). https://www.businessofapps.com/marketplace/push-notifications/research/push-notifications-statistics/.

[40] Manabu Ota, Yasuo Morinaga, Masakatsu Tsukamoto, and Takeshi Higuchi. 2013. Portable terminal and gripping-feature learning method. US Patent App. 13/881,386.

[41] Saumay Pushp, Yunxin Liu, Mengwei Xu, Changyoung Koh, and Junehwa Song. 2018. PrivacyShield: A Mobile System for Supporting Subtle Just-in-time Privacy Provisioning through Off-Screen-based Touch Gestures. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–38.

[42] Yanzhi Ren, Yingying Chen, Mooi Choo Chuah, and Jie Yang. 2014. User Verification Leveraging Gait Recognition For Smartphone Enabled Mobile Healthcare Systems. *IEEE Transactions on Mobile Computing* (2014).

[43] Douglas A Reynolds and Richard C Rose. 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE transactions on speech and audio processing* 3, 1 (1995), 72–83.

[44] SM Series. 2017. Spectrum occupancy measurements and evaluation. (2017).

[45] SkyPaw. 2020. Decibel X. http://skypaw.com/decibelx.html.

[46] Mark Spoonauer. 2017. iPhone X Face ID Slower Than Touch ID (But There's a Fix). https://www.tomsguide.com/us/iphone-x-face-id-speed-up,news-26060.html.

[47] The Engineering Toolbox. 2004. Sound Pressure. https://www.engineeringtoolbox.com/sound-pressure-d$_7$11.html.

[48] Constantine Tsikos. 1982. Capacitive fingerprint sensor. US Patent 4,353,056.

[49] Sebastian Uellenbeck, Markus Dürmuth, Christopher Wolf, and Thorsten Holz. 2013. Quantifying the security of graphical passwords: the case of android unlock patterns. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*. 161–172.

[50] Tavish Vaidya, Yuankai Zhang, Micah Sherr, and Clay Shields. 2015. Cocaine noodles: exploiting the gap between human and machine speech recognition. In *9th {USENIX} Workshop on Offensive Technologies ({WOOT} 15)*.

[51] Chen Wang, S Abhishek Anand, Jian Liu, Payton Walker, Yingying Chen, and Nitesh Saxena. 2019. Defeating hidden audio channel attacks on voice assistants via audio-induced surface vibrations. In *Proceedings of the 35th Annual Computer Security Applications Conference*. 42–56.

[52] Yi-Chu Wang and Kwang-Ting Cheng. 2011. Energy-optimized mapping of application to smartphone platform - a case study of mobile face recognition. In *CVPR 2011 WORKSHOPS*. IEEE, 84–89.

[53] Arnold S Weksler, Nathan J Peterson, and Russell Speight VanBlon. 2015. Grip signature authentication of user of device. US Patent App. 14/098,180.
[54] Mengwei Xu, Jiawei Liu, Yuanqiang Liu, Felix Xiaozhu Lin, Yunxin Liu, and Xuanzhe Liu. 2019. A first look at deep learning apps on smartphones. In *The World Wide Web Conference*. 2125–2136.
[55] Yilin Yang, Chen Wang, Yingying Chen, and Yan Wang. 2020. EchoLock: Towards Low Effort Mobile User Identification. *arXiv preprint arXiv:2003.09061* (2020).
[56] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. 2017. Dolphinattack: Inaudible voice commands. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 103–117.
[57] Nan Zheng, Kun Bai, Hai Huang, and Haining Wang. 2014. You Are How You Touch: User Verification on Smartphones via Tapping Behaviors.. In *ICNP*, Vol. 14. 221–232.
[58] Sergey Zhidkov, Andrey Sychev, Alexander Zhidkov, and Alexander Petrov. 2018. On smartphone power consumption in acoustic environment monitoring applications. *Applied System Innovation* 1, 1 (2018), 8.