Proximity-Echo: Secure Two Factor Authentication Using Active Sound Sensing

Yanzhi Ren[1], Ping Wen[1], Hongbo Liu[1], Zhourong Zheng[1], Yingying Chen[2], Pengcheng Huang[1], Hongwei Li[1]
[1]Dept. of CS, University of Electronic Science and Technology of China [2]WINLAB, Rutgers University, USA
[1]{renyanzhi05, hongweili}@uestc.edu.cn [2]yingche@scarletmail.rutgers.edu

IEEE INFOCOM 2021 - IEEE Conference on Computer Communications | 978-1-6654-0325-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/INFOCOM42981.2021.9488866

***Abstract*—The two-factor authentication (2FA) has drawn increasingly attention as the mobile devices become more prevalent. For example, the user's possession of the enrolled phone could be used by the 2FA system as the second proof to protect his/her online accounts. Existing 2FA solutions mainly require some form of user-device interaction, which may severely affect user experience and creates extra burdens to users. In this work, we propose Proximity-Echo, a secure 2FA system utilizing the proximity of a user's enrolled phone and the login device as the second proof without requiring the user's interactions or pre-constructed device fingerprints. The basic idea of Proximity-Echo is to derive location signatures based on acoustic beep signals emitted alternately by both devices and sensing the echoes with microphones, and compare the extracted signatures for proximity detection. Given the received beep signal, our system designs a period selection scheme to identify two sound segments accurately: the chirp period is the sound segment propagating directly from the speaker to the microphone whereas the echo period is the sound segment reflected back by surrounding objects. To achieve an accurate proximity detection, we develop a new energy loss compensation extraction scheme by utilizing the extracted chirp periods to estimate the intrinsic differences of energy loss between microphones of the enrolled phone and the login device. Our proximity detection component then conducts the similarity comparison between the identified two echo periods after the energy loss compensation to effectively determine whether the enrolled phone and the login device are in proximity for 2FA. Our experimental results show that our Proximity-Echo is accurate in providing 2FA and robust to both man-in-the-middle (MiM) and co-located attacks across different scenarios and device models.**

I. Introduction

The mobile two-factor authentication (2FA) becomes increasingly critical as mobile devices (e.g., smartphones, tablets, wearables) are used extensively in our daily lives. In mobile 2FA, a user logs into the system from a login device, which can be an arbitrary networked device such as a laptop, a smartphone, a tablet or even a public computer using his/her username and password. Such system then further utilizes the user's enrolled phone or other enrolled mobile devices (e.g., tablets) as the second security proof to protect the online accounts as using passwords alone is vulnerable to spidering or steal attacks [1]. For instance, when a user tries to log into an online bank account which employs mobile 2FA, the system verifies the user's possession of his/her enrolled smartphone after he/she enters the username and password. So in such systems, the smartphone serves as a second proof of the user's identity and the system can still keep safe even if the username and password have been leaked.

In these 2FA solutions, active interactions between a user and his/her phone is usually required. For example, commercial 2FA systems such as Duo Mobile App and Google 2-step Verification [2], [3], which can be integrated with various online systems, either send a random passcode to the enrolled device for the user to input on the interface or call a pre-registered phone for the user's answer to finish the login process. These 2FA techniques need users' active participation and could be cumbersome for user experience or even add additional burdens to senior citizens and people with disabilities.

Some studies have been proposed to improve the usability of mobile 2FA by eliminating the explicit user interaction. For instance, recent studies demonstrated that the ambient sound can be utilized to detect the proximity of the enrolled phone and login device [4] without user interaction for 2FA. However, this scheme may become invalid if the adversary can guess or generate very similar ambient sound at the login device's end [5]. Han *et al.* [6] shows the initial success of mobile 2FA via the acoustic ranging. However, this technique requires each device to pre-construct an acoustic fingerprint to thwart the man-in-the-middle (MiM) attack, which makes it less suitable for large-scale deployment.

When the enrolled phone and the login device are in close proximity, they should share highly similar surrounding environments. Such similarity, if captured, could be utilized for the proximity detection of two devices. (It is worth noting that here we assume the enrolled phone and the login device are two distinct devices, otherwise it can be easily detected by the system server when sending the request messages, and in this case, the login request will be accepted directly. The details of the system flow are presented in Section III-C). These observations trigger our idea from acoustic sensing [7]–[10] to use the reflected beep sound, which contains rich information of surrounding objects, as the proximity proof of two devices for 2FA. Specifically, in this paper, we proposed Proximity-Echo, a mobile 2FA system that uses the acoustic location signatures, which are derived from both devices by emitting acoustic beep signals with their speakers and sensing the echoes with their microphones, as the second authentication factor.

However, several unique challenges need to be addressed when developing such a system: First, due to different electronic features or manufacturing imperfections, the energy loss (i.e., the energy gain or attenuation measurement at each frequency) of microphones on the enrolled phone and the login

978-1-6654-0325-2/21/$31.00 ©2021 IEEE
Authorized licensed use limited to: Rutgers University. Downloaded on October 01,2021 at 19:51:48 UTC from IEEE Xplore. Restrictions apply.

device may vary significantly, making it hard to compare the echoes received by two devices directly for accurate proximity detection. Second, the received beep signal can be significantly affected by the distance between the devices' speaker and microphone or the relative position of objects in the surrounding environment, making it hard to identify the segment of reflected sound from the received signal accurately. Third, the 2FA process should be able to complete at the minimal efforts without the involvement of users' extra interactions or pre-constructed device fingerprints. Last but not least, an attacker can relay the beep signal between the enrolled phone and the login device or be physically co-located with the victim in an attempt to pass the 2FA. Our system should be resilient to such MiM attack or co-located attack attempts.

To cope with these challenges, our proposed Proximity-Echo consists of three main components: *Period Selection*, *Energy Loss Compensation Extraction* and *Proximity Detection*. Given the input microphone samplings, *Period Selection* is first performed to identify two sound segments named the chirp period and the echo period from the received beep signal accurately, which correspond to the sound segment that propagates directly from the speaker to the microphone and the sound segment reflected back by surrounding objects, respectively. *Energy Loss Compensation Extraction* is the core component that derives the compensation which estimates the intrinsic difference of energy loss between microphones of the enrolled phone and the login device using the energy spectrum of chirp periods. Such compensation is only tie to the microphone-microphone pair and it remains invariant even if the distance between two devices varies. During *Proximity Detection*, after the energy loss compensation, the similarity comparison is performed between echo periods extracted by two devices to determine whether the enrolled phone and the login device are in proximity for 2FA. We summarize our main contributions as follows:

- We design Proximity-Echo, a new mobile 2FA system which utilizes the proximity between the enrolled phone and the login device as the second proof.
- We propose to use acoustic location signatures, which are derived from devices by emitting a pre-designed acoustic beep signal and sensing its echoes, as the proximity proof for 2FA.
- We design a period selection scheme to identify the chirp period and the echo period from the received signal accurately by exploiting the inherent correlation between the original beep signal and the received beep signal.
- To achieve an accurate proximity detection, we develop a new energy loss compensation extraction scheme to estimate the intrinsic differences of energy loss between microphones of the enrolled phone and the login device using the energy spectrum of the identified chirp periods.
- We show that our Proximity-Echo is robust to adversarial behaviors of relaying signals between the enrolled phone and a remote adversarial login device, or being physically co-located with the victim in an attempt to pass the 2FA.
- Our extensive experimental results show that our proposed Proximity-Echo is accurate and robust across different device models under various real world scenarios.

II. Related Work

Traditional software based authentication mechanisms such as Duo Mobile App and Google 2-step Verification [2], [3] have been developed for 2FA. Such systems send a passcode to the enrolled device for the user to input on an interface to finish the login process. The main advantage of these mechanisms is that they can be easily integrated with online systems. However, they require the user to interact with his/her device explicitly and it can severely affect the user experience and bring extra burdens to him/her.

Some recent studies have been developed to improve the usability of 2FA mechanism by eliminating the need of user interaction. Some technique [11] designs a challenge-response based protocol between the login device and the enrolled phone for 2FA without user interaction. However, the Bluetooth function is required in this scheme and such function is not always available on login devices. In addition, Sound-Proof [4] utilizes the ambient sound as the proximity proof between devices for 2FA. However, this technique may become invalid if the adversary can generate the same ambient sound at the login device's end [5]. Listening Watch [12] addresses the limitation of Sound-Proof by populating a short random number encoded into sound to defeat the attacker unless it is extremely close to the device. However, such scheme still requires an extra smart watch equipped with a low sensitivity microphone to pick up nearby sounds.

There are also recent works dedicated for acoustic sensing. In these techniques, a device's speaker emits a pre-designed signal and then uses its microphone to capture the echoes to identify or distinguish the surrounding environment. Tung *et al.* [7] and Song *et al.* [8] propose to recognize indoor locations by transmitting a sound signal and sensing its reflected signal with the phone's microphone. Pradhan *et al.* [9] develop a smartphone-based indoor space mapping system that enables a user quickly map the indoor space via acoustic sensing.

The most similar work to our own is by [6]. They propose a technique to extract fingerprints for the speaker and the microphone on each device from the acoustic signal for 2FA. However, such system requires to pre-construct a legitimate acoustic fingerprint for each device, which may bring extra burden to users. Unlike the aforementioned work, we aim to develop a 2FA mechanism by using the proximity of the user's enrolled phone and the login device as the second authentication factor via acoustic sensing. Our proposed system does not need the user's explicit participation and is also easy-to-use without requiring any dedicated hardware or pre-constructed device fingerprints.

III. Framework Overview

In this section, we first describe the design goals of our 2FA system. We then introduce the adversary model and provide an overview of our proposed system.

Authorized licensed use limited to: Rutgers University. Downloaded on October 01,2021 at 19:51:48 UTC from IEEE Xplore. Restrictions apply.

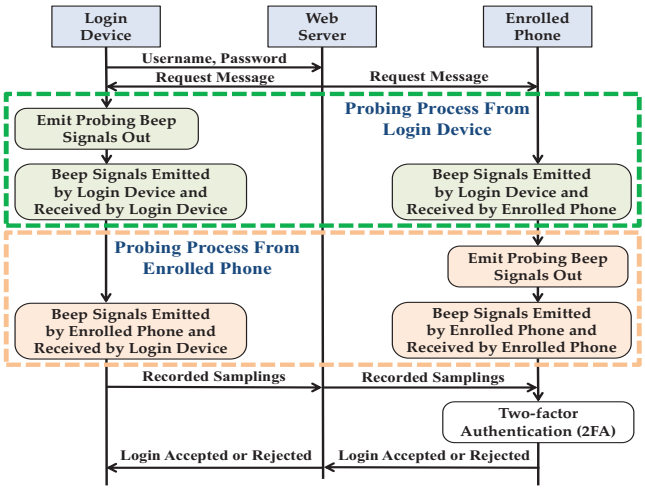


Fig. 1. 2FA system model.

A. Design Goals

We consider a mobile 2FA system which uses the proximity of the enrolled phone and the login device as the second proof. Specifically, our system derives acoustic location signatures from both devices through sensing the reflected sound of the beep signal emitted by speakers, and uses such signatures as proximity proof for 2FA. In particular, our system is designed to meet following requirements:

Robust to Energy Loss Variations. The energy loss of microphones on the login device and the enrolled phone may vary significantly due to their different electronic features or manufacturing imperfections, making the direct comparison between echoes received by two devices be a challenging task. Our system should be able to estimate and compensate such difference between microphones on two devices to achieve an accurate proximity detection.

Robust to Different Environments and Device Distances. The received beep signal can be easily affected by the relative position of surrounding objects and the distance between speaker-microphone pairs from devices. Our system should be able to identify the segment for reflected sound from the received beep signal accurately for proximity detection.

Easy to Use. Our system should be able to complete the 2FA process with minimal efforts: no explicit user interaction or pre-constructed device fingerprints are needed as it may bring extra burdens to users.

Secure Against to Various Attacks. An adversary can obtain the victim's username and password via the leakage of the password database or phishing [4] and try to pass our 2FA system by launching attacks. Our system should be secure to various attacks on our mobile 2FA schemes, including the MiM attack and the co-located attack.

B. Adversary Model

An adversary has compromised the victim's username and password in an attempt to pass the 2FA on behalf of the victim. The attack is successful if the adversary can convince the system that he/she holds the second authentication factor of the victim. For instance, in this work, such factor is the proximity between the login device and the enrolled phone associated with the victim's username. As prior works [4], [6], [13], we also adopt assumptions as follows: First, the adversary cannot compromise the victim's phone, otherwise the security of the 2FA system reduces to the security of a regular password based authentication system. Second, the communication channel between the server and the login device (or the enrolled phone) is secured using the TLS-like mechanisms. Based on these assumptions, we consider two possible adversarial behaviors as described below:

- **Man-in-the-middle Attack:** The adversary is far away from the victim and his/her enrolled phone. However, the adversary puts an eavesdropping device near the victim's enrolled phone and sets up an invisible channel with high speed between the enrolled phone and the adversarial login device. When the adversary tries to login, the triggered beep sounds emitted by two devices will be relayed to each other in real time via the adversarial channel.
- **Co-located Attack:** In this case, the adversary is physically co-located with the victim in some public places such as libraries, restaurants or cafeterias. The adversary's attempts to login the system triggers the adversarial login device and the victim's enrolled phone to emit beep signals alternatively for 2FA, which can be directly received by both devices' microphones.

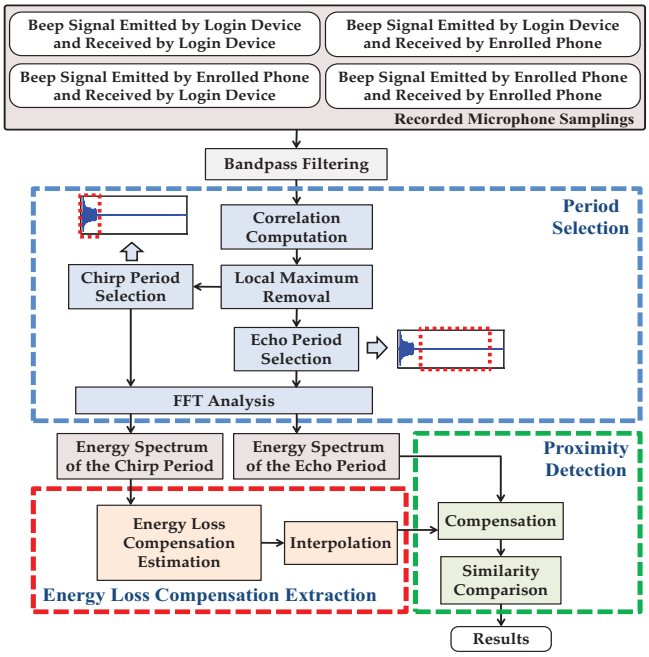


Fig. 2. System flow of our Proximity-Echo.

C. System Overview

As prior works [4], [5], we consider a mobile 2FA system model in which a user has his/her username and password for login and the core 2FA mechanism has been implemented on the user's enrolled phone. As shown in Figure 1, the user logs into the system from a login device, which could

Authorized licensed use limited to: Rutgers University. Downloaded on October 01,2021 at 19:51:48 UTC from IEEE Xplore. Restrictions apply.

be any networked device such as a laptop, a smartphone, a tablet or even a public computer. When he/she attempts to login, the username and password are required to input on the interface of the login device, which is relayed to the server via a secure channel. The server then verifies the validity of the password and sends request messages to the login device and the enrolled phone (or other enrolled mobile devices such as tablets) which is associated with the username to validate the second authentication factor (i.e., the proximity of two devices). Specifically, probing beep signals are emitted out by the login device and the enrolled phone alternatively, and they are then received by both devices' microphones. The login device then encrypts the recorded microphone samplings using the public key and sends them to the enrolled phone using the server as a proxy. The phone then decrypts these received samplings and uses them together with the microphone samplings recorded locally to perform our proposed 2FA. If the authentication process passes, the enrolled phone concludes that it is in proximity with the device from which the user is logging in and informs the server to accept the login process.

As shown in Figure 2, our Proximity-Echo consists of three major components: *Period Selection*, *Energy Loss Compensation Extraction* and *Proximity Detection*. The system takes as input the recorded microphone samplings from both the enrolled phone and the login device. In the period selection phase, to deal with variations caused by the distances between the speaker and the microphone or the relative positions of surrounding objects, a correlation based technique is proposed to accurately identify the chirp period and the echo period from the received beep signal, which correspond to the sound segment that propagates directly from the speaker to the microphone and the sound segment reflected back by surrounding objects, respectively. The energy loss compensation extraction is conducted to capture the differences of energy loss between microphones of two devices from the energy spectrum of identified chirp periods. Such extracted compensation is usually only tie to the microphone-microphone pair and it remains invariant even if the distance between two devices changes. After the energy loss compensation, the proximity detection is performed by calculating a correlation value between echo periods extracted from the enrolled phone and the login device respectively. Based on the correlation value, our system makes decision on whether to accept or reject the login request.

IV. Two-factor Authentication (2FA) system

In this section, we present the detailed system implementation of our Proximity-Echo.

A. Beep Design

When designing the probing beep signal played through the speaker for proximity detection, we mainly consider three factors: frequency band, length and time interval.

Frequency Band. Studies show that human can hear acoustic signals of frequency up to 20 kHz [7]. Thus, it may be desirable to set the frequency above 20 kHz to make the emitted sound inaudible (to avoid annoyance) to users. However, due to the hardware's imperfection, the frequency response of most mobile devices decays quickly when the frequency is beyond 20 kHz [10]. On the other hand, the frequency below 11 kHz may not be used since it contains most environmental noises of human activities [14]. In summary, given the trade-off of all factors, our system adopts the 14 kHz to 15 kHz bandwidth beep acoustic signal. Even though this frequency selection makes the beep signal audible to humans, the impact is not obvious because our 2FA triggers the authentication process infrequently.

Length. The length of beep signal also impacts the accuracy and reliability of our 2FA system. The speaker and microphone on mobile devices cannot generate or pick up too short beep signals. Thus, it seems best to set a longer length of the beep signal since more energy at each frequency could be collected. However, a too long duration of the emitted beep signal could cause severe multipath distortions since reflections which are from far away objects will also be collected during this long sensing process [7]. Therefore, in this work, we empirically set the length of the beep signal as 0.02s. This selection reduces the multipath distortions but keeps enough energy at each frequency for proximity detection.

Time Interval. The last parameter we consider is the time interval between two consecutive beep signals. This parameter is related to the sensing speed of our system: a larger time interval results in a longer time our system needs to take for 2FA. On the other hand, a short interval causes detection errors since the reflected signals might accidentally overlap with each other. Based on our observations, the reflected sound could still exist even after 0.2s to 0.4s from the beep sound. Therefore we set the interval to be 0.5s in this work.

B. Period Selection

The basic idea underlying our Proximity-Echo is to use the reflected beep sounds, which contains rich information of surrounding patterns, as the proximity proof for 2FA. However, the received beep signal not only contains the sound segment which reflects back from surrounding objects (named the echo period) but also includes the sound segment which directly propagates from the speaker to the microphone (named the chirp period). Identifying both the chirp period and the echo period from the received beep signal accurately is a challenging task because such received signal can be easily affected by the distance between the speaker and microphone or the relative positions of reflective objects in the surrounding environment. In the commonly used period identification techniques [7], [8], the chirp period and echo period are determined by truncating acoustic data out from the received beep signal using fixed windows. The problem of these schemes is that the beginning points of these periods may vary significantly and such techniques cannot adapt to these changes.

To solve this problem, in this paper, we propose a correlation based technique for period selection by utilizing the fact that the original beep signal should have a good match with the received copies of the beep sound embedded in the

Authorized licensed use limited to: Rutgers University. Downloaded on October 01,2021 at 19:51:48 UTC from IEEE Xplore. Restrictions apply.

received signal. The basic idea of our scheme is to evaluate the correlations between the received signal and the original beep sound, and the peaks in the correlation sequence allow us to detect the beginning points of the chirp period and the echo period accurately. Specifically, the received beep signal is first sampled with a frequency of 48 kHz and a bandpass filter with lower and upper cutoff frequencies, which are 14 kHz and 15 kHz respectively, is applied to remove environmental noises and extract signal components which fall into the frequency range of the beep signal. We assume that $e_l(t)$ represents the received signal for the l-th beep signal ($1 \leq l \leq L$) after filtering and let $s(t)$ be the original beep signal. Thus, to conduct the period selection, the original signal $s(t)$ is slid across the acoustic readings of $e_l(t)$ with a moving window and the correlation is calculated based on the matched filter as follows:

$$C_l(t) = \int_{-\infty}^{+\infty} e_l(\tau)h(t-\tau)d\tau \tag{1}$$

where $h(t)$ denotes the conjugated and time-reversed version of the original beep signal $s(t)$ (i.e., $h(t) = s^*(-t)$). To capture the overall trend changes of $C_l(t)$, we identify the envelope of $C_l(t)$ using the envelope detection schemes proposed in [15] and denote it as $E_l(t)$. Thus, the periodical peaks within envelope $E_l(t)$ can be used as candidates to identify the beginning points of the chirp period and echo period, respectively.

To identify these peaks, we search for a set of local maximums from $E_l(t)$ by varying t and denote it as $MaxSet$ which consists of W local maximum points: $MaxSet = \{\tau_k \mid 1 \leq k \leq W\}$. For each $\tau_k \in MaxSet$, it satisfies that $E_l(\tau_k) > E_l(t)$ for any $t \in (\tau_k - d, \tau_k + d)$ and $E_l(\tau_k) > th$, where d is a pre-defined small distance and th is a threshold. Ideally, the first local maximum point τ_1 in $MaxSet$ should therefore correspond to the reception of sound traveled directly from the speaker to the microphone (i.e., the possible beginning point of chirp period) and the subsequent local maximum points $\tau_2, ...\tau_W$ represent the reception of the reflected sounds (i.e., the possible beginning points of echo period).

However, although the local maximum points are roughly identified, it is still not applicable for accurate period selection due to that the detected local maximums in $MaxSet$ could be affected by the noise existed in the received signal. More importantly, the local maximums in $MaxSet$ that correspond to reflections from insignificant obstacles should be filtered out, and adjacent local maximums that correspond to reflections from close objects should also be merged. For these reasons, we utilize the fact that the direct received signal or signal reflections from significant obstacles should have a higher similarity value with the original beep signal. Thus, the 'important' local maximums that corresponds to them should hold a relatively higher value than other nearby local maximum points. So it is quite natural for us to think about whether we could use a sliding window to remove the 'unimportant' local maximums with relatively smaller values. In addition, another benefit of using the sliding window is that it could also help us to merge peaks corresponding to very close objects as one peak. Specifically, given the window length P and $E_{j,l} = \{E_l(t) \mid t \in [j, j+P]\}$, for all local maximums in $E_{j,l}$, we only keep the local maximum with the largest value and remove all other local maximums from the set $MaxSet$. The detailed algorithm of local maximum removal is provided in Algorithm 1.

Algorithm 1 Local Maximum Removal

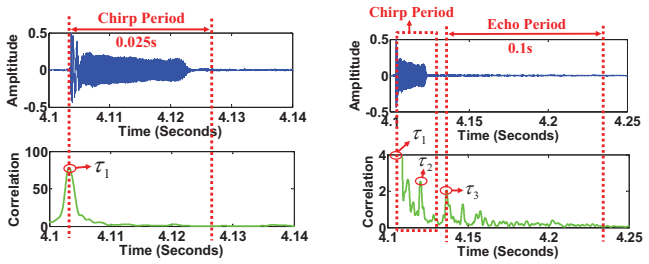
INPUT:
$E_l(t)$; *The envelope of $C_l(t)$*
$MaxSet = \{\tau_k \mid 1 \leq k \leq W\}$; *$W$ local maximum of $E_l(t)$*
P; *Length of the sliding window*
$T_{\max}$; *The maximum search range*
PROCEDURES:
for All $j \in [0, T_{\max} - P]$ **do**
  $E_{j,l} = \{E_l(t) \mid t \in [j, j+P]\}$;
  for All $k \in [1, W]$ **do**
    if $\tau_k \in [j, j+P] \& E_l(\tau_k) < max(E_{j,l})$ **then**
      delete τ_k from $MaxSet$
    end if
  end for
end for
Return $MaxSet$

Provided with the knowledge about the environmental information, we can then determine suitable values for P and $T_{\max}$. In this work, we deem that the two local maximums of two objects which are within about 1.5 meters should be merged. In addition, as the purpose of our work is to sense the surroundings for proximity detection, it is unnecessary to collect reflections from far-away objects (say roughly 30 meters away) in an indoor environment. Thus, in this work, given the sampling frequency of 48 kHz, we empirically set the $T_{\max}$ as 9600 samples and the P as 480 samples in Algorithm 1, respectively.

After the removal process, the 0.025s period (i.e., 0.02s chirp length plus 0.005s safeguard region) of the received signal after the first local maximum τ_1 derived from the matched filter, which corresponds to the sound which traveled directly from the device's speaker to the microphone, will be detected as chirp period and we denote it as $r_l(t)$. Its corresponding energy spectrum can thus be denoted as $R_l(f)$ via the fast Fourier transform (FFT). Next, the first identified local maximum τ_k ($k \in [2, W]$) after the chirp period will then be identified as the beginning point of the reception of signal reflections. Specifically, the 0.1s period after such τ_k will be detected as the echo period and we represent it as $r'_l(t)$. Similarly, its corresponding energy spectrum is denoted as $R'_l(f)$.

Example. Figure 3 shows an example on how the chirp period and echo period are selected using correlation values from a real experiment. Specifically, the original beep signal is slide across the acoustic readings with a moving window and the correlation values based on matched filter are calculated. The local maximum removal process has also been conducted to remove the 'unimportant' local maximums from the set $MaxSet$. From Figure 3(a), we can observe that the 0.025s period of the received beep signal after the first peak τ_1 is

Authorized licensed use limited to: Rutgers University. Downloaded on October 01,2021 at 19:51:48 UTC from IEEE Xplore. Restrictions apply.



(a) (b)

Fig. 3. Illustration of (a) chirp period selection and (b) echo period selection using peaks within correlation values derived from a matched filter.

identified as the chirp period. During this period, the acoustic signal propagates directly from the speaker to the microphone. Figure 3(b) shows the zoom-in view of the correlation values and we can observe that the first peak after the identified chirp period (i.e., τ_3) is detected as the beginning point of the echo period. Thus, the 0.1s period of the received beep signal after τ_3 can then be identified as the echo period. Such encouraging result confirms the feasibility of using our proposed scheme for the chirp and echo period selection.

C. Energy Loss Compensation Extraction using Chirp Periods

Due to different electronic features and manufacturing imperfections, the energy loss (i.e., the energy gain or attenuation measurement at each frequency) of microphones on the login device and the enrolled phone may vary significantly, making it hard to conduct a direct comparison between echo periods of the received beep signals for accurate proximity detection. Existing works [7]–[9] did not discuss on how to cope with such differences in acoustic sensing. However, from Section IV-B, we notice that the identified chirp periods actually contain information about the energy loss of microphones on two devices. Inspired by this observation, in this part, we design a new energy loss compensation extraction scheme to estimate the differences between microphones by utilizing chirp periods to achieve an accurate proximity detection for 2FA.

To simplify the description of our proposed algorithm, we first use A and B to denote the login device and the enrolled phone, respectively. We then use $R_{l,AB}(f)$ to represent the energy spectrum of chirp period extracted from the l-th received beep signal emitted by device A and received by device B, and similar expressions can also be derived for $R_{l,BA}(f)$, $R_{l,AA}(f)$ and $R_{l,BB}(f)$. Thus, we have the following equations by adopting the direct sound propagation model proposed in [6], [16]:

$$R_{l,AA}(f) = P_{l,A}(f)S_A(f)M_A(f)e^{\lambda(x_{AA})} \quad (2)$$

$$R_{l,BB}(f) = P_{l,B}(f)S_B(f)M_B(f)e^{\lambda(x_{BB})} \quad (3)$$

$$R_{l,AB}(f) = P_{l,A}(f)S_A(f)M_B(f)e^{\lambda(x_{AB})} \quad (4)$$

$$R_{l,BA}(f) = P_{l,B}(f)S_B(f)M_A(f)e^{\lambda(x_{BA})} \quad (5)$$

where $P_{l,A}(f)$ denotes device A's transmission energy at frequency f for the l-th beep signal (similar expressions can be derived for $P_{l,B}(f)$), $S_A(f)$ and $M_A(f)$ represent the energy

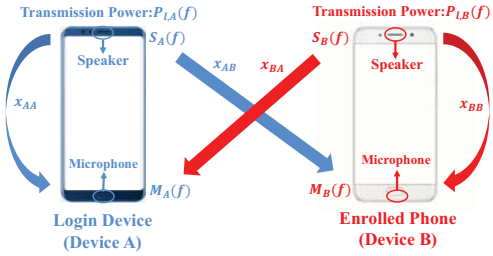


Fig. 4. An illustration of acoustic propagation model for the chirp period of the beep signal.

loss at frequency f of device A's speaker and the microphone respectively (similar expressions can be derived for $S_B(f)$ and $M_B(f)$), x_{AB} denotes the distance between device A's speaker and device B's microphone (similar expressions can be derived for x_{AA}, x_{BB} and x_{BA}) and $\lambda(x)$ is a function of distance x that can be derived by fitting the experimental data. Figure 4 shows such propagation model for clarity.

Direct deriving the relationship between $M_A(f)$ and $M_B(f)$ for energy loss compensation from above equations involves getting accurate values of $S_A(f)$, $S_B(f)$, $P_{l,A}(f)$, $P_{l,B}(f)$, x_{AA}, x_{BB}, x_{AB} and x_{BA}, which is very difficult. Thus, we propose a new method to derive such relationships without involving exact value estimation. In particular, we divide Equation (2) by Equation (4) and we can get:

$$\frac{R_{l,AA}(f)}{R_{l,AB}(f)} = \frac{P_{l,A}(f)S_A(f)M_A(f)e^{\lambda(x_{AA})}}{P_{l,A}(f)S_A(f)M_B(f)e^{\lambda(x_{AB})}} = \frac{M_A(f)e^{\lambda(x_{AA})}}{M_B(f)e^{\lambda(x_{AB})}} \quad (6)$$

Similarly, we divide Equation (3) by Equation (5):

$$\frac{R_{l,BB}(f)}{R_{l,BA}(f)} = \frac{P_{l,B}(f)S_B(f)M_B(f)e^{\lambda(x_{BB})}}{P_{l,B}(f)S_B(f)M_A(f)e^{\lambda(x_{BA})}} = \frac{M_B(f)e^{\lambda(x_{BB})}}{M_A(f)e^{\lambda(x_{BA})}} \quad (7)$$

Because sizes of devices are usually comparable, so we roughly have: $x_{AA} \approx x_{BB}$. In addition, it is also obvious that $x_{AB} \approx x_{BA}$. So from Equation (6) we have:

$$\frac{e^{\lambda(x_{BB})}}{e^{\lambda(x_{BA})}} \approx \frac{e^{\lambda(x_{AA})}}{e^{\lambda(x_{AB})}} = \frac{R_{l,AA}(f)}{R_{l,AB}(f)}\frac{M_B(f)}{M_A(f)} \quad (8)$$

We then put Equation (8) into Equation (7):

$$\begin{aligned}\frac{R_{l,BB}(f)}{R_{l,BA}(f)} &\approx \frac{M_B(f)}{M_A(f)}\frac{R_{l,AA}(f)}{R_{l,AB}(f)}\frac{M_B(f)}{M_A(f)} \\ &= \frac{R_{l,AA}(f)}{R_{l,AB}(f)}\left(\frac{M_B(f)}{M_A(f)}\right)^2\end{aligned} \quad (9)$$

We then have the compensation between $M_B(f)$ and $M_A(f)$ as follows:

$$M_B(f) \approx \sqrt{\frac{R_{l,BB}(f)}{R_{l,BA}(f)}\frac{R_{l,AB}(f)}{R_{l,AA}(f)}}M_A(f) \quad (10)$$

As described in Section III-C, the login device will encrypt its recorded microphone samplings using the public key and send them to the enroll phone. Thus, from Equation (10), the enrolled phone can estimate such compensation using its derived energy spectrums $R_{l,BB}(f)$ and $R_{l,AB}(f)$ along with energy spectrums $R_{l,BA}(f)$ and $R_{l,AA}(f)$ extracted from the login device's microphone samplings for energy loss compensation after the data decryption.

Authorized licensed use limited to: Rutgers University. Downloaded on October 01,2021 at 19:51:48 UTC from IEEE Xplore. Restrictions apply.

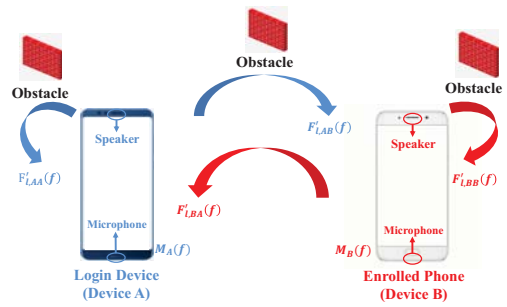


Fig. 5. An illustration of acoustic propagation model for the echo period of the beep signal.

D. Proximity Detection using Echo Periods

The energy spectrum of the echo period is characterized by uneven attenuations occurring at different frequencies. There are two main causes of this uneven attenuation: First, when the emitted beep sound reaches the surrounding objects, the surface material absorbs the signal at some frequencies and different materials have different absorption properties. Second, the combination of reflections makes the received signal constructive at some frequencies and destructive at other frequencies. This phenomena is also akin to the frequency selective fading in wireless communication. Thus, such energy spectrum contains rich information of the surrounding environment and the similarity between energy spectrums of echo periods derived from the login device and the enrolled phone could represent their proximity information. To capture this observation in a quantitative way, we propose to first conduct the energy loss compensation for microphones of two devices using the compensation estimation derived from Equation (10), and then compare the energy spectrums of echo periods using the correlation coefficient for accurate proximity detection.

1) Energy Loss Compensation: We let $R'_{l,AB}(f)$ denote the energy spectrum of the echo period derived from the l-th received beep signal emitted by the login device (i.e., device A) and received by the enrolled phone (i.e., device B) (similar expressions for $R'_{l,AA}(f)$, $R'_{l,BB}(f)$ and $R'_{l,BA}(f)$). Thus, they can be represented as:

$$R'_{l,AA}(f) = F'_{l,AA}(f)M_A(f) \tag{11}$$

$$R'_{l,BB}(f) = F'_{l,BB}(f)M_B(f) \tag{12}$$

$$R'_{l,AB}(f) = F'_{l,AB}(f)M_B(f) \tag{13}$$

$$R'_{l,BA}(f) = F'_{l,BA}(f)M_A(f) \tag{14}$$

where the $F'_{l,AB}(f)$ represents the energy spectrum of the reflected beep signal which is emitted by device A and just arrives the microphone of device B (similar expressions for $F'_{l,AA}(f)$, $F'_{l,BB}(f)$ and $F'_{l,BA}(f)$) and the detailed propagation model is shown in Figure 5.

Note that the basic idea of our scheme is to conduct the similarity comparison between $F'_{l,AA}(f)$ and $F'_{l,AB}(f)$ for proximity detection if the beep signal is emitted by device A. Ideally, from the relationships derived from Equation (11) and (13), this comparison could be conducted via comparing known energy spectrums $R'_{l,AA}(f)$ and $R'_{l,AB}(f)$ if $M_A(f) \approx M_B(f)$. However, due to $M_A(f)$ and $M_B(f)$ are unknown,

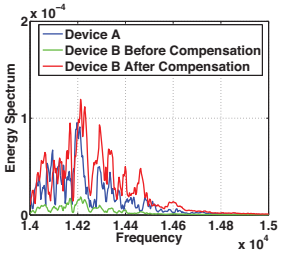


Fig. 6. An illustration of energy loss compensation for the login device (i.e., device A) and the enrolled phone (i.e., device B).

making it impossible for us to make accurate comparisons. Thus, we adopt the compensation between $M_B(f)$ and $M_A(f)$ derived from Equation (10) and reformulate the equation as:

$$R'_{l,AA}(f) = F'_{l,AA}(f)M_A(f) \tag{15}$$

$$R''_{l,AB}(f) = \frac{R'_{l,AB}(f)}{\sqrt{\frac{R_{l,BB}(f)}{R_{l,BA}(f)}\frac{R_{l,AB}(f)}{R_{l,AA}(f)}}} \approx F'_{l,AB}(f)M_A(f) \tag{16}$$

Note that the cubic spline interpolation [17] has been performed on the compensation to make its frequency resolution consistent with $R'_{l,AB}(f)$. From the above equations, we can observe that the similarity comparison between $F'_{l,AA}(f)$ and $F'_{l,AB}(f)$ can be conducted via comparing $R'_{l,AA}(f)$ and $R''_{l,AB}(f)$ because they are multiplied by the same energy loss $M_A(f)$. Thus, this new comparison compensates the difference between the energy loss $M_A(f)$ and $M_B(f)$ using the compensation estimated in Equation (10). We further refer to $R'_{l,AA}(f)$ and $R''_{l,AB}(f)$ as location signatures and repeat such compensation procedure to derive location signatures $R'_{l,BB}(f)$ and $R''_{l,BA}(f)$ when the beep signal is emitted by device B. To capture the pattern of all the beep signals, we further average all location signatures over L beep signals to derive the average location signatures: $\bar{R}'_{AA}(f)$, $\bar{R}''_{AB}(f)$, $\bar{R}'_{BB}(f)$ and $\bar{R}''_{BA}(f)$.

2) Similarity Comparison: We propose to use the Pearson correlation coefficient [18] to conduct the similarity comparison between location signatures for proximity detection. However, the presence of random noise may add small variations to location signatures, making such comparison inaccurate. To cope with this problem, we propose to use an average filter to remove such small variations. Specifically, we set the number of frequency points in the average filter as 20 and compute the Pearson correlation coefficient c_A between the filtered average location signatures $\bar{R}'_{AA}(f)$ and $\bar{R}''_{AB}(f)$ (c_B for $\bar{R}'_{BB}(f)$ and $\bar{R}''_{BA}(f)$, resp) for proximity detection. If the average value of c_A and c_B is higher than a pre-defined threshold c_{th}, the system will declare that two devices are in close proximity and the login request will be accepted.

Feasibility Study. We provide a feasibility study on how the energy spectrum changes before and after the energy loss compensation for device A and device B. Specifically, we place two smartphones in close proximity, use them as device A and B respectively and collect 10 beep periods for energy loss compensation purpose. The average energy spectrums $\bar{R}'_{AA}(f)$, $\bar{R}'_{AB}(f)$ and $\bar{R}''_{AB}(f)$ over 10 beep signals are then derived from the received signal and presented

Authorized licensed use limited to: Rutgers University. Downloaded on October 01,2021 at 19:51:48 UTC from IEEE Xplore. Restrictions apply.

in Figure 6. Before the energy loss compensation, due to the energy loss of microphones (i.e., $M_A(f)$ and $M_B(f)$) differs significantly, $\bar{R}'_{AA}(f)$ and $\bar{R}'_{AB}(f)$ varies and their corresponding correlation value is only 0.42. However, after the energy loss compensation process, the energy spectrums $\bar{R}'_{AA}(f)$ and $\bar{R}''_{AB}(f)$ become very similar and such correlation value increases significantly (i.e., from 0.42 to 0.91). These observations strongly confirm the feasibility of using our proposed energy loss compensation scheme to conduct an accurate proximity detection.

V. Performance Evaluation

In this section, we conduct experiments to evaluate the performance of our 2FA system over a period of six months.

A. Experimental Setup

We use a ThinkPad X280 laptop along with four smartphones including Samsung Galaxy s7, Huawei Mate 10, Huawei Mate 30 and Honor 10 for evaluations. These devices differ in both RAM sizes or processors and the detailed information of each device is shown in Table I. During the experiment, we set the bandwidth of the beep signal as 14 to 15 kHz with a length of 0.02s as described in Section IV-A. We conduct our experiments under three representative environments: the medium sized conference room with some chairs and tables (i.e., *conference room*), the spacious office room with a large number of cubicles or desks (i.e., *office*) and the long hallway with few tables (i.e., *hallway*). Unless otherwise specified, the results presented in this work are using the acoustic data collected from the conference room. We then develop applications to collect the acoustic data which is written into a sound file stored in the smartphone or laptop during the authentication process.

1) Evaluation Scenarios: We evaluate our system under three scenarios including one regular authentication scenario and two representative attack scenarios.

Regular Authentication: A legitimate user is told to place his enrolled phone besides the login device and tries to pass the 2FA process after inputting his own username and password. In this scenario, we use the X280 laptop as the login device and other smartphones as enrolled phones.

Man-in-the-middle (MiM) Attack: A remote adversary sets up a high-speed channel between the victim's enrolled phone and the adversarial login device and relay probing beep signals between them in attempt to pass our proposed 2FA. More specifically, in this scenario, we use one Mate 30 smartphone as the enrolled phone and two iPhone 6s smartphones as relay devices to launch the MiM attack.

Co-located Attack: An adversary is physically co-located with the victim. The adversary's login attempts triggers both the login device and the victim's enrolled phone to emit probing beep signals for 2FA, which can be directly received by both devices' microphones. More specifically, in this scenario, we use one Mate 10 smartphone as the adversary login device and one Mate 30 smartphone as the victim's enrolled phone to perform the co-located attack.

Device	Processor	Memory	OS
Galaxy s7	Exynos 8890 Octa	4 GB	Android 6.0
Mate 10	Kirin 970	4 GB	Android 8.0
Mate 30	Kirin 990	6 GB	Android 10
Honor 10	Kirin 970	6 GB	Android 8.1
X280	i7 – 8650U	8 GB	Win 10

TABLE I
Summary of experimental devices.

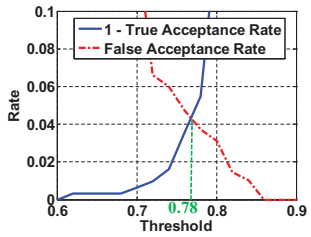


Fig. 7. The $1 - TAR$ and FAR as a function of the threshold c_{th}. The EER is about 0.043 at $c_{th} = 0.78$.

2) Metrics: We use the following metrics to evaluate the effectiveness of our 2FA system.

True Acceptance Rate (TAR): the ratio of the number of legitimate login attempts accepted by our system to the total number of legitimate login attempts.

False Acceptance Rate (FAR): the ratio of the number of fraudulent login attempts accepted by our system to the total number of fraudulent login attempts.

Equal Error Rate (EER): it is defined as the rate at which the FAR is equal to one minus TAR (i.e., $FAR = 1 - TAR$). The EER shows the trade-off between two error types and it can help us to choose the value of threshold c_{th} via a statistical study.

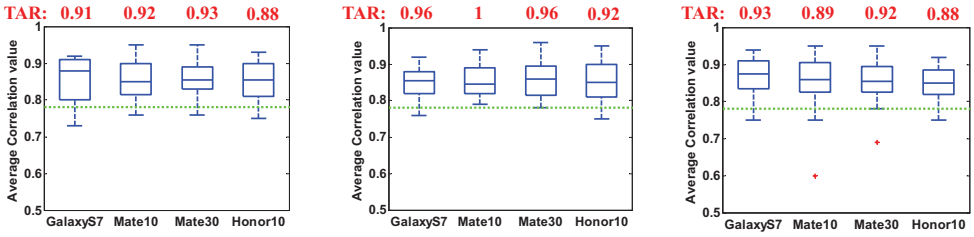
B. Impact of Threshold Settings

In the first set of experiments, we evaluate the effectiveness of our Proximity-Echo by adopting different threshold values for user authentication. Specifically, we evaluate the system performance by varying the threshold c_{th} from 0.6 to 0.9. Figure 7 plots the $1 - TAR$ and FAR when varying the value of threshold c_{th}. The TAR and FAR are calculated from correlation values derived from the regular authentication scenario and attack scenarios, respectively. We observe that the $1 - TAR$ increases whereas the FAR decreases as the value of threshold increases. This is because with a higher detection threshold, fewer login attempts (regardless of legitimate or fraudulent attempts) could be accepted by our system. In addition, the crossing point of $1 - TAR$ and FAR shows that our system has around 0.043 EER when the threshold is set as about 0.78, which shows a good trade-off between TAR and FAR. Thus, unless otherwise specified, we choose 0.78 as the threshold c_{th} in this work.

C. Robustness to Experimental Environments

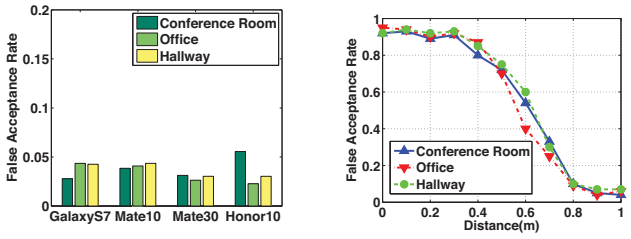
We next study the robustness of Proximity-Echo for the regular authentication scenario when experiments are conducted under different environments. In this study, four smartphones including Galaxy s7, Mate 10, Mate 30 and Honor 10 are used as enrolled phones and the number of beep signals is set as 20.

Authorized licensed use limited to: Rutgers University. Downloaded on October 01,2021 at 19:51:48 UTC from IEEE Xplore. Restrictions apply.



(a) Conference room (b) Office (c) Hallway

Fig. 8. Robustness study under different experimental environments.



(a) MiM Attack (b) Co-located Attack

Fig. 9. Performance study under different attacks.

Figure 8 (a) to (c) display the distribution of correlation values when the experiments are conducted under different environments. We observe that most correlation values remain higher than the threshold c_{th} (i.e., 0.78 as the green dot line shown in the figure) across all environments and enrolled phone usages. In addition, these figures demonstrate that our system can achieve a satisfactory TAR (i.e., over 0.88) in all scenarios. These observations illustrate that our system is effective in 2FA and robust across different environments.

D. Performance Evaluation Under the MiM Attack

We further evaluate our Proximity-Echo under MiM attacks. Specifically, without loss of generality, we choose one Mate 30 smartphone as the victim enrolled phone and two iPhone 6s smartphones as relay devices to conduct the MiM attack. The victim's enrolled phone is also placed under three experimental environments and the adversary login device is placed in a separate room which is far away from the enrolled phone.

Figure 9(a) presents the FAR of MiM attacks under different environments. We observe that the overall FAR remains less than about 0.05 across all scenarios. Further, this figure also shows that the performances from different scenarios are comparable, indicating our system operating 2FA is robust to MiM attacks under different experimental environments and phone models. This is due to the fact that the login device would obtain the environmental characteristics of the adversary's location instead of the victim's location via the reflected beep signal under MiM attacks. Such illegitimate environmental patterns differ significantly from the victim's real environmental patterns and it cannot pass the 2FA. Therefore, the MiM attack can be thwarted effectively.

E. Performance Evaluation Under the Co-located Attack

Finally, we study the performance of our Proximity-Echo under co-located attacks. In this study, we use one Mate 30 smartphone as the victim enrolled phone and one Mate 10 smartphone as the adversary login device to launch the co-located attack. We vary the distance between two devices from 0 to 1 meters with a step length of 0.1 meters and run the 2FA for each distance.

Figure 9(b) presents the FAR under co-located attacks when the distance between the victim enrolled phone and the adversary login device increases from 0 to 1 meter. We observe that the FAR remains higher than 0.9 when the distance is less than 0.1 meters. However, the FAR drops significantly as the distance increases and it remains lower than 0.1 when the distance is larger than 0.8 meters. It indicates that a distance of 0.8 meters is sufficient to thwart the co-located attack and almost none of the login attempts can pass the authentication process. Further, we observe that a slightly higher FAR is achieved when experiments are conducted in the hallway. This is because the reflected acoustic signal contains less information of the surrounding characteristics in an indoor environment with a smaller number of obstacles. These results show that our system is secure against the co-located attacks.

VI. Conclusion

In this paper, we present Proximity-Echo, a secure system leveraging the proximity of a user's enrolled phone and the login device as the second proof for 2FA. The proposed system extracts location signatures by emitting acoustic beep signals alternately by both devices with speakers and sensing the reflections with microphones, and compares the extracted signatures for proximity detection. Our designed period selection scheme identifies two sound segments named the chirp period and the echo period accurately from the received beep signal. To achieve an accurate proximity detection, our system further develops a new energy loss compensation extraction scheme by using the identified chirp periods to estimate the intrinsic differences of energy loss between microphones of the enrolled phone and the login device. Moreover, our proximity detection component conducts the similarity comparison between the identified echo periods to determine whether two devices are in proximity for 2FA. Extensive experiments are conducted to show that our proposed system is accurate in providing 2FA and robust to both man-in-the-middle (MiM) and co-located attacks across different scenarios and device models.

Acknowledgments: This work is supported in part by National Natural Science Foundation of China under Grants 61802051, 62020106013, Sichuan Science and Technology Program under Grants 2020$JDTD$0007 and 2020YFG0298.

Authorized licensed use limited to: Rutgers University. Downloaded on October 01,2021 at 19:51:48 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] "How Hackers Steal Passwords," https://cybriant.com/heres-how-hackers-steal-passwords/, 2020.

[2] "Duo Mobile App," https://duo.com/product/multi-factor-authentication-mfa/duo-mobile-app, 2020.

[3] "Google 2-step Verification," https://www.google.com/landing/2step/, 2020.

[4] N. Karapanos, C. Marforio, C. Soriente, and S. Čapkun, "Sound-proof: Usable two-factor authentication based on ambient sound," in *Proceedings of the 24th USENIX Conference on Security Symposium*, 2015.

[5] B. Shrestha, M. Shirvanian, P. Shrestha, and N. Saxena, "The sounds of the phones: Dangers of zero-effort second factor login based on ambient audio," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, 2016.

[6] D. Han, Y. Chen, T. Li, R. Zhang, Y. Zhang, and T. Hedgpeth, "Proximity-proof: Secure and usable mobile two-factor authentication," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, 2018.

[7] Y.-C. Tung and K. G. Shin, "Echotag: Accurate infrastructure-free indoor location tagging with smartphones," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, 2015.

[8] Q. Song, C. Gu, and R. Tan, "Deep room recognition using inaudible echos," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2018.

[9] S. Pradhan, G. Baig, W. Mao, L. Qiu, G. Chen, and B. Yang, "Smartphone-based acoustic indoor space mapping," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2018.

[10] Z. Wang, S. Tan, L. Zhang, and J. Yang, "Obstaclewatch: Acoustic-based obstacle collision detection for pedestrian using smartphone," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2018.

[11] C. Alexei, D. Michael, K. Tadayoshi, W. Dan, and B. Dirk, "Strengthening user authentication through opportunistic cryptographic identity assertions," in *Proceedings of the ACM Conference on Computer and Communications Security*, 2012.

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

[13] M. Shirvanian, S. Jarecki, N. Saxena, and N. Nathan, "Two-factor authentication resilient to server compromise using mix-bandwidth devices," in *Proceedings of the Network and Distributed System Security (NDSS) Symposium*, 2014.

[14] S. P. Tarzia, P. A. Dinda, R. P. Dick, and G. Memik, "Indoor localization without infrastructure using the acoustic background spectrum," in *ACM MobiSys*, 2011.

[15] Y. Ren, C. Wang, Y. Chen, J. Yang, and H. Li, "Noninvasive fine-grained sleep monitoring leveraging smartphones," *IEEE Internet of Things Journal*, 2019.

[16] D. Chen, N. Zhang, Z. Qin, X. Mao, Z. Qin, X. Shen, and X. Li, "S2m: A lightweight acoustic fingerprints-based wireless device authentication protocol," *IEEE Internet of Things Journal*, 2017.

[17] R. V. Dukkipati, *Numerical Methods*. New Age International Pvt Ltd Publishers, 2010.

[18] G. Casella, R. Berger, and R. Berger, *Statistical inference*. Duxbury Press Belmont, Calif, 1990.

Authorized licensed use limited to: Rutgers University. Downloaded on October 01,2021 at 19:51:48 UTC from IEEE Xplore. Restrictions apply.