Secure and Robust Two Factor Authentication via Acoustic Fingerprinting

Yanzhi Ren[1], Tingyuan Yang[1], Zhiliang Xia[1], Hongbo Liu[1], Yingying Chen[2]
Nan Jiang[3], Zhaohui Yuan[3], Hongwei Li[1]
[1]Dept. of CS, University of Electronic Science and Technology of China [2]WINLAB, Rutgers University, USA
[3]East China Jiaotong University
[1]{renyanzhi05, hongweili}@uestc.edu.cn [2]yingche@scarletmail.rutgers.edu

IEEE INFOCOM 2023 - IEEE Conference on Computer Communications | 979-8-3503-3414-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/INFOCOM53939.2023.10229054

***Abstract*—The two-factor authentication (2FA) has become pervasive as the mobile devices become prevalent. Existing 2FA solutions usually require some form of user involvement, which could severely affect user experience and bring extra burdens to users. In this work, we propose a secure 2FA that utilizes the individual acoustic fingerprint of the speaker/microphone on enrolled device as the second proof. The main idea behind our system is to use both magnitude and phase fingerprints derived from the frequency response of the enrolled device by emitting acoustic beep signals alternately from both enrolled and login devices and receiving their direct arrivals for 2FA. Given the input microphone samplings, our system designs an arrival time detection scheme to accurately identify the beginning point of the beep signal from the received signal. To achieve a robust authentication, we develop a new distance mitigation scheme to eliminate the impact of transmission distances from the sound propagation model for extracting stable fingerprint in both magnitude and phase domain. Our device authentication component then calculates a weighted correlation value between the device profile and fingerprints extracted from run-time measurements to conduct the device authentication for 2FA. Our experimental results show that our proposed system is accurate and robust to both random impersonation and Man-in-the-middle (MiM) attack across different scenarios and device models.**

I. Introduction

The mobile two-factor authentication (2FA) becomes pervasive as mobile devices (e.g., smartphones, tablets, wearables) play increasingly significant roles in our daily lives. When a user logs into an online system employing 2FA using his/her username and password, the system will further verify whether the user has a pre-enrolled device (e.g., smartphone or tablets), and further use it as the second security proof to protect the online accounts [1]. For instance, when a user tries to log into an online shopping account, after receiving the username and password, the system will further verify the user's possession of his/her pre-registered mobile device. Thus, such online system can still keep safe even if the user's username and password have been compromised.

However, in most commercial 2FA systems, certain user involvement is usually required. For example, the leading online 2FA service such as Duo Mobile App and Google 2-step Verification [2], [3], either call the enrolled phone for the user's answer or send a notification message for the user to approve on the screen to pass the authentication process. Similar authentication methods are also adopted in many other mobile 2FA solutions. Such requirements for users' active participation could seriously affect the user's experience, e-specially for the senior or disabled people.

To improve the usability of mobile 2FA, some studies have been proposed by eliminating the explicit user interaction. For instance, some technique [4], [5] designs a secure challenge-response protocols over a Bluetooth channel. However, the Bluetooth functionalities may not be supported by standard web browsers or micro devices. Later, to avoid this weakness, some acoustic-based methods have been explored. Sound-Proof [6] utilizes the ambient sound to detect the proximity of the enrolled phone and login device without user interaction for 2FA. However, this scheme is vulerable to the adversary if he/she could generate similar ambient sound at the login device's end [7]. Proximity-Proof [8] and SoundID [9] show the initial success of mobile 2FA via device fingerprinting. They propose techniques to derive the acoustic fingerprints for the acoustic elements on mobile devices. However, the extracted fingerprints differ significantly when the environments or transmission distance change, which could threaten the accuracy of these approaches.

When the sound propagates directly from the speaker to the microphone in the air, the received acoustic signal should contain unique fingerprint information of the devices. Such acoustic fingerprints, if captured, could be utilized as the proof for authentication. These observations trigger our idea from acoustic fingerprinting [10], [11] to use the received beep sound, which contain rich information of acoustic elements on mobile devices, as the authentication factor for 2FA. Specifically, in this paper, we propose a mobile 2FA system that uses acoustic fingerprints of the speaker/microphone on enrolled device as the second authentication factor.

However, several unique challenges need to be addressed when developing such a system: First, the received beep signal could be significantly affected by the transmission distance between the speaker and microphone pairs of devices, requiring our scheme to be able to extract robust fingerprints for accurate 2FA even though the distance varies. Second, due to the nature of sound transmitting and receiving, the received signal is always tied to a pair of speaker and microhone on devices, posing significant difficulty to extract fingerprints for individual speaker or microphone. Third, the beep signal is relatively weak and could also be significantly affected by the interference caused by multi-path environments, making it a challenging task to identify the beginning point of beep

segment from the received signal accurately. Last but not least, our proposed 2FA process should be completed with the minimal efforts without the involvement of the user's extra interactions.

To cope with these challenges, our proposed system consists of three main components: Arrival Time Detection, Acoustic Fingerprinting and Device Authentication. Given the input microphone samplings, our arrival time detection component proposes a cumulative sum based technique to accurately identify the beginning point of beep signal. Acoustic fingerprinting component is the core component that could mitigate the effect of transmission distance from the sound propagation model, and then extract both magnitude and phase fingerprints to capture the unique and robust frequency response patterns of acoustic elements on the enrolled device. Such fingerprints are only tied to the individual microphone or speaker and it remains invariant even if the distance between two devices changes. During the device authentication, a weighted correlation value between the device profile and fingerprints extracted from run-time measurements is calculated to conduct the authentication for 2FA. We summarize our main contributions as follows:

- We design a new mobile 2FA system which utilizes the acoustic fingerprint of individual speaker/microphone on enrolled device as the second proof.
- We propose to use both magnitude and phase fingerprints, which are derived from the frequency response of the enrolled device by emitting acoustic beep signals alternately from both enrolled and login devices and receiving their direct arrivals, as the proof for 2FA.
- We design a sequential change-point detection scheme to accurately identify the beginning point of beep signal from the received signal by computing the cumulative sum of difference to the ambient noise level.
- To achieve a robust authentication, we develop a new distance mitigation scheme to eliminate the impact of transmission distances from the sound propagation model in both magnitude and phase domain for stable fingerprint extraction.
- We show that our system is robust to adversarial behaviors of impersonating the enrolled phone with other mobile devices, or relaying signals between the enrolled phone and a remote adversarial login device in an attempt to pass our 2FA.
- Our extensive experimental results show that our proposed 2FA is accurate and robust across different device models under various real world scenarios.

II. Related Work

There have been some software based authentication schemes such as Google 2-step Verification or Duo Mobile App [2], [3] for 2FA. In these systems, a passcode is sent to the enrolled device and the user then inputs it on an interface to finish the login process. These mechanisms can be easily integrated with online systems. However, such schemes require users to interact with their devices explicitly and it may severely affect the user experience and bring extra burdens to them.

To improve the usability of 2FA mechanism, some studies have been proposed by eliminating the need of user interaction. In [4], [5], the Bluetooth channel between the login device and the enrolled phone is utilized to design a challenge-response based protocol for 2FA without user involvement. However, the Bluetooth function is required for this scheme and it is not always available for all login devices. Along this direction, in [6], the ambient sound is used as the proximity proof between devices for 2FA. However, this scheme may become invalid if the adversary could generate similar ambient sound at the login device's end.

There are also recent works dedicated for acoustic device fingerprinting. Zhou et al [10], Anupam et al [11] and Chen [12] studied the device authentication/identification in detail and they also show that the energy loss could be used as the acoustic fingerprint of the device. However, the extracted fingerprint is always associated with a pair of speaker and microphone (i.e., the emitting speaker and the recording microphone) rather than with an individual speaker or microphone. In addition, these schemes also cannot efficiently handle the effect of the variable transmission distances between the sender and receiver, and the system must obtain the acoustic fingerprint associated with multiple distances in the enrollment phase, which is very burdensome. SoundID [9] develop a scheme that can derive acoustic fingerprints to defend against the man-in-the-middle (MITM) attack. However, the extracted fingerprints is dynamic and it could differ significantly when the environments changes.

The most similar work to our own is by Proximity-proof [8]. They propose a technique to derive the acoustic fingerprint with frequency response for the acoustic elements on mobile devices. However, the extracted fingerprint also differs with the change of transmission distance, which may threaten the accuracy of this approach. Unlike the aforementioned work, we aim to develop an authentication mechanism which can mitigate the effect of transmission distance in fingerprint extraction to achieve a robust and accurate 2FA. Our proposed system does not need the user's explicit participation and is also easy-to-use without requiring any dedicated hardware.

III. Framework Overview

In this section, we introduce the adversary model and provide an overview of our proposed system.

A. Adversary Model

We assume that the victim's username and password have been compromised by an adversary, with which he/she attempts to log into the victim's account on an arbitrary networked device. This attack is successful if the adversary can convince the system that he/she has the enrolled phone associated with the victim. As prior works [6], [8], [13], we also adopt the following assumptions: First, the communication channel between the server and devices is secured using the TLS-like mechanisms. Second, the victim's enrolled

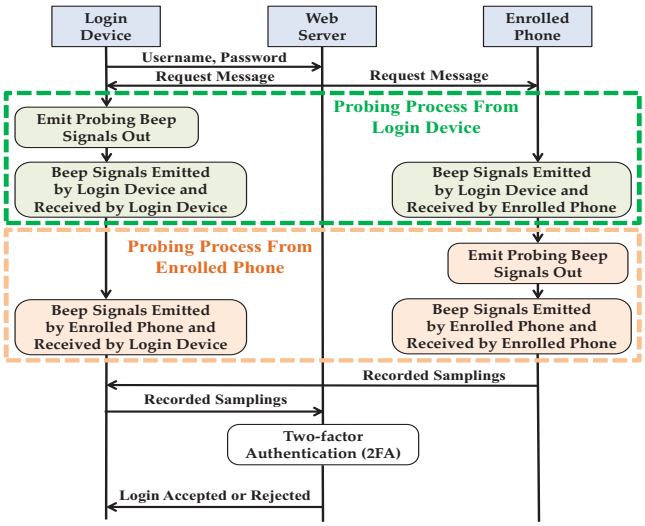


Fig. 1. 2FA system model.

phone cannot be compromised by the adversary, otherwise the 2FA system reduces a regular password based authentication system. Third, we assume that the adversary only utilizes commercial off-the-shelf (COTS) mobile devices to launch attacks since these devices could pose more realistic threats to our system. Forth, since the 2FA is a time sensitive task, the timing information is always included in the challenge and response messages of both devices. So we do not consider the replay attack in this work. Fifth, the legitimate user should always possess his enrolled phone. Thus, there should be a line-of-sight channel between the login device and the enrolled phone. Last but not least, according to [14], no device fingerprinting scheme alone can well defeat the co-located attack. So we also do not consider this attack in this work and it can be addressed by the cross-device ranging method proposed in [8]. Based on these assumptions, we consider two possible adversarial behaviors as described below:

- **Random Impersonation:** The adversary impersonates the enrolled phone with his/her own mobile devices. The adversary could even obtain the detailed model information of the enrolled phone, and uses a device of the same model to launch the attack.
- **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.

B. System Overview

As prior works [6], [7], we consider a standard mobile 2FA system model to lay out the foundation for the subsequent illustration. As shown in Figure 1, the user logs into the system via a login device, and it can be any networked device such as a smartphone, a tablet, a laptop or even a public computer. When attempting to login, the user is required to input the corresponding username and password on the interface of the login device, and they are relayed to the server via a secure channel. The server then checks the validity of the username and password, and sends request messages to the login device and the enrolled device which is associated with the username to validate the second authentication factor (i.e., the acoustic fingerprint). Specifically, the login device and the enrolled phone emit probing beep signals alternatively, and they are then received by both devices' microphones. Both the enrolled phone and the login interface then encrypts the recorded microphone samplings using their public key and sends them to the server using the login device as a proxy. The server then decrypts these received samplings to perform our proposed 2FA. If the authentication process passes, the server informs the login device to accept the login process.

As shown in Figure 2, our system consists of three major components: *Arrival Time Detection*, *Acoustic Fingerprinting* and *Device Authentication*. The system takes as input the recorded microphone samplings from both the enrolled phone and the login device. In the arrival time detection phase, to deal with the interference caused by multi-path environments, a sequential change-point detection technique is proposed to accurately identify the beginning point of the beep signal from the received signal. Given the identified beep signal, in acoustic fingerprinting component, our system first mitigates the effect of transmission distance from the sound propagation model, and then extracts both magnitude and phase fingerprints to capture the unique and robust frequency response patterns of acoustic elements on the enrolled device. Such extracted fingerprint is usually only tie to the individual microphone or speaker and it remains invariant even if the distance between two devices changes. After the fingerprint extraction, the device authentication is performed by calculating a weighted correlation value between the device profile and fingerprints extracted from run-time measurements. Based on such 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 proposed system.

A. Parameter Setting of the Beep Signal

In this work, we choose the Linear-Frequency Modulated (LFM) chirp signal as our beep signal. LFM chirp is a type of modulated sinusoidal signal whose instantaneous frequency increases or decreases linearly over time. When designing the probing beep signal played through the speaker for 2FA, we mainly consider three factors: frequency band, length and time interval.

Frequency Band. The frequency response of acoustic elements are quite uneven and drastically different in the high frequency range (e.g.,higher than 14 kHz) [12]. In addition, due to the hardware's imperfection, the performance of most mobile devices decays quickly when the frequency is beyond

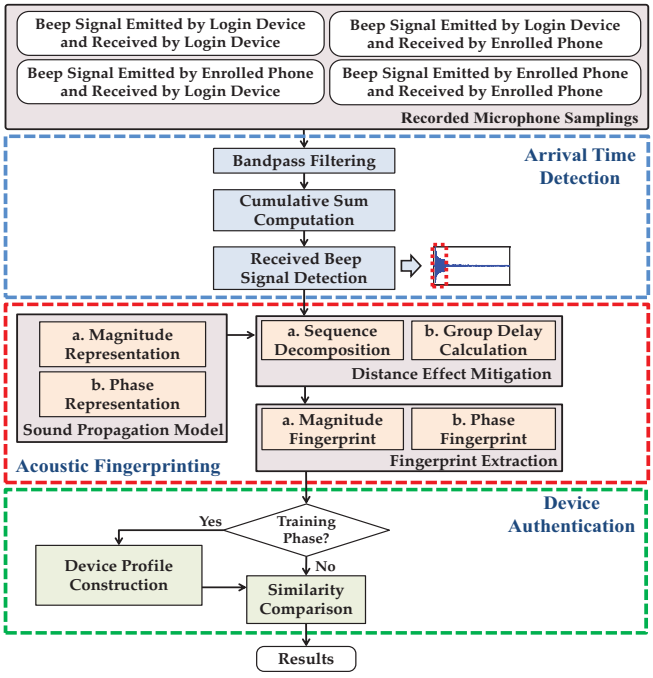


Fig. 2. System flow of our system.

18 kHz [15]. In summary, given the trade-off of all factors, our system adopts the 14 kHz to 15 kHz bandwidth beep acoustic signal.

Although this selection makes the beep signal audible to humans, the impact of this selection is minimal since our system triggers the authentication process infrequently. Moreover, we further conduct a post-use survey after our experimental evaluation and the results show that most participants in our usability study did not consider the emitted sound annoying in the authentication process.

Length. The length of beep signal also impacts the accuracy and reliability of our authentication system. The acoustic elements on mobile devices usually cannot generate or pick up too short beep signals. However, setting a too long beep signal may cause severe multipath distortions since reflections from surrounding objects will also be collected during this long sensing process [16]. Thus, in this work, we empirically set the length of the beep signal as 0.02s.

Time Interval. The third parameter 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 authentication, and a short interval may cause errors since the transmitted signals might accidentally overlap with each other. Based on our observations, we set the interval to be 0.5s in this work.

B. Detecting Arrival Time of the Beep Signal

Our system aims to use the received beep signal propagating directly from the speaker to the microphone, which contains unique fingerprint information of mobile devices, as the proof for 2FA. However, detecting the arrival of beep signal is challenging because the received signal could be easily affected by the interference caused by multi-path environments. In particular, the commonly used correlation based techniques, which detects the maximum correlation point between the received signal and the original beep signal, is also susceptible to such distortions.

To solve this problem, in this paper, we propose a sequential change-point detection scheme. The basic idea of our scheme is to identify the first strong signal in our frequency band that deviates from the background noise, and its corresponding point allow us to detect the arrival time of the beep signal accurately. Specifically, the received signal is first sampled with a frequency of 48 kHz, and a bandpass filter with lower and upper cutoff frequencies of 14 kHz and 15 kHz respectively is then 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. Initially, without the beep signal, the received signal only contains background noise, which follows a distribution with the density function of p_0. Later on, at certain time t_p, the distribution changes to the density function of p_1 due to the arrival of beep signal. To identify t_p, we formulate the problem as sequential change-point detection by computing the cumulative sum of difference to the averaged noise level. Specifically, suppose that the mobile device first estimates the mean value μ of the background noise before transmitting the beep, the cumulative sum of difference $s(t)$ over a window of length W at time t is calculated as:

$$s(t) = \frac{1}{W}\int_{t}^{t+W} |e_l(t) - \mu| dt \tag{1}$$

Intuitively, the cumulative sum should have a small and stable drift if the received signal $e_l(t)$ is smaller than the mean value of the background noise, and a large drift after the presence of the beep signal. Thus, the 'change-point' that corresponds to the beep arrival should be the earliest point whose cumulative sum is larger than a threshold, and this trend remains the same for its subsequent points. So the arrival time of the beep signal is given by:

$$t_p = \inf(t\,|s(\tau) \geq h, \forall\tau \in [\mathrm{t}, \mathrm{t}+\mathrm{W}]) \tag{2}$$

where h is the threshold, W is the window used to reduce the false alarm. After this detection process, the 0.02s period (i.e., the length of the beep signal) of the received signal after the identified arrival time t_p, which corresponds to the sound segment which traveled directly from the speaker to the microphone, will be detected as the received beep signal and we denote it as $r_l(t)$. Its corresponding frequency-domain representation could thus be denoted as $R_l(f)$ via the fast Fourier transform (FFT).

Example. Figure 3 shows an example on how the arrival time of the beep signal is detected using our proposed sequential change-point detection scheme. Specifically, a moving window is slide across the acoustic readings and the cumulative sum of difference is calculated. The arrival time of the beep signal t_p is then determined by searching for the earliest point at which the signal significantly deviates from

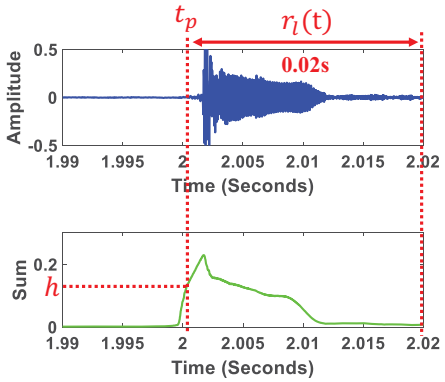


Fig. 3. An illustration of received beep signal detection.

the noise using Equation (2). From Figure 3, we can observe that the received beep signal $r_l(t)$ can be accurately identified as the 0.02s period of the received beep signal after t_p. In addition, during this period, the acoustic signal propagates directly from the speaker to the microphone. Such encouraging result confirms the feasibility of using our proposed scheme for detecting the arrival time of beep signal.

C. Sound Propagation Model

Sound is a sequence of waves which propagates through a transmission medium such as air or water. During the sound propagation, sound waves would be attenuated and delayed by the medium. To simplify the description of our propagation model, we first use A and B to denote the enrolled phone and the login device, respectively. We then use $R_{l,AB}(f)$ to represent the Fourier transform of 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 magnitude representation of the Fourier transform [17] by adopting the direct sound propagation model [8], [12] as follows:

$$|R_{l,AA}(f)| = |P_{l,A}(f)|\,|S_A(f)|\,|M_A(f)|\,e^{\lambda(x_{AA},f)} \quad (3)$$

$$|R_{l,BB}(f)| = |P_{l,B}(f)|\,|S_B(f)|\,|M_B(f)|\,e^{\lambda(x_{BB},f)} \quad (4)$$

$$|R_{l,AB}(f)| = |P_{l,A}(f)|\,|S_A(f)|\,|M_B(f)|\,e^{\lambda(x_{AB},f)} \quad (5)$$

$$|R_{l,BA}(f)| = |P_{l,B}(f)|\,|S_B(f)|\,|M_A(f)|\,e^{\lambda(x_{BA},f)} \quad (6)$$

where $P_{l,A}(f)$ denotes the Fourier transform of the transmission signal for the l-th beep (similar expressions can be derived for $P_{l,B}(f)$), $S_A(f)$ and $M_A(f)$ represent the frequency response 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 $\lambda(x, f)$ is a function of distance x and frequency f. Specifically, the acoustic waves traveling follows the power-law attenuation and thus the $\lambda(x, f)$ can be represented as $\lambda(x, f) = -\alpha_0(2\pi f)^\eta x$ where parameters α_0 and η can be obtained by fitting the experimental data [12]. Figure 4 shows such propagation model for

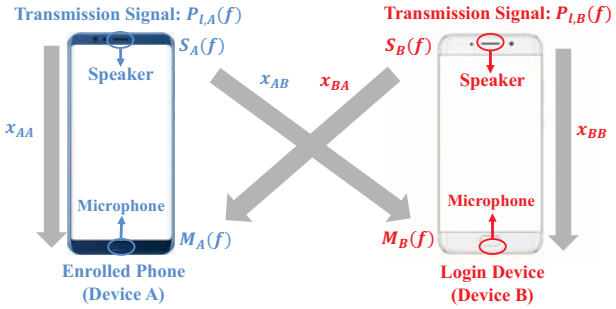


Fig. 4. An illustration of acoustic propagation model of the beep signal.

clarity. From Equation (3) to (6), we can observe that the effect of transmission distance in magnitude representation could be modeled as an exponential trend function of distance x and frequency f (i.e., $e^{\lambda(x,f)}$).

Similarly, the phase representation of the Fourier transform [17] of the direct sound propagation model [8], [12] can be represented as:

$$\angle R_{l,AA}(f) = \angle P_{l,A}(f) + \angle S_A(f) + \angle M_A(f) - \frac{x_{AA} f}{c} \quad (7)$$

$$\angle R_{l,BB}(f) = \angle P_{l,B}(f) + \angle S_B(f) + \angle M_B(f) - \frac{x_{BB} f}{c} \quad (8)$$

$$\angle R_{l,AB}(f) = \angle P_{l,A}(f) + \angle S_A(f) + \angle M_B(f) - \frac{x_{AB} f}{c} \quad (9)$$

$$\angle R_{l,BA}(f) = \angle P_{l,B}(f) + \angle S_B(f) + \angle M_A(f) - \frac{x_{BA} f}{c} \quad (10)$$

where c represents the speed of sound in the air, $\frac{x}{c}$ denotes the delay of the sound propagation for the distance of x. Thus, from Equation (7) to (10), we can observe that the effect of transmission distance (or equivalently, non-synchronization) in phase representation could be modeled as a linear function and its corresponding slope represents the length of the time delay.

D. Acoustic Fingerprinting

The imperfect manufacturing process could introduce unique electronic features to each acoustic element (e.g., the microphone or speaker) on mobile devices, making it feasible to utilize their frequency response as acoustic fingerprints for authentication [8], [10]–[12]. In addition, as illustrated in Section IV-C, the frequency response is a complex number, which can be represented in terms of its magnitude and phase. However, almost all existing works [8], [10]–[12] either only consider to use the magnitude information (i.e., energy gain or loss) as fingerprint for authentication, or neglect the fact that the physical distance between devices actually has a substantial effect on the extracted fingerprint, which could threaten the accuracy of these approaches. For these reasons, in this part, we utilize both magnitude and phase information of the frequency responses which are affiliated with the enrolled phone as fingerprints, and further mitigate the impact of transmission distances on the extracted acoustic fingerprints to achieve an accurate and robust device authentication for 2FA.

Specifically, remind that in Section IV-C, we use Equation (3) to (10) to describe the direct sound propagation model

in terms of the magnitude and phase representation, respectively. However, direct deriving the effect of transmission distances from these equations is challenging due to that the transmission distances (i.e.,x_{AA}, x_{BB}, x_{AB} and x_{BA}) between speakers and microphones are unknown to us, whereas they could have an uncertain effect on the magnitude/phase of the received signal. So it is natural for us to think about whether we could design a new scheme to first mitigate such effects in magnitude/phase domain, and then 'distill' robust fingerprint for 2FA.

1) Acoustic Fingerprint Extraction in Magnitude Domain: Remind that from the sound propagation model as described in Equation (3) to (6), the effect of transmission distance in magnitude domain could be well modeled as an exponential trend function with large-scale attenuations [18], [19]. However, magnitude distortions caused by devices' unique electronic features can usually be represented as irregular small-scale fluctuations [12]. Thus, inspired by this observation, in this part, we first propose a new mitigation scheme to eliminate the impact of distance by using the idea of sequence decomposition, and then extract robust magnitude fingerprints for 2FA.

The data sequence can exhibit a variety of patterns, and it is thus often helpful to decompose a sequence into several components, each representing an underlying pattern category [20]. Towards this direction, we assume that a multiplicative decomposition could be conducted on the magnitude sequence of Fourier transform, and they could be represented as the product of two components: the trend component and the non-trend component. The trend component shows the general tendency of the sequence to increase or decrease over a wide range of frequencies, and it's relative fluctuations are also orders of smoother than the non-trend component. Thus, the trend component should actually contain most of the large-scale attenuations caused by the transmission distance [17], and it only includes a small amount of information for 2FA. So it is quite natural for us to think about whether we could decompose the magnitude sequence and remove the trend component from Equation (3) to (6) to mitigate the effect of transmission distance.

Specifically, if we conduct a multiplicative decomposition on the magnitude sequence of frequency transformation/response of the received beep signal, the device's speaker and microphone respectively (e.g., $|R_{l,AB}(f)|$, $|S_A(f)|$ and $|M_B(f)|$), they could be represented as follows:

$$|R_{l,AB}(f)| = \left|R^T_{l,AB}(f)\right| \left|R^N_{l,AB}(f)\right| \tag{11}$$

$$|S_A(f)| = \left|S^T_A(f)\right| \left|S^N_A(f)\right| \tag{12}$$

$$|M_B(f)| = \left|M^T_B(f)\right| \left|M^N_B(f)\right| \tag{13}$$

where $\left|R^T_{l,AB}(f)\right|$ (or $\left|S^T_A(f)\right|$, $\left|M^T_B(f)\right|$) and $\left|R^N_{l,AB}(f)\right|$ (or $\left|S^N_A(f)\right|$, $\left|M^N_B(f)\right|$) represent the corresponding trend and non-trend components, respectively. Put Equation (11) to (13) back to Equation (5), we have:

$$\begin{aligned}|R_{l,AB}(f)| &= \left|R^T_{l,AB}(f)\right| \left|R^N_{l,AB}(f)\right| \\ &= |P_{l,A}(f)| \left|S^T_A(f)\right| \left|S^N_A(f)\right| \left|M^T_B(f)\right| \left|M^N_B(f)\right| e^{\lambda(x_{AB},f)}\end{aligned} \tag{14}$$

Since our system uses a flat stimulation (i.e., $|P_{l,A}(f)|$) as the input to the speaker, and remind again that the effect of device distance could also be described as large-scale attenuations (i.e., an exponential decay function $e^{\lambda(x_{AB},f)}$), both stimulation function and exponential decay function should be included in the trend component. So we can derive the trend and non-trend components using Equation (14) as:

$$\left|R^T_{l,AB}(f)\right| = |P_{l,A}(f)| \left|S^T_A(f)\right| \left|M^T_B(f)\right| e^{\lambda(x_{AB},f)} \tag{15}$$

$$\left|R^N_{l,AB}(f)\right| = \left|S^N_A(f)\right| \left|M^N_B(f)\right| \tag{16}$$

To mitigate the effect of transmission distance, we need to estimate the trend component and eliminate it from Equation (14). Specifically, we adopt a moving average scheme. The moving average $ma(\cdot)$ is commonly used for sequential data to smooth out small-scale fluctuations and extract large-scale trends from the sequence. In this part, we set the size of the average filter as 20 and calculate the un-weighted mean from an equal number of data on either side of each frequency point in $|R_{l,AB}(f)|$. The magnitude sequence after moving average is denoted as $ma(|R_{l,AB}(f)|)$. Thus, the trend components could be estimated as:

$$\left|R^T_{l,AB}(f)\right| \approx ma(|R_{l,AB}(f)|) \tag{17}$$

To eliminate the impact of transmission distance, we divide both sides of Equation (11) by Equation (17):

$$\left|R^N_{l,AB}(f)\right| = \frac{|R_{l,AB}(f)|}{\left|R^T_{l,AB}(f)\right|} \approx \frac{|R_{l,AB}(f)|}{ma(|R_{l,AB}(f)|)} \tag{18}$$

Next, we put Equation (18) to Equation (16), and then repeat this similar decomposition and estimation process for $|R_{l,AA}(f)|$, $|R_{l,BB}(f)|$ and $|R_{l,BA}(f)|$ respectively, we have the following equations:

$$\frac{|R_{l,AA}(f)|}{ma(|R_{l,AA}(f)|)} \approx \left|S^N_A(f)\right| \left|M^N_A(f)\right| \tag{19}$$

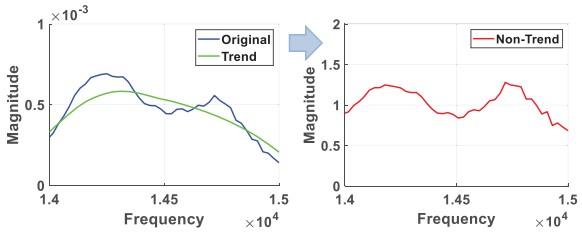
$$\frac{|R_{l,BB}(f)|}{ma(|R_{l,BB}(f)|)} \approx \left|S^N_B(f)\right| \left|M^N_B(f)\right| \tag{20}$$

$$\frac{|R_{l,AB}(f)|}{ma(|R_{l,AB}(f)|)} \approx \left|S^N_A(f)\right| \left|M^N_B(f)\right| \tag{21}$$

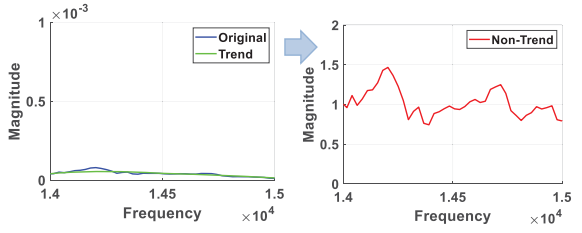
$$\frac{|R_{l,BA}(f)|}{ma(|R_{l,BA}(f)|)} \approx \left|S^N_B(f)\right| \left|M^N_A(f)\right| \tag{22}$$

By solving these equations, the login device can derive the estimated fingerprints $r_1(f,l) = \left|S^N_{l,A}(f)\right|$ and $r_2(f,l) = \left|M^N_{l,A}(f)\right|$ of the enrolled phone by utilizing the data of the l-th received beep signal, and we further use them as magnitude fingerprints for 2FA.

Example. Figure 5 shows an example on how the non-trend components are derived when the transmission distance between two devices (i.e., A and B) is 0.1 m or 3 m, respectively. Specifically, a moving window is slide across the magnitudes of Fourier transform of the received beep



(a) $x_{AB} = 0.1$ m.



(b) $x_{AB} = 3$ m.

Fig. 5. An example of non-trend component estimation using Equation (18) under different transmission distances (i.e., $x_{AB} = 0.1$ m or 3 m): the original $|R_{l,AB}(f)|$, the estimated trend component $\left|R^T_{l,AB}(f)\right| \approx ma(|R_{l,AB}(f)|)$ and the corresponding extracted non-trend component $\left|R^N_{l,AB}(f)\right|$.

signal (i.e., $|R_{l,AB}(f)|$) and the trend component $\left|R^T_{l,AB}(f)\right|$ is calculated using the moving average scheme. The non-trend component is then extracted using Equation (18). From Figure 5 (a) and (b), we can observe that after mitigating the effect of transmission distance, the non-trend components are similar regardless of the transmission distances. This result is encouraging as it confirms the feasibility of using our proposed scheme to mitigate the effect of transmission distance in magnitude domain.

2) Acoustic Fingerprint Extraction in Phase Domain: Remind that the effect of transmission distance (or non-synchronization) in phase domain could be modeled as a linear function (i.e., $f \times \frac{x}{c}$) as illustrated in Equation (7) to (10). In particular, the derivative of this linear function with respect to frequency f (i.e., $\frac{x}{c}$) indicates the delay of the sound propagation for the distance of x, which can also denote as the size of the group delay [17], and such concept can also be naturally extended to other nonlinear functions. Thus, inspired by these observations, we design a new scheme to mitigate the effect of transmission distance in phase representation by utilizing the idea of group delay derivation. Specifically, we first calculate the derivative on both side of Equation (9) with respect to f. They could be represented as follows:

$$\angle R'_{l,AB}(f) = \angle P'_{l,A}(f) + \angle S'_A(f) + \angle M'_B(f) - \frac{x_{AB}}{c} \quad (23)$$

Note that the effect of distance $\frac{x_{AB}}{c}$ becomes a constant value. Thus, it is natural for us to think about whether we could do zero-centering to eliminate this constant value. In particular, we apply the zero-centering function $zc(\cdot)$ on both side of Equation (23) by subtracting the mean value from the data, and further repeat the above process for Equation (7), Equation (8) and Equation (10), we then have the following equations:

$$zc(\angle R'_{l,AA}(f)) = zc(\angle P'_{l,A}(f)) + zc(\angle S'_A(f)) + zc(\angle M'_A(f)) \quad (24)$$

$$zc(\angle R'_{l,BB}(f)) = zc(\angle P'_{l,B}(f)) + zc(\angle S'_B(f)) + zc(\angle M'_B(f)) \quad (25)$$

$$zc(\angle R'_{l,AB}(f)) = zc(\angle P'_{l,A}(f)) + zc(\angle S'_A(f)) + zc(\angle M'_B(f)) \quad (26)$$

$$zc(\angle R'_{l,BA}(f)) = zc(\angle P'_{l,B}(f)) + zc(\angle S'_B(f)) + zc(\angle M'_A(f)) \quad (27)$$

where $zc(x) = x - mean(x)$. Since $\angle P_{l,A}(f)$ and $\angle P_{l,B}(f)$ of the input signal are known, we can solve the above equations to derive the estimated phase fingerprints $r_3(f,l) = zc(\angle S'_{l,A}(f))$ and $r_4(f,l) = zc(\angle M'_{l,A}(f))$ of the enrolled phone by utilizing the data of the l-th received beep signal for 2FA.

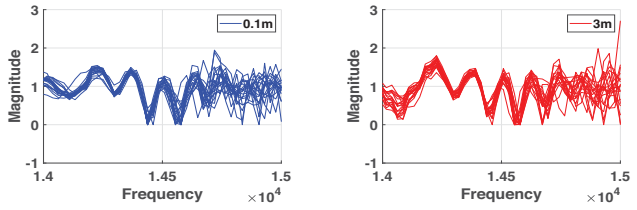
3) Feasibility Study: We provide a feasibility study to show the effectiveness of our acoustic fingerprint extraction scheme. Specifically, we place two smartphones in an empty room, use them as device A and B respectively and collect beep signals for acoustic fingerprint extraction purpose. The magnitude/phase fingerprints for the microphone of device A (i.e., $r_2(f,l)$ and $r_4(f,l)$) under different transmission distances over 20 beep signals are then derived from the received signal and presented in Figure 6.

From Figure 6, irrespective of whether the transmission distance x_{AB} (or x_{BA}) is set as 0.1m or 3m, the fingerprint extracted from both magnitude and phase domain remain similar. These observations strongly confirm the feasibility of using our proposed fingerprint extraction scheme to conduct an accurate 2FA. This result is also consistent with our expectation since it demonstrates that our proposed scheme could mitigate the effect of transmission distance for the acoustic fingerprint extraction.

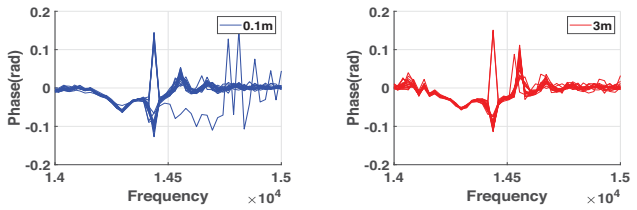
E. Device Authentication

We propose to use the fingerprints $r_1(f,l)$ to $r_4(f,l)$ extracted from the enrolled phone as acoustic fingerprints to conduct the device authentication for 2FA. However, a fingerprint could be decomposed into sub-segments, and only a part of these sub-segments remain invariant across a set of fingerprints generated by the same enrolled device. Such 'stable' sub-segments should be treated more significantly since they could better represent the uniqueness of the device's fingerprint patterns. Thus, to capture this observation in a quantitative way, we propose to use weighted Pearson correlation coefficients when conducting the similarity comparison between the extracted fingerprints and the device profile.

We next illustrate how to calculate the weight for each sub-segment from the derived fingerprints of the enrolled phone. Based on the frequency range (i.e., 1000 Hz band from 14000 Hz to 15000 Hz) used in this work, we first equally divide them into K (e.g., $K = 10$) sub-segments: $\{P_n, ..., P_{n+1}\}, n = 0, ..., K-1$ with $P_0 = 14000$, $P_1 = 14100$, ..., and $P_{10} = 15000$. Thus, for the m-th fingerprint



(a) $r_2(f,l)$ when $x_{AB} = 0.1$ m (b) $r_2(f,l)$ when $x_{AB} = 3$ m



(c) $r_4(f,l)$ when $x_{AB} = 0.1$ m (d) $r_4(f,l)$ when $x_{AB} = 3$ m

Fig. 6. An illustration of feasibility study.

($m \in [1,4]$), the average distance over these blocks can be represented as: $Dist = \{\bar{d}_{n,m}, n = 0, ..., K-1, m = 1, ..., 4\}$, where each $\bar{d}_{n,m}$ is defined as:

$$\bar{d}_{n,m} = \frac{\sum\limits_{\substack{l_1,l_2 \in [1,L] \\ l_1 \neq l_2}} \int_{P_n}^{P_{n+1}} |r_m(f,l_1) - r_m(f,l_2)|\, df}{(L-1) \times L \times (P_{n+1} - P_n)} \tag{28}$$

Each $\bar{d}_{n,m}$ in $Dist$ measures the average sample distance in the n-th sub-segments between each pair of L fingerprints for the m-th fingerprint. Based on the sample distance, we define *weights* over these blocks as $\{w_{n,m}, n = 0, ..., K-1, m = 1, ..., 4\}$, where each $w_{n,m}$ is defined as: $w_{n,m} = 1/\bar{d}_{n,m}$.

We then define the similarity score between the fingerprints obtained from run-time measurement $I_g = \{r_m^g(f,l), m = 1, ..., 4\}$ and the device profile $I_p = \{r_m^p(f), m = 1, ..., 4\}$ by computing weighted Pearson correlation coefficient with the weight $\{w_{n,m}, n = 0, ..., K-1, m = 1, ..., 4\}$ as:

$$\mathbf{C}(I_p, I_g) = \frac{1}{4} \sum_{m=1}^{4} \sum_{n=0}^{K-1} \frac{corr(r_m^g(f,l), r_m^p(f))(u(f-P_n) - u(f-P_{n+1}))w_{n,m}}{\sum\limits_{n=0}^{K-1} w_{n,m}} \tag{29}$$

where $u(f)$ is the unit step function with $u(f) = 1$ when $f \geq 0$ and $u(f) = 0$ when $f < 0$. Thus, if the similarity scores is larger than a pre-defined threshold, our system will pass the authentication process for this particular beep l. To further improve the accuracy, we adopt a plurality vote based criteria, which chooses the most frequently occurring results among consecutive beeps, to determine the final authentication results.

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 different smartphone models including Samsung Galaxy s7, Huawei Mate 10, Huawei Mate 30, and Nova 4 for evaluations. During the experiment, according to Section IV-A, we set the bandwidth of the beep signal as 14 to 15 kHz with a length of 0.02s. The experiments are conducted under three representative environments: the quiet living room (i.e., *Quiet*), the relatively noisy conference room with music on (i.e., *Music*), with the talking noise (i.e., *Talk*). Unless otherwise specified, the results presented in this work are using the acoustic data collected from the quiet living room.

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

Device Authentication: A legitimate user is told to place his/her enrolled phone besides the login device to conduct the normal authentication after inputting his/her own username and password.

Random Impersonation Attack: A adversary obtain the model of the legitimate enrolled phone and impersonates it with his/her own mobile device of the same or different models.

Man-in-the-middle (MiM) Attack: A high-speed channel between the adversarial login device and the victim's enrolled phone is set up. The adversary then relay the probing beep signals between them in attempt to pass our 2FA scheme.

2) Metrics: The following metrics are adopted to evaluate the performance of our proposed 2FA system:

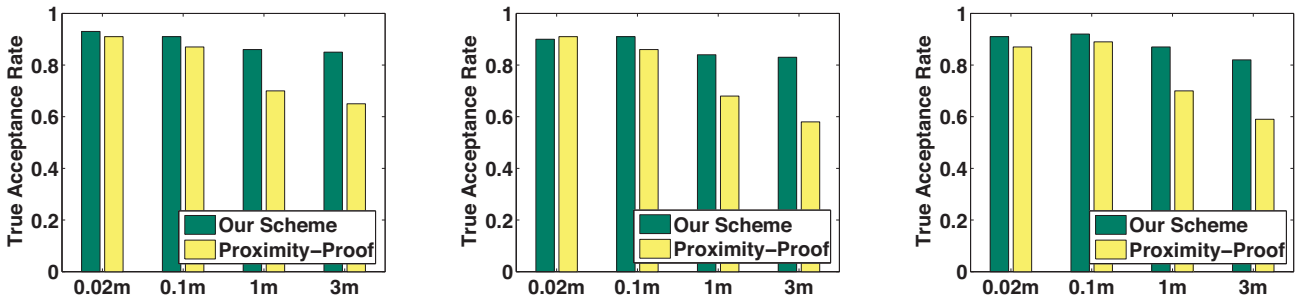
True Acceptance Rate (TAR): the ratio of the number of legitimate login attempts accepted by our system to the total number of legitimate login attempts.

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.

B. Performance Comparison with Existing Schemes

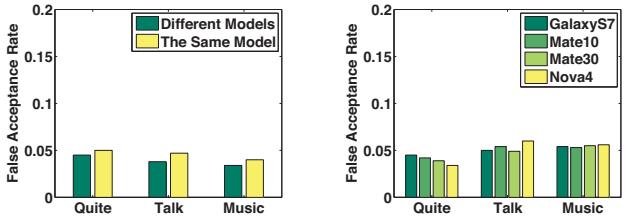
In the first set of experiments, we study the performance of our proposed system for the regular device authentication scenario by comparing it with the state-of-the-art Proximity-Proof system [8] in Figure 7. This scheme also derives the acoustic fingerprints of the microphone and speaker from devices for 2FA. However, this technique only considers the use of magnitude information and the extracted fingerprint also differs with the change of transmission distance. In addition, in this study, experiments are conducted under different environments. The X280 laptop is used as the login device and the Mate 30 is used as the enrolled phone.

From Figure 7(a) to (c), we can observe that the TAR of our proposed system is higher than the Proximity-Proof scheme under most scenarios. In addition, it also remains stable over 0.8 across all scenarios when the transmission distance increases from 0.02m to 3m. However, this metric decreases significantly when the transmission distance increases for the Proximity-Proof. This is because in Proximity-Proof, only the magnitude information is used and the transmission distance



(a) Quite (b) Talk (c) Music

Fig. 7. Performance comparison with the existing scheme under different environments.



(a) Random Impersonation Attack (b) MiM Attack

Fig. 8. Performance study under different attacks.

between devices could also significantly impact the extracted fingerprint for 2FA. However, our proposed system has the capability to mitigate the effect of transmission distance for both magnitude and phase fingerprints. Therefore, the distance cause little disturbance to our proposed system. Overall, these results show that our system is effective in 2FA and also robust to distance changes between devices.

C. Performance Evaluation Under the Random Impersonation Attack

Next, we study how the performance of our system changes under the random impersonation attack. In this study, we use one Mate 30 smartphone as the legitimate enrolled phone and other smartphones as fake enrolled phones to launch the random impersonation attack.

Figure 8(a) presents the FAR of random impersonation attacks under different environments. We observe that the overall FAR remains less than about 0.05 across all scenarios. This shows that our proposed scheme can thwart the random impersonation attack under different environments even if the legitimate and fake enrolled phones come with the same model. Further, we can also observe that better performance can be achieved when the legitimate and fake enrolled phones are with different models. This is because when two devices are with the same model, they still tend to have more similar fingerprint patterns. Overall, these results show that our system is effective in thwarting the random impersonation attack across different environments.

D. Performance Evaluation Under the MiM Attack

Finally, we evaluate our 2FA system under MiM attacks. Specifically, 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 enrolled phone is placed in different experimental environments and the adversary login device is placed in a separate room which is far away from the enrolled phone.

Figure 8(b) presents the FAR of MiM attacks under different environments. We observe that the overall FAR remains less than about 0.06 across all environments. Further, this figure also illustrates that similar performances could be achieved from different scenarios, indicating our 2FA system is robust to MiM attacks under different experimental environments and phone models. This is because the login device would obtain the relay device's fingerprints instead of the enrolled phone's fingerprint under MiM attacks. Such illegitimate fingerprint patterns differ significantly from the real fingerprint and it cannot pass the 2FA. Therefore, the MiM attack can be thwarted effectively.

VI. Conclusion

In this work, we propose a secure 2FA that utilizes the individual acoustic fingerprint of the speaker/microphone on enrolled device as the second proof. The main idea behind our system is to use both magnitude and phase fingerprints derived from the frequency response of the enrolled device for 2FA. Given the input microphone samplings, our system designs an arrival time detection scheme to accurately identify the beginning point of the beep signal by computing the cumulative sum of difference to the ambient noise level. To achieve an accurate authentication, we develop a new distance mitigation scheme to eliminate the impact of transmission distances from the sound propagation model for extracting robust fingerprint in both magnitude and phase domain. Our device authentication component then calculates a weighted correlation value between the device profile and fingerprints extracted from run-time measurements to conduct the device authentication for 2FA. Our experimental results show that our proposed system is accurate and robust to both random impersonation and Man-in-the-middle (MiM) attack across different scenarios and device models.

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, 2022.

[3] "Google 2-step Verification," https://www.google.com/landing/2step/, 2022.
[4] M. Shirvanian, S. Jarecki, N. Saxena, and N. Nathan, "Two-factor authentication resilient to server compromise using mix-bandwidth devices," in *Proceedings of NDSS*, 2014.
[5] 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.
[6] N. Karapanos, C. Marforio, C. Soriente, and S. Čapkun, "Soundproof: Usable two-factor authentication based on ambient sound," in *Proceedings of the 24th USENIX Conference on Security Symposium*, 2015.
[7] 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.
[8] 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.
[9] D. Liu, Q. Wang, M. Zhou, P. Jiang, Q. Li, C. Shen, and C. Wang, "Soundid: Securing mobile two-factor authentication via acoustic signals," *IEEE Transactions on Dependable and Secure Computing*, 2022.
[10] Z. Zhou, W. Diao, X. Liu, and K. Zhang, "Acoustic fingerprinting revisited: Generate stable device id stealthily with inaudible sound," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security(CCS)*, 2014.
[11] A. Das, N. Borisov, and M. Caesar, "Do you hear what i hear? fingerprinting smart devices through embedded acoustic components," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, 2014.
[12] 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.
[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] D. Han, A. Li, L. Zhang, Y. Zhang, J. Li, T. Li, R. Zhang, and Y. Zhang, "(in)secure acoustic mobile authentication," *IEEE Transactions on Mobile Computing*, 2021.
[15] 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.
[16] 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.
[17] S. H. Alan V. Oppenheim, Alan S. Willsky, *Signals and Systems*. Prentice Hall, 1996.
[18] E. J. G. Stordal, "Power-law attenuation of acoustic waves in random stratified viscoelastic media," Master's thesis, University of Oslo, 2011.
[19] L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of Acoustics*, forth ed. John Wiley and Sons, Inc., 2000.
[20] C. Chatfield and H. Xing, *The Analysis of Time Series An Introduction with R*. Chapman and Hall/CRC, 2019.