(In)secure Acoustic Mobile Authentication

Dianqi Han, *Student Member, IEEE,* Ang Li, Tao Li, Lili Zhang, Yan Zhang, Jiawei Li, *Student Member, IEEE,* Rui Zhang, *Member, IEEE,* Yanchao Zhang, *Fellow, IEEE,*

Abstract—Acoustic fingerprinting aims to identify a mobile device based on its internal microphone(s) and speaker(s) which are unique due to manufacturing imperfection. This paper seeks a thorough understanding of the (in)security of exploring acoustic fingerprints for achieving distributed mobile authentication. Our contributions are threefold. First, we present a new acoustic fingerprint-emulation attack and demonstrate that it is a common vulnerability of acoustic mobile authentication systems. Second, we propose a dynamic challenge-response defense to secure acoustic mobile authentication systems against the acoustic fingerprint-emulation attack. Finally, we thoroughly investigate existing acoustic fingerprinting schemes and identify the best option for accurate, secure, and deployable acoustic mobile authentication systems.

Index Terms—Acoustic authentication, fingerprint-emulation attack, dynamic challenge-response

1 Introduction

Acoustic fingerprinting aims to identify a mobile device based on its internal microphone(s) and speaker(s). It is promising for two primary reasons. First, typical smartphones have multiple microphones and at least one speaker, and the latest smartwatches also have a built-in speaker-microphone pair to support phone calls. Second, every microphone or speaker is a multi-stage audio signal processing system consisting of multiple hardware elements, so it can be quite unique due to the hardware imperfection introduced in the manufacturing process.

Different acoustic fingerprints have been explored. The Frequency Response Curve (FRC), which refers to the normalized output gains of a speaker or microphone over a given frequency range, was used in [1], [2], [3], [4]. Das *et al.* used Mel-Frequency Cepstral Coefficients (MFCCs) of the output audio to identify a speaker or microphone [5]. Finally, NAuth [6] distinguishes different speaker-microphone pairs with a nonlinear feature called Acoustic Nonlinear Pattern (ANP). The hardware features of a device's speaker, microphone, or speaker-microphone pair can be used as the acoustic fingerprint, and we term the corresponding fingerprints as S-Print, M-Print, and SM-Print, respectively.

Mobile authentication is one of the most appealing application scenarios of acoustic fingerprinting. A mobile authentication system considers a user as legal if (s)he can prove the possession of a registered mobile device. Fig. 1 shows a generic acoustic mobile authentication system. It consists of three parties: the prover $\mathcal{P}$ (the registered mobile device of the user), the verifier $\mathcal{V}$, and the server $\mathcal{S}$. Without ambiguity, we also denote the user owning the prover device by $\mathcal{P}$. $\mathcal{P}$ starts an authentication instance by sending a request with $\mathcal{P}$'s ID to $\mathcal{S}$. Then $\mathcal{S}$ sends a challenge to $\mathcal{P}$ via $\mathcal{V}$, and $\mathcal{P}$ returns a response corresponding to the challenge. $\mathcal{P}$ is authenticated if $\mathcal{S}$ verifies that the response is associated with one of the registered devices, and vice versa. The mobile device can act as both $\mathcal{P}$ and $\mathcal{V}$ in self-proof scenarios like the online account login on a mobile device, in which case $\mathcal{S}$ directly communicates with the mobile device.

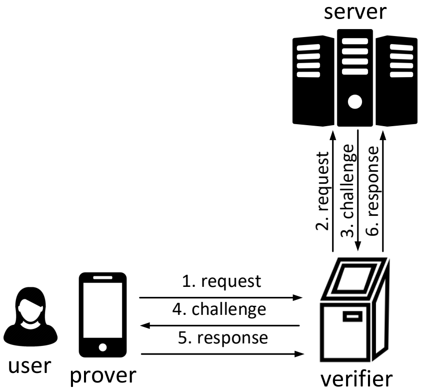


Fig. 1. A generic mobile authentication system.

Acoustic fingerprinting has been applied to mobile authentication in [3], [4], but there are still two open questions: 1) which acoustic fingerprint technique is most suitable for mobile authentication; 2) the fingerprint of which acoustic element(s) (the speaker, microphone, or speaker-microphone pair) should be used. To answer these two questions, we identify the following three essential requirements for a sound acoustic mobile authentication system.

- **Accurate**: the system can accurately identify mobile devices.
- **Deployable**: it is low-cost and can extract *verifier-agnostic* acoustic fingerprints. In particular, the fingerprint of a mobile device should not be tied to a specific verifier, which is very important in a large distributed system with many verifiers such as smart door locks.

D. Han, A. Li, L. Zhang, Y. Zhang, J. Li, and Y. Zhang are with the School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, AZ 85287 (e-mail: dqhan, anglee, lilizhang, yanzhangyz, jwli, yczhang@asu.edu).
T. Li is with the Computer and Information Technology Department, Indiana University-Purdue University Indianapolis, Indianapolis, IN 46202 (tli6@iupui.edu)
R. Zhang is with the Computer and Information Sciences Department, University of Delaware, Newark, DE 19716 (e-mail: ruizhang@udel.edu).

- **Secure**: it is highly resilient to possible attacks.

To extract a verifier-agnostic MFCC fingerprint of a prover, the verifier must be equipped with a high-fidelity speaker or microphone which usually costs a few hundred dollars or more. Since MFCC fingerprints do not satisfy the deployable requirement, we focus on studying FRC and ANP fingerprints henceforth.

We consider a powerful adversary that launches a highly risky fingerprint-emulation attack. In this attack, some acoustic fingerprints of the prover $\mathcal{P}$ are exposed to and can be successfully emulated by the adversary. We experimentally show that all existing acoustic fingerprinting techniques are vulnerable to this attack. We also propose a dynamic acoustic challenge-response scheme as a strong defense based on the motivation that the fingerprint-emulation attack can be successfully defeated as long as the challenges corresponding to exposed fingerprints are not reused. The efficacy of this defense relies on a large number of distinct acoustic fingerprints. So we quantify the fingerprint space for each fingerprinting scheme and find that only the ANP fingerprints of the microphone are highly resilient to the fingerprint-emulation attack.

We summarize the contribution of this paper as follows.

- We present a new fingerprint-emulation attack that is overlooked in previous work and demonstrate that it is a common vulnerability of acoustic mobile authentication systems.
- We propose a dynamic challenge-response defense to secure acoustic mobile authentication systems against the acoustic fingerprint-emulation attack;
- We thoroughly investigate existing acoustic fingerprinting schemes and identify the best option for accurate, secure, and deployable acoustic mobile authentication systems.

The rest of this paper is organized as follows. Section 2 reviews the background knowledge about acoustic elements and acoustic fingerprints. Section 3.1 outlines the system model, the adversary model, and our defense strategy. Section 4 shows the vulnerability of FRC fingerprints and also presents our corresponding defense. Section 5 demonstrates the vulnerability of ANP fingerprints and details our countermeasure as well. Section 6 discusses the trade-off between FRC and ANP fingerprinting schemes. Section 7 briefs the related work. Section 8 concludes this work.

2 Background

2.1 Linear and Nonlinear Behaviors of Acoustic Elements

Microphones and speakers are pervasive on mobile devices. A typical MEMS microphone converts sound into an electrical signal through a sequence of modules, including a pressure-sensitive diaphragm, a pre-amplifier, a low-pass filter, and an analog-to-digital converter (ADC) that produces a digital audio signal. In contrast, a speaker turns an electrical signal into sound waves through a reverse sequence of modules.

Ideal microphones and speakers are expected to be linear and frequency-independent systems. In particular, let S_{in} and S_{out} denote the input and output signals of a microphone or speaker, respectively. A linear microphone or speaker satisfies

$$S_{\text{out}} = g_1 S_{\text{in}}, \tag{1}$$

where g_1 denotes the gain factor, i.e., the amplification or attenuation ratio of the microphone or speaker.

In practice, the gain factor of an acoustic element varies across different frequencies due to hardware imperfection [1]. The Frequency Response Curve (FRC), which represents the magnitudes of the gain factor over a frequency range, formulates the linear characteristic of an acoustic element.

Moreover, practical microphones and speakers on commodity mobile devices are only approximately linear in the audible range due to cost considerations and exhibit nonlinearity in the non-audible range. In particular, we have

$$S_{\text{out}} = \sum_{i=1}^{\infty} g_i S_{\text{in}}^i, \tag{2}$$

where g_i is called the ith-order non-constant nonlinear coefficient. According to [7], $\{g_i | i \geq 1\}$ are sensitive to the frequencies in S_{in}, and $\{g_i | i \geq 2\}$ are also sensitive to the power of individual frequency components in S_{in}. Given a specific S_{in}, g_i is determined by the nonlinear characteristic of the acoustic element.

2.2 Basics of Acoustic Fingerprints

The FRC of each acoustic element is unique due to manufacture imperfection and can thus be used as the linear acoustic fingerprint. To improve accuracy and reduce latency, existing work preselected multiple frequencies and measures the gain factors at the selected frequencies as FRC [1], [3], [4].

Existing work also proposed to identify a microphone-speaker pair with nonlinear features [6]. Due to the nonlinear relation between the input and output signals shown in Eq. 2, the output signal contains new frequency components not present in the input signal [12], [13], and those new frequency components are referred to as distortion components. Given a specific input signal, the amplitudes of distortion components in an acoustic element's output signal are unique due to manufacture imperfection and can be used as nonlinear features to identify this element [6].

3 System Model, Adversary Model, and Overview of Defense

3.1 System Model

A generic acoustic mobile authentication system consists of the prover $\mathcal{P}$, the verifier $\mathcal{V}$, and the server $\mathcal{S}$, as shown in Fig. 1. $\mathcal{P}$ is the user's mobile device and is registered to the system in the initialization stage. The registration can only be conducted when the communication between $\mathcal{P}$ and $\mathcal{S}$ is guaranteed to be secure, i.e., when the attacker can neither overhear nor tamper with the communication. To register a mobile device, the user first sends $\mathcal{S}$ a registration request which contains an identification proof such as the username and password. $\mathcal{S}$ verifies the request and returns a challenge, and the user generates a response corresponding to the challenge with $\mathcal{P}$ and submits the response. $\mathcal{S}$ extracts

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2021.3053282, IEEE Transactions on Mobile Computing

$\mathcal{P}$'s acoustic fingerprint from the response and stores the fingerprint for future verification.

We make the following assumptions for the acoustic mobile authentication system. First, the verifier $\mathcal{V}$ can communicate with the server $\mathcal{S}$ through a secure wireless or wired channel. Second, the prover $\mathcal{P}$ and $\mathcal{V}$ can communicate via a short-range wireless channel (e.g., Bluetooth, WiFi, or acoustic channels) which is not necessarily secure.

The system can identify $\mathcal{P}$ with the acoustic fingerprint of its speaker, microphone, or speaker-microphone pair which is termed as S-Print, M-Print, or SM-Print of $\mathcal{P}$ using different challenges and responses. If S-Print is used, the challenge specifies the input to $\mathcal{P}$'s speaker whose output audio is the response. $\mathcal{V}$ records the audio with a built-in microphone and forwards the response to $\mathcal{S}$. If $\mathcal{P}$ is identified with its M-Print, the challenge is an audio generated by $\mathcal{V}$'s speaker. $\mathcal{P}$ records the challenge audio with the its microphone and submits the recorded audio as the response to $\mathcal{S}$. Finally, SM-Print involves $\mathcal{P}$'s speakers and microphones. The challenge specifies the input to $\mathcal{P}$'s speaker, and $\mathcal{P}$ records the output audio with its microphone as response. In addition, the response in S-Print is an audio, and those in M-Print and SM-Print are audio files. To clarify the difference, we term the response in S-Print as an A-response and that in M-Print or SM-Print as an F-response.

S-Print, M-Print, and SM-Print target different authentication scenarios. S-Print and M-Print are suitable for proximity-based authentication systems in which a standalone verifier is available to verify the proximity of $\mathcal{P}$ to $\mathcal{V}$. The verifier can be a smart lock in an access control system or a login terminal such as a laptop with which the user tries to log into his online account. SM-Print is suitable for self-proof authentication scenarios in which $\mathcal{P}$ directly communicates with $\mathcal{S}$. For example, a user may log into his online account on the mobile device which is also used as his prover. In this case, the challenge audio specified by $\mathcal{S}$ is played by $\mathcal{P}$'s speaker and recorded by $\mathcal{P}$'s microphone.

3.2 Adversary Model

We consider an attacker $\mathcal{A}$ who attempts to be authenticated as $\mathcal{P}$ by the system. We have the following assumptions about $\mathcal{A}$: 1) $\mathcal{A}$ has no access to $\mathcal{P}$ and cannot compromise $\mathcal{P}$, $\mathcal{S}$, or $\mathcal{V}$; 2) $\mathcal{A}$ is aware of the used fingerprinting scheme and has acquired some fingerprint(s) of $\mathcal{P}$; 3) $\mathcal{A}$ can launch the attack with advanced equipment like high-fidelity speakers and microphones. Our work focuses on the (in)security of exploring acoustic fingerprints for mobile authentication, and other security mechanisms such as encryption and biometric-based verification are beyond the scope of this paper. We thus consider the authentication system compromised if $\mathcal{A}$ can bypass the acoustic-fingerprint verification.

$\mathcal{A}$ may obtain $\mathcal{P}$'s acoustic fingerprints through three practical ways. First, $\mathcal{A}$ can overhear the communication between $\mathcal{P}$ and $\mathcal{V}$ if the channel between them is insecure. For example, the prover in an S-Print authentication system transmits the response audio to $\mathcal{V}$ through the insecure acoustic channel, and the audio can be captured by any nearby microphones. $\mathcal{A}$ can thus obtain the response audio by deploying a microphone around $\mathcal{V}$ and then use it to infer $\mathcal{P}$'s acoustic fingerprint. In M-Print and SM-Print systems, $\mathcal{P}$ may communicate with $\mathcal{V}$ through Wi-Fi or Bluetooth, which are more secure than the acoustic channel. However, $\mathcal{A}$ still gets a chance to obtain the communication content by launching some advanced attacks such as the Man-in-the-Middle attack proposed in [9]. Second, $\mathcal{A}$ can deploy a phishing website or application which also adopts acoustic authentication and requires the user to reveal $\mathcal{P}$'s acoustic fingerprints. Finally, $\mathcal{S}$ must store $\mathcal{P}$'s fingerprint for verification which may be exposed to $\mathcal{A}$ due to data leakage.

We consider two existing attacks (random impersonation and replay) proposed by previous work [4] and a new attack (fingerprint-emulation).

Random Impersonation. $\mathcal{A}$ impersonates $\mathcal{P}$ with his own mobile device $\hat{\mathcal{P}}$. $\mathcal{A}$ can obtain the model of $\mathcal{P}$ and uses a device of the same model to launch the attack.

Replay Attack. $\mathcal{A}$ manages to obtain $\mathcal{P}$'s response and replays it to the system. In particular, $\mathcal{A}$ starts an authentication instance with his own device $\hat{\mathcal{P}}$ and sends a request containing $\mathcal{P}$'s ID to $\mathcal{S}$. When being asked for a response, $\mathcal{A}$ submits $\mathcal{P}$'s response to $\mathcal{S}$. An F-response, which is an audio file, can be directly submitted through the short range wireless channel between $\hat{\mathcal{P}}$ and $\mathcal{V}$. To submit an A-response, $\mathcal{A}$ plays the response audio to $\mathcal{V}$ with $\hat{\mathcal{P}}$'s speaker. The Man-in-the-Middle attack proposed in Proximity-Proof [4] is essentially a real-time replay attack, so we do not investigate it individually.

Fingerprint-Emulation Attack.$\mathcal{A}$ manages to obtain one fingerprint of $\mathcal{P}$ and then emulates it with his own mobile device. Specifically, $\mathcal{A}$ first starts an authentication instance with his own mobile device and then submits a forged response corresponding to the challenge to $\mathcal{S}$ for verification. If the target authentication system uses M-Print or SM-Print for verification, $\mathcal{A}$ can submit the forged F-response through the short-range wireless channel between $\mathcal{A}$'s mobile device and $\mathcal{V}$. If the target authentication system adopts S-Print, $\mathcal{A}$ plays the forged A-response with a high-fidelity speaker which does not distort the forged A-response. Since $\mathcal{P}$'s fingerprints corresponding to the same challenge are highly consistent with a subtle variance, the classification methods adopted by acoustic fingerprinting techniques can tolerate this variance [1], [4], [6]. By perfectly emulating one fingerprint of the prover $\mathcal{P}$, $\mathcal{A}$ can be authenticated just as $\mathcal{P}$. We demonstrate more details of this attack in Section 4.2 and Section 5.3.

Another relevant attack is the co-located attack mentioned in [4], in which $\mathcal{P}$ and $\mathcal{A}$ are co-located around $\mathcal{V}$. No acoustic fingerprinting scheme alone can defeat this attack, but the cross-device ranging method used in [8] can be easily incorporated used as an effective defense. So we do not consider this attack in this paper due to space constraints.

3.3 Defense Strategy

The fingerprint-emulation attack is a common vulnerability of acoustic mobile authentication systems, so we propose a dynamic challenge-response mechanism as a defense. A mobile device has multiple acoustic fingerprints corresponding to different challenges. The attack can thus be thwarted if the fingerprint used in each authentication session has never been used before. Specifically, $\mathcal{S}$ stores multiple fingerprints of $\mathcal{P}$ referred to as the fingerprint pool and randomly selects a fingerprint for each authentication instance. Every fingerprint can be used only once. The system can update the fingerprint pool when a secure channel between $\mathcal{P}$ and $\mathcal{S}$ is available. To eliminate the risk of fingerprint exposure, the user must update the the fingerprint pool with fingerprints that are not revealed to any authentication system.

The amount of $\mathcal{P}$'s distinct fingerprints (referred to as the fingerprint space) is the primary concern about the dynamic challenge-response defense. People may use a mobile device for years, and thousands of authentication sessions may be conducted during this period. If the fingerprint space is not big enough, the dynamic challenge-response mechanism cannot be adopted. In this paper, we investigate the physical properties of the microphone and speaker and quantify $\mathcal{P}$'s fingerprint space for each fingerprinting scheme. Then we identify which schemes are more suitable for the dynamic challenge-response mechanism.

4 Frequency Response Curve (FRC)

In this section, we study whether FRC fingerprints are suitable for acoustic mobile authentication. We first outline the authentication process in an FRC authentication system. Then we discuss how to launch the fingerprint-emulation attack against FRC. Next, we analyze the accuracy, deployability, and security of FRC authentication systems. Finally, we analyze the feasibility of using dynamic acoustic challenges and responses to defend the FRC authentication system against the fingerprint-emulation attack.

4.1 Authentication System

FRC is a hardware feature of a speaker or microphone associated with its linear properties. The gain factor of a speaker or microphone is sensitive to frequency, and the FRC represents the magnitudes of the gain factor over a frequency range. Due to manufacture imperfection, the FRC of every acoustic element is unique and can be used to identify the element.

FRC authentication systems based on S-Print, M-Print, and SM-Print identify $\mathcal{P}$ with the FRC of its speaker, microphone, and speaker-microphone pair, respectively. For S-Print and M-Print, the challenge specifies the input signal to the fingerprinted acoustic element, and the system obtains the FRC by measuring the amplitude ratio of the output signal to the input signal at each frequency. For SM-Print, $\mathcal{P}$ generates an audio with its speaker and also records the audio with its microphone. The joint FRC of the speaker-microphone pair can be obtained by measuring the amplitude ratio of the microphone's output signal to the speaker's input signal.

To improve accuracy and reduce latency, authentication systems preselect multiple frequencies and measure the gain factors at the selected frequencies as FRC. The input and output signals of the fingerprinted acoustic element(s) are discrete in the frequency domain and only contain measurements at the selected frequencies. Since manufactures typically do not care about the mobile acoustic element's performance in the inaudible range (>18 kHz), the FRCs of mobile acoustic elements are quite uneven and drastically different in the inaudible range. Authentication systems usually adopt the FRC between 18 kHz and 22 kHz (the cut off frequencies of most mobile acoustic elements) as the fingerprint. For example, Proximity-Proof [4] selects 21 frequencies ranging from 18 kHz to 20 kHz with a step length of 100 Hz and uses the microphone and speaker's gain factors at the selected frequencies as the device's acoustic fingerprint. An FRC fingerprint can be denoted by a vector $\langle \alpha_1, \alpha_2, ..., \alpha_n \rangle$, where each α_i denotes the gain factor at the ith selected frequency.

4.2 Fingerprint-Emulation Attack

The process of this attack has been presented in Section 3.2. So we only explain how to conduct it in FRC authentication systems.

SM-Print. Being aware of the prover $\mathcal{P}$'s SM-Print and the challenge, the attacker $\mathcal{A}$ can obtain the FRC of $\mathcal{P}$'s speaker-microphone pair and the input signal to the speaker. Then $\mathcal{A}$ multiplies the spectrum of the input signal by the FRC of $\mathcal{P}$'s speaker-microphone pair to obtain the spectrum of the response. Specifically, the input signal consists of multiple frequency components, and $\mathcal{A}$ multiplies the amplitude of each frequency component by the corresponding gain factor of $\mathcal{P}$'s speaker-microphone pair. Finally, $\mathcal{A}$ can reproduce the response by applying Inverse Fast Fourier Transform (IFFT) to the obtained spectrum.

M-Print. $\mathcal{A}$ multiplies the spectrum of the challenge audio by the FRC of $\mathcal{P}$'s microphone to obtain the spectrum of the response. For this purpose, $\mathcal{A}$ uses a high-fidelity microphone with flat FRC to capture the challenge audio and obtains the spectrum through Fast Fourier Transform (FFT). The FRC of $\mathcal{P}$'s microphone can be obtained from $\mathcal{P}$'s M-Print. Then $\mathcal{A}$ reproduces the response by applying IFFT to the response's spectrum.

S-Print. $\mathcal{A}$ can obtain the FRC of $\mathcal{P}$'s speaker from $\mathcal{P}$'s S-Print and the spectrum of the speaker's input from the challenge. Then $\mathcal{A}$ multiples the spectrum of the input signal by the FRC of $\mathcal{P}$'s speaker and applies IFFT to the obtained spectrum to recover the response audio. Finally, $\mathcal{A}$ reproduces the A-response by playing the response audio with a high-fidelity speaker whose FRC is flat.

4.3 Pros & Cons of FRC Authentication

Accuracy. The Accuracy of FRC fingerprinting schemes has been thoroughly investigated. The results in [1], [3], [4] show that mobile devices can be identified with the FRCs of their speakers, microphones, or speaker-microphone pairs with accuracy above 98%.

Deployability. The verifier for SM-Print does not affect the response generation, so the system is naturally verifier-agnostic and deployable. The method proposed in Proximity-Proof can be used to extract a verifier-agnostic S-Print and M-Print of the prover $\mathcal{P}$ [4]. This method only requires that $\mathcal{P}$ and the verifier $\mathcal{V}$ both be equipped with COTS microphones and speakers that each costs at most a few dollars. Therefore, S-Print and M-Print system options are both deployable.

Security. Mobile devices can be accurately distinguished based on their acoustic fingerprints, so FRC authentication systems are robust to random impersonation.

M-Print and SM-Print system options are both vulnerable to the replay attack, which is nevertheless ineffective against S-Print. To launch the replay attack against an M-Print or SM-Print system, the attacker $\mathcal{A}$ can directly submit the acquired F-response of $\mathcal{P}$ through the short-range wireless channel between $\mathcal{A}$'s mobile device $\hat{\mathcal{P}}$ and $\mathcal{V}$, and $\hat{\mathcal{P}}$'s acoustic components are not involved in this process. Since the response indeed contains $\mathcal{P}$'s fingerprint, the system identifies $\mathcal{A}$ as $\mathcal{P}$ and is thus compromised. To launch the replay attack against an S-Print system, $\mathcal{A}$ replays the obtained A-response of $\mathcal{P}$ with $\hat{\mathcal{P}}$'s speaker. The FRC of $\hat{\mathcal{P}}$'s speaker distorts the replayed response audio as well as the extracted fingerprint. Consequently, $\mathcal{A}$ fails to reproduce a response containing $\mathcal{P}$'s fingerprint and thus cannot compromise the S-Print system.

As a more severe threat, the fingerprint-emulation attack can compromise S-Print, M-Print, and SM-Print. Specifically, the reproduction of F-response is essentially the inverse process of fingerprint extraction. Therefore, the fingerprint extracted from the reproduced F-response is identical to $\mathcal{P}$'s fingerprint, and thus $\mathcal{A}$ can pass the verification. Reproducing an A-response is more challenging because $\mathcal{A}$'s speaker is involved in the response generation. The flat FRC of the high-fidelity speaker has no impact on the spectrum of response, and thus the fingerprint extracted from the reproduced A-response is identical to $\mathcal{P}$'s S-Print.

We conducted an experiment to show the feasibility of reproducing an A-response. We used 20 devices as targeted prover devices in the experiment, and Table 1 lists the device models. The number in the bracket following each model indicates the quantity of devices of this model. The 20 devices listed in Table 1 were used for all the remaining experiments in this paper. We used a laptop as $\mathcal{V}$ and adopted the flat stimulation specified in [4] as the challenge. We reproduced the response with a Pettersson L400 ultrasound speaker [10]. The Pettersson L400 ultrasound speaker can generate sound waves from 10 kHz to 110 kHz and has a relatively flat FRC between 18 kHz and 22 kHz. We adopted the stimulation specified in [4] as the challenge. The ultrasound speaker was connected to a laptop, and we used Audacity [11] to specify the audio generated by the speaker.

In the experiments, the 20 devices were chosen as $\mathcal{P}$ one by one. For a chosen $\mathcal{P}$, we first used $\mathcal{P}$ to generate the benign A-response. Specifically, the input to $\mathcal{P}$'s speaker was the flat stimulation which contained multiple tones at pre-selected frequencies with the same amplitude [4]. $\mathcal{P}$'s speaker generated the response audio with the maximum sampling frequency (44.1 kHz for iOS devices and 48 kHz for Android devices) for 20 ms, and $\mathcal{V}$'s microphone recorded the audio at a sampling rate of 44.1 kHz. The fingerprint extracted from the benign A-response was used as the fingerprint profile stored in $\mathcal{S}$. Then we inferred the response spectrum and used the ultrasound speaker to forge the response. Since the flat stimulation contains multiple tones at pre-selected frequencies with the same amplitude, the response audio contains tones at the same frequencies with the stimulation. The amplitude of each tone can be obtained by multiplying the amplitude of the stimulation tone by the gain factor of the speaker at the corresponding frequency, which can be obtained from $\mathcal{P}$'s S-Print and is known to $\mathcal{A}$. We used the ultrasound speaker to reproduce the inferred response spectrum. We generated 100 forged responses for each prover device and obtained totally 2,000 forged response from which we extracted 2,000 forged fingerprint samples. We calculated the Euclidean distances between the 2,000 fingerprint samples and the corresponding fingerprint profile. Only two distances were larger than 0.2, which is the threshold adopted by Proximity-Proof to distinguish different devices. In other words, only two forged responses were correctly identified as illegal. The high-fidelity ultrasound speaker has a flat FRC in the frequency range used by acoustic fingerprints, and thus it can reproduce the response spectrum and emulate the FRC of $\mathcal{P}$'s speaker with only a subtle distortion. Therefore, the fingerprint-emulation attack can achieve a success rate as high as 99.8%.

TABLE 1
Mobile devices in experiments.

Android devices (12)	Nexus 5 (2), Nexus 7 (2), Google Pixel 2 (2), Google Pixel 3 (2), Samsung S5 (2), and Samsung S7 (2)
iOS devices (8)	iPhone 5 (1), iPhone 5s (1), iPhone 6 (3), iPhone XR (1), iPad 2 (2), and iPad 4 (1)

4.4 Dynamic Challenge-Response

A dynamic challenge-response mechanism outlined in Section 3.3 can enhance the security of FRC authentication systems. Both the replay and fingerprint-emulation attacks can be defeated because the fingerprints that may have been exposed to the attacker $\mathcal{A}$ are never reused.

The primary concern about the dynamic challenge-response defense is the amount of distinct acoustic fingerprints of the prover $\mathcal{P}$. Two distinct FRC fingerprints should not contain any common gain factor to be distinguishable. In what follows, we first demonstrate how we quantify distinct S-Prints of $\mathcal{P}$ for FRC systems and then provide the fingerprint spaces of M-Print and SM-Print obtained through similar processes.

Since each S-Print contains the speaker's gain factors at multiple frequencies, we first investigate the minimal number $\mathcal{K}$ of gain factors needed for accurate device identification through experiments. The results in Proximity-Proof show that a mobile device can be accurately identified with its speaker's gain factors at 21 frequencies ranging from 18 kHz to 20 kHz [4]. In our experiments, we selected the first m gain factors as the S-Print and calculated the corresponding accuracy for different values of m. For each m, we extracted the S-Prints of each device in Table 1 for 20 times and obtained 400 fingerprint samples. We then used the method proposed in Proximity-Proof [4] to identify the device associated with each fingerprint sample and calculated the accuracy. We tested 20 values from 2 to 21 for m. The accuracy increases with m. When m is larger than 10, mobile devices can be identified with accuracy above 95%, and the benefit of further increasing m is insignificant when m exceeds 10. We therefore chose $\mathcal{K}$ to be 10.

Next, we investigate the number of distinct gain factors of a speaker. We assume that a fingerprint $\langle \alpha_1, \alpha_2, \ldots, \alpha_{10} \rangle$ is chosen by the system. Here, α_i is the ith gain factor contained in the fingerprint, and we denote the frequency corresponding to α_i by χ_i. Under the dynamic challenge-response mechanism, $\mathcal{A}$ cannot obtain any α_i. However, $\mathcal{A}$ may have obtained $\hat{\alpha_i}$ whose corresponding frequency $\hat{\chi_i}$ is close to χ_i and then use $\hat{\alpha_i}$ as α_i to launch the fingerprint-emulation attack. The difference between $\hat{\chi_i}$ and χ_i is denoted by $\Delta\chi_i$. Without loss of generality, we assume that $\Delta\chi_1 = \Delta\chi_2 = \cdots = \Delta\chi_n = \Delta\chi$.

The gain-factor variance of an acoustic component within a small frequency range is insignificant even in the high frequency domain [14]. If $\Delta\chi$ is not sufficiently large, the two fingerprints $\langle \alpha_1, \alpha_2, \ldots, \alpha_{10} \rangle$ and $\langle \hat{\alpha_1}, \hat{\alpha_2}, \ldots, \hat{\alpha_{10}} \rangle$ are very likely to be indistinguishable, and thus $\mathcal{A}$ is identified as $\mathcal{P}$ and authenticated. We conducted an experiment to obtain the minimal $\Delta\chi$ to defeat the attack. We tested 10 values ranging from 10 Hz to 100 Hz with a step length of 10 Hz and measured the success rate of the attack for each value of $\Delta\chi$. More specifically, the 20 devices were chosen as $\mathcal{P}$ one by one. For a chosen $\mathcal{P}$, we randomly selected 10 frequencies $(\chi_1, ..., \chi_{10})$ from the 21 frequencies used in Proximity-Proof. The gain factors of $\mathcal{P}$'s speaker on the selected frequencies were extracted as the S-Print $\mathcal{F}$. We then extracted the speaker's fingerprint $\hat{\mathcal{F}}$ on frequencies $\langle \chi_1 + \Delta\chi, ..., \chi_{10} + \Delta\chi \rangle$. The experiment was repeated 10 times for each device, and we calculated the ratio that $\hat{\mathcal{F}}$ is not distinguishable from $\mathcal{F}$ (i.e., the success rate of the attack) for each $\Delta\chi$. Fig. 2 shows our experiment results. The success rate of the attack decreases with the increase of $\Delta\chi$, and the fingerprint-emulation attack can be defeated (i.e., the success rate below 5%) when $\Delta\chi$ is larger than 60 Hz. We meet the requirement for $\Delta\chi$ by choosing the fingerprint frequencies from a set of predetermined values with sufficient gaps. In particular, 66 frequencies ranging from 18 kHz to 21.96 kHz with a step length of 60 Hz are chosen as the candidate frequencies. For each authentication attempt, the system randomly selected 10 frequencies from the candidate frequencies. Since each candidate frequency can be chosen only once, the speaker has $\lfloor 66/10 \rfloor = 6$ distinct FRC fingerprints. This fingerprint space is obviously too small for mobile authentication.

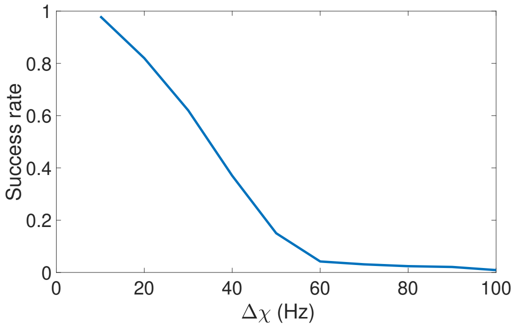


Fig. 2. Success rates of the fingerprint-emulation attack.

We conducted similar experiments and derived the fingerprint space of M-Print and SM-Print as 5 and 6, respectively. Therefore, FRC authentication systems based on M-Print, S-Print, and SM-Print are all still vulnerable to the fingerprint-emulation attack due to the very small fingerprint space.

4.5 Summary of FRC System Performance

We summarize the pros and cons of FRC authentication systems in Table 2. The accuracy is evaluated based on the ratio of acoustic fingerprints whose corresponding devices are correctly identified. The deployability is evaluated based on the hardware with which $\mathcal{V}$ must be equipped. The security is evaluated based on whether the system is resilient to specific attacks (indexed as 'yes' or 'no') and whether the dynamic challenge-response (shortened as dynamic C-R in the table) is adoptable.

TABLE 2
Pros & cons of FRC authentication systems.

Fingerprint scheme		S-Print	M-Print	SM-Print
Accuracy		99.5% [4]	99.5% [4]	98% [3]
Deployability		COTS microphones and speakers [4]		none [3]
Security	Impersonation	yes		
	Replay attack	yes	no	no
	Fingerprint-emulation	no		
	Dynamic C-R	not adoptable		

5 Acoustic Nonlinear Pattern (ANP)

ANP is a hardware feature of an acoustic element related to its nonlinear properties. Due to the nonlinear relation between the input and output signals shown in Eq. 2, the output signal contains new frequency components not present in the input signal [12], [13], and those new frequency components are referred to as distortion components. Our subsequent discussion refers to ANP as the amplitudes of distortion components produced by the nonlinearity of the speaker, microphone, or both. Different speakers or microphones have distinct ANPs for the same input signals.

In the device-to-device authentication system NAuth [6], ANP is used to distinguish different speaker-microphone pairs. NAuth is quite effective in the targeted application scenarios but is not verifier-agnostic. Here we extend NAuth [6] by introducing two ANP authentication systems, M-ANP and SM-ANP, which identify $\mathcal{P}$ using the ANP fingerprints of its microphone and speaker-microphone pair, respectively. We do not consider identifying $\mathcal{P}$ with its speaker's ANP fingerprint because it is difficult to extract verifier-agnostic ANP fingerprints of a speaker. In particular, in order to extract a speaker's fingerprint, a microphone must be used to capture the speaker's output audio. Most microphones, including high-quality ones, exhibit significant nonlinearity in the high frequency domain. The high-frequency audio that can invoke the speaker's nonlinear distortion can also cause distortion at the microphone. Therefore the distortion components in the recorded audio is affected by the microphone and cannot be used to fingerprint the speaker alone.

In this section, we first illustrate M-ANP and SM-ANP and then present a tailored fingerprint-emulation attack. After evaluating M-ANP and SM-ANP, we study whether the dynamic challenge-response defense can be adopted by those two systems.

5.1 M-ANP

M-ANP is an M-Print authentication system, and the system model has been demonstrated in Section 3. Here, we only discuss how to extract $\mathcal{P}$'s ANP fingerprint.

5.1.1 Challenge audio

M-ANP uses a high-frequency audio with two tones as the challenge audio played by verifier $\mathcal{V}$ to prover $\mathcal{P}$. In particular, the challenge audio S_{in} is generated as

$$S_{\text{in}} = A_1 \cos(2\pi f_1 t) + A_2 \cos(2\pi f_2) \ . \tag{3}$$

The nonlinearity of the microphone in COTS smartphones and smartwatches is more significant in the high frequency range above 18 kHz [13]. So we require $f_2 > f_1 \geq 18$ kHz. Since the nonlinear coefficient g_i in Eq. (2) of a common microphone is negligible for $i \geq 3$ [12], the nonlinear output of $\mathcal{A}$'s microphone before low-pass filtering can be approximated by

$$\begin{aligned} S_{\text{out}} &\approx g_1 S_{\text{in}} + g_2 S_{\text{in}}^2 \\ &= \frac{g_2}{2}(A_1^2 + A_2^2) + g_1 A_1 \cos(2\pi f_1 t) + g_1 A_2 \cos(2\pi f_2) \\ &\quad + \frac{g_2 A_1^2}{2} \cos(4\pi f_1 t) + \frac{g_2 A_2^2}{2} \cos(4\pi f_2 t) \\ &\quad + g_2 A_1 A_2 \Big(\cos(2\pi(f_2 + f_1)t) + \cos(2\pi(f_2 - f_1)t) \Big) \ . \end{aligned} \tag{4}$$

Since a typical microphone's cutoff frequency is 22 kHz, the frequency components at $2f_1$, $2f_2$, and $f_2 + f_1$ in S_{out} cannot be recorded. We additionally require $f_2 - f_1 < 18$ kHz so that the distortion component $g_2 A_1 A_2 \cos(2\pi(f_2 - f_1)t)$ can not only be recorded but also be differentiated from the two tones at f_1 and f_2, respectively. As we will see shortly, this distortion component is used to construct the ANP fingerprint of $\mathcal{P}$.

5.1.2 Challenge audio generation

Verifier $\mathcal{V}$ cannot use an ordinary speaker to generate S_{in}. In particular, different COTS speakers exhibit distinct and significant nonlinearity in the high-frequency range above 18 kHz. So S_{in} would invoke the speaker's nonlinear distortion that would further result in many low-frequency distortion components in its output. Such unwanted distortion components can be recorded and mixed with those induced by $\mathcal{P}$'s microphone. The fingerprint extracted from the recorded audio would thus be tied to both $\mathcal{P}$'s microphone and $\mathcal{V}$'s speaker, which violates the verifier-agnostic requirement.

We propose a cost-effective solution based on COTS ultrasound transducers which each costs at most several US dollars. In particular, we let each verifier use two ultrasound transducers with each generating a unique tone in S_{in}. Although ultrasound transducers also exhibit nonlinearity, the resulting distortion components are in the high-frequency range above 22 kHz and thus cannot be recorded by $\mathcal{P}$'s microphone. To see this more clearly, consider an arbitrary transducer $i \in [1, 2]$. The input to transducer i is an electrical signal $A'_i \cos(2\pi f_i t)$, and the corresponding nonlinear output can be modeled as

$$\begin{aligned} T_i &\approx g_{1,i} A'_i \cos(2\pi f_i t) + g_{2,i}(A'_i \cos(2\pi f_i t))^2 \\ &= g_{1,i} A'_i \cos(2\pi f_i t) + \frac{g_{2,i} A'^2_i}{2}(1 + \cos(4\pi f_i t)) \ , \end{aligned} \tag{5}$$

where $g_{1,i}$ and $g_{2,i}$ denote the first-order and second-order coefficients of transducer i, respectively. Since we require that $f_i \geq 18$ kHz, the distortion component at $2f_i$ cannot be recorded by $\mathcal{P}$'s microphone. In addition, the DC component can be easily filtered from the audio recording.

We further use a simple calibration to extract transducer-agnostic and thus verifier-agnostic fingerprints. In particular, each $g_{1,i}$ corresponds to the gain of transducer i which is a standard parameter in the technical specification of the transducer. Since different transducers may have distinct gain factors, we set $A'_i = A_i / g_{1,i}$. Therefore, the effective output from transducer i with regard to $\mathcal{A}$'s microphone is $g_{1,i} A'_i \cos(2\pi f_i t) = A_i \cos(2\pi f_i t)$, which is exactly the challenge tone T_i we need in Eq. (3).

5.1.3 Fingerprint extraction and matching

The absolute amplitude of the distortion component at frequency $f_2 - f_1$ cannot be directly used as $\mathcal{P}$'s fingerprint due to the Automatic Gain Control (AGC) system in common microphones. Specifically, the system automatically adjusts the microphone gain according to the perceived sound volume. So the measured amplitude at frequency $f_2 - f_1$ may vary considerably for different verifiers and/or verifier-$\mathcal{A}$ distances instead of equaling the ideal constant $g_2 A_1 A_2$.

Since the AGC system affects all the frequency components almost equally [15], we propose to use the relative amplitude as $\mathcal{P}$'s fingerprint. For this purpose, we add a reference tone $A_0 \cos(2\pi f_0 t)$ to S_{in}, which is played by an additional transducer at the verifier. Here, A_0 and f_0 are both system constants. We require f_0 much below 18 kHz and also any possible $f_2 - f_1$ so that $A_0 \cos(2\pi f_0 t)$ incurs negligible nonlinear distortion at the microphone. Then we define the **fingerprint element** as the absolute amplitude of frequency $f_2 - f_1$ divided by that of frequency f_0.

M-Print uses $\kappa \geq 1$ different challenge audios that differ in frequencies and/or amplitudes in each authentication session, leading to κ fingerprint elements. $\mathcal{P}$'s fingerprint is extracted as $\boldsymbol{\Theta}_{\text{M}} = \langle \boldsymbol{\theta_1, ..., \theta_\kappa} \rangle$, where θ_i denotes the fingerprint element corresponding to the ith challenge audio. The larger κ, the longer the authentication time, the higher distinguishable Θ_{M}, the more reliable the authentication result, and vice versa.

We also need to mitigate the impact of ambient noise to extract Θ_{M}. For this purpose, we let the verifier play each challenge audio for a duration of ω and then keep silent for ω. Meanwhile, $\mathcal{P}$'s microphone kept recording with a sampling frequency of 44.1 kHz. The ambient noise can be considered constant during this short duration (e.g., $\omega = 50$ ms in our experiment). After applying fast Fourier transform to the audio captured by $\mathcal{P}$'s microphone, we subtracted the noise spectrum in the silent period from the audio spectrum in the non-silent period. The resulting differential spectrum was used to extract the "noise-free" fingerprint for this challenge audio. This process was repeated multiple times, and the average result was used as Θ_{M} for final verification by the authentication server.

We use the scaled Euclidean distance to compare two fingerprints to avoid the dominance of large-valued elements. In particular, assume that the authentication server stores an authentic fingerprint Θ'_{M} for the κ challenge audios. It compares Θ'_{M} with the extracted Θ_{M} by computing

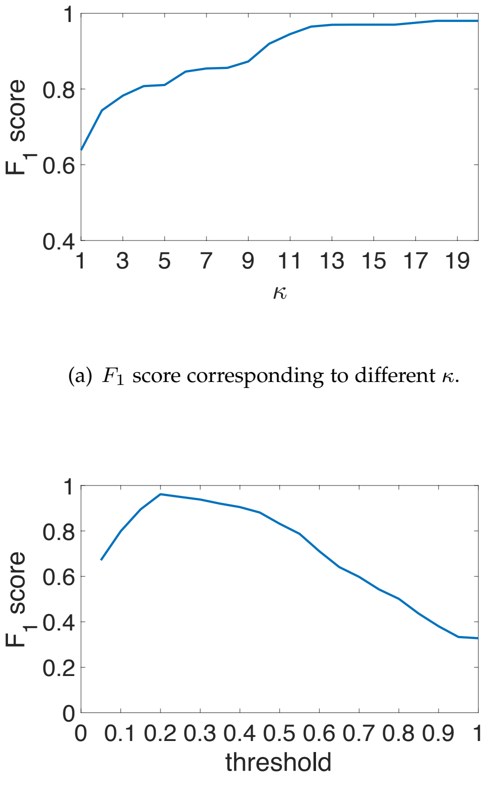
$$\mathsf{diff}(\Theta_{\mathrm{M}}, \Theta'_{\mathrm{M}}) = \sqrt{\sum_{l=1}^{\kappa} \left(\frac{\theta_l - \theta'_l}{\theta_l + \theta'_l} \right)^2 } . \tag{6}$$

If $\mathsf{diff}(\Theta_{\mathrm{M}}, \Theta'_{\mathrm{M}})$ is no larger than a system threshold τ_{M}, the authentication server considers the responses from $\mathcal{P}$ and authenticate the request. κ and τ_{M} are obtained through experiments. We tested 20 candidate values ranging from 1 to 20 for κ. For each value, we obtained the corresponding τ_{M} and calculated the identification accuracy.

We use the F_1 score to obtain τ_{M} corresponding to a specific κ. Two Prowave 250ST160 transducers [16] were used to generate the challenge, and two Agilent 33220A signal generators [17] were used to power the transducers. We chose 10 challenges that each contains κ challenge audios. The frequencies of the challenge tones were randomly selected, and the output voltages of the signal generators were fixed as 10 V. For each of the 20 devices, we extracted its fingerprints corresponding to each of those 10 challenges for 20 times and totally got 4,000 testing fingerprint samples. The distance between transducers and the device's microphone, denoted by $\boldsymbol{d}$, may also has impacts on the amplitudes of distortion components and thus affects the ANP fingerprint. We randomly chose a value between 10 cm and 25 cm as $\boldsymbol{d}$ in each experiment so that the obtained κ and τ_{M} are robust to slight changes of $\boldsymbol{d}$. Then the 20 devices were chosen as $\mathcal{P}$ one by one. When a devices was chosen as $\mathcal{P}$, the rest 19 devices were considered unauthenticated, and we extracted $\mathcal{P}$'s fingerprints corresponding to the 10 selected challenges one more time as the reference fingerprints for later classification. Then we tried 20 values ranging from 0.05 to 1 with a step of 0.05 as τ_{M} to identify whether each testing sample comes from the prover. Based on the classification result, we calculated the F_1 scores corresponding to each τ_{M} as follows:

$$\begin{aligned} F_1 \text{ score} &= \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}} \\ \text{Precision} &= \frac{TP}{TP+FP} \\ \text{Recall} &= \frac{TP}{TP+FN} . \end{aligned} \tag{7}$$

Here, TP denotes the number of fingerprint samples correctly recognized as being the fingerprints of $\mathcal{P}$; FP and FN denote the number of fingerprint samples incorrectly recognized as being and not being the fingerprints of $\mathcal{P}$, respectively. The F_1 score is an important metric to evaluate the accuracy of the binary classification method. A high Precision value guarantees that malicious devices are detected with a high possibility, and a high Recall value guarantees that the legal device is misidentified as illegal with a low possibility. A high F_1 score ensures that both Precision and Recall are high. For one evaluated value of τ_{M}, we obtained 20 F_1 scores with different devices chosen as $\mathcal{P}$. We chose the evaluated value with the highest average F_1 score as τ_{M}.



(a) F_1 score corresponding to different κ.

(b) F_1 score corresponding to different threshold (κ = 12).

Fig. 3. F_1 score corresponding to different κ and threshold.

The maximum average F_1 scores corresponding to different κ are shown in Fig. 3(a). With the increase of κ, the maximum average F_1 score increases. The system can achieve an average F_1 score of 0.96 with an average Precision of 97.6% and an average Recall of 95.2% when κ is 12. However, the benefit of increasing κ is insignificant when κ is larger than 12. In particular, the system achieve average F_1 scores of 0.964 and 0.965 with κ equaling 13 and 14, respectively. Compared with the performance adopting $\kappa = 12$, the average F_1 score increases by less than 0.5%, while the challenge audio length increases by more than 8%. To avoid unnecessary time consumption, M-ANP adopts $\kappa = 12$. Fig. 3(b) shows the average F_1 scores corresponding to different thresholds when κ is 12. A threshold of 0.2 achieves the highest average F_1 score, so we adopt 0.2 as τ_{M}. With $\kappa = 12$ and $\tau_{\mathrm{M}} = 0.2$, we can identify the devices associated with the 4,000 fingerprint samples with accuracy of 96.4%.

5.1.4 Overall performances of M-ANP

We further evaluated M-ANP in three common scenarios with different noise volumes: the office, the store, and the restaurant. Due to hardware constraints, we can only conduct the experiment in the lab. So we recorded the ambient noise in those scenarios and played the recording when we conducted the experiments to emulate those scenarios.

TABLE 3
The overall performance of M-ANP.

M-Print	office	store	restaurant
Precision	97.2%	94.7%	94.1%
Recall	95.1%	93.7%	92.5 %
F_1 score	0.96	0.94	0.93

We used the 20 mobile devices listed in 1 to evaluate the performance of M-ANP. We randomly selected 10 challenges and extracted every device's fingerprints corresponding to each challenges for 80 times (20 times in each scenario). For each scenario, we had obtained a testing set that contained 400 fingerprint samples. The 20 mobile devices were selected as the prover one by one, and the rest 19 devices were considered unauthentic. We extracted the chosen devices fingerprints corresponding to those 10 challenges without playing the recorded noise and used the extracted fingerprints as the references. Then we distinguished whether the fingerprints in the testing sets are associated to the prover or not and calculated the Precision and Recall. After all the 20 devices had been chosen as the prover, we calculated the averaged Precision and Recall for each scenario, and the results are shown in Table 3. M-ANP performs best in the office scenario. The performances in noisy scenarios, such as the store and restaurant scenarios, are comparable to that in the office.

5.2 SM-ANP

5.2.1 Challenge audio

SM-ANP adopts the AM modulated signal used in NAuth [6] as the input to $\mathcal{P}$'s speaker. The challenge signal is obtained by modulating a baseband signal of frequency f_b upon a carrier signal of frequency f_c and is represented by

$$S_{\text{in}} = A_{f_c}\sin(2\pi f_c t)(1 + A_{f_b}\sin(2\pi f_b t)) \ . \tag{8}$$

The challenge in SM-ANP specifies f_c, f_b, A_{f_c}, and A_{f_b}. We invoke the speaker on $\mathcal{P}$ through a standard API which takes a discrete sequence of amplitude values sampled from an sound wave as input and outputs the corresponding audio. For this purpose, we sample S_{in} at a common speaker's maximum sample rate $f_s = 48$ kHz to obtain the following sequence as the input to the speaker API:

$$\hat{S}_{\text{in}}[i] = A_{f_c}\sin(2\pi f_c i/f_s)(1 + A_{f_b}\sin(2\pi f_b i/f_s)) \ , \tag{9}$$

where $\hat{S}_{\text{in}}[i]$ denotes the ith element for all $i = 1, 2, \dots$. S_{in} can invoke significant nonlinear distortion of the speaker-microphone pair on $\mathcal{P}$, which results in many distortion components in the output of $\mathcal{P}$'s microphone with a cutoff frequency of 22 kHz. According to Eq. (2) and trigonometric expansion, these distortion components are at frequencies nf_c, mf_b, and $nf_c \pm mf_b$, where $k, m, n \in \mathbb{N}$ and $nf_c, mf_b, nf_c \pm mf_b <$ 22 kHz [6].

We carefully select f_c and f_b to enhance the nonlinear distortion components in the microphone's output. In particular, we find that many distortion frequencies are actually the same for different combinations of k, m, and n. For example, if $f_c = 20$ kHz and $f_b = 5$ kHz, we have $2f_b = f_c - 2f_b = 10$ kHz. Based on this observation, we can stack up the distortion components so that their combined effect is more profound. For this purpose, we set $f_c = nf_b$ ($n \in \mathbb{N}+$), resulting in $\lfloor 22\text{ kHz}/f_b \rfloor$ distortion components at frequencies $\{kf_b | 1 \leq k \leq \lfloor 22\text{ kHz}/f_b \rfloor\}$).

The next issue is to decide the feasible values for f_b and f_c. Assume that A_{f_c} and A_{f_b} are fixed for the time being. The microphone can record $\lfloor 22\text{ kHz}/f_b \rfloor$ distortion components, each corresponding to one fingerprint element of prover $\mathcal{P}$. If f_b is set too large, very few fingerprint elements can be obtained and thus may be insufficient to distinguish a large number of devices. On the other hand, if f_b is set too small, there may be too many distortion components whose amplitudes may be too small given the fixed total power of the recorded audio, and such weak distortion components may be indistinguishable from noise and thus would dramatically decrease the identification accuracy. So we select f_b from a range $[f_{b,\min}, f_{b,\max})$. In addition, the nonlinearity of speakers and microphones is more significant in the inaudible domain, and the cutoff frequency of the speaker in common mobile devices is 24 kHz. So we set 18 kHz $\leq f_c <$ 24 kHz with f_c being a multiple of f_b.

In SM-ANP, prover $\mathcal{P}$ generates the audio with the highest possible volume to maximize the nonlinear distortion. In particular, the strength of nonlinear distortions is significantly affected by the modulation depth defined as $\zeta = A_{f_b}/A_{f_c}$ [6], [13], which should neither be too large nor too small. SM-ANP selects ζ from a predetermined range $[\zeta_{\min}, \zeta_{\max})$, which can be empirically determined as well. Once ζ is chosen, we maximize A_{f_c} and A_{f_b} under the constraint that the maximum value of sample $\hat{S}_{\text{in}}[i]$ does not exceed the default peak value defined by the operating system (e.g., 32767 in Android).

5.2.2 Fingerprint extraction and matching

Let θ'_i denote the amplitude of the ith distortion component at frequency if_b for all $1 \leq i \leq \beta$. We define the fingerprint of $\mathcal{P}$ for specific f_b, f_c, and ζ as $\Theta_{\text{SM}} = \langle \theta_1, \dots, \theta_\beta \rangle$, where

$$\theta_i = \frac{\theta'_i}{\sqrt{\sum_{j=1}^{\beta} {\theta'_j}^2}} \ . \tag{10}$$

We use the normalized amplitudes instead of absolute values to counteract the impact of the AGC system. We also adopt the method demonstrated in Section 5.1.3 to mitigate the impact of ambient noise.

SM-ANP also uses the scaled Euclidean distance as in Eq. (6) to measure the fingerprint similarity. If the calculated distance does not exceed a system threshold τ_{SM}, the authentication server considers prover $\mathcal{P}$ and thus authentic, and vice versa.

5.2.3 Parameters for SM-ANP

We now explain how to obtain the pre-determined parameters of SM-ANP.

$\boldsymbol{f_b^{\min}}$**,** $\boldsymbol{f_b^{\max}}$**, and** $\boldsymbol{\tau_{\text{SM}}}$**.** The 12 android devices listed in Table 1 were used in the experiment. We fixed ζ as 100% and tried 30 frequencies ranging from 100 Hz to 3 kHz with a step length of 100 Hz as f_b. For each frequency f_b, we set $f_c = \lceil \frac{18\text{ kHz}}{f_b} \rceil f_b$. We obtained totally 30 challenges and used

TABLE 4
The overall performance of SM-ANP.

SM-Print	office	store	restaurant
Precision	96.3%	92.6%	93.1%
Recall	95.2%	92.1%	91.7 %
F_1 score	0.96	0.92	0.92

the F_1 $mesurement$ as in M-ANP to obtain the corresponding τ_{SM} and average F_1 $score$. The average F_1 $score$ is above 0.95 when f_b is between 700 Hz and 2 kHz and dramatically lower when f_b is out of this range. We therefore chose 700 Hz and 2 kHz as $f_b^{\min}$ and $f_b^{\max}$, respectively. Among the 16 frequencies within $[f_b^{\min},\ f_b^{\max}]$, we found that 0.15 is the optimal threshold for 10 of them, and the average F_1 $score$ under this threshold is above 0.95 for the rest 6 frequencies as well. Based on this finding, we chose $\tau_{\text{SM}} = 0.15$. With $\tau_{\text{SM}} = 0.15$, we can identify the devices associated with the collected fingerprint with an accuracy of 95.3%.

$\zeta_{\min}$ and $\zeta_{\max}$. The same 12 android devices were used in this experiment. The impact of ζ is more significant when the amplitudes of the distortion components are small. So we fixed f_b and f_c as 700 Hz and 18.2 kHz, respectively. We increased ζ from 50% to 150 % with a step length of 5% and obtained the corresponding averaged F_1 $score$ for each value. We found that the averaged F_1 $score$ is above 0.95 when M_d is between 75% and 110% and therefore chose $\zeta_{\min} = 75\%$ and $\zeta_{\max} = 110\%$.

5.2.4 Overall performances of SM-ANP

We further evaluated SM-ANP in office, store, and restaurant scenarios with the 12 android devices. We randomly selected 10 challenges and used the same way as we did with M-ANP to obtain the Precision and Recall of SM-ANP in three scenarios. To avoid redundancy, we omit the description of the experiment details and only show the results in Table 4. Similarly to M-ANP, SM-ANP performs best in the office and comparably well in the store and restaurant. Since the speaker's power is weaker compared with the transducer's, the dynamic noise's impact to SM-ANP is more significant.

5.3 Fingerprint-Emulation Attack

In S-ANP and SM-ANP, prover $\mathcal{P}$'s fingerprint partially reveals the spectrum of the response. In S-ANP, the fingerprint reveals the amplitude ratio θ of the distortion component to the reference tone. Attacker $\mathcal{A}$ can forge a response by setting the amplitude of the distortion component as θ multiplied by the reference tone's amplitude. In SM-ANP, the fingerprint reveals the normalized amplitudes of distortion components. Attacker $\mathcal{A}$ can forge the response by using each fingerprint element as the amplitude of the corresponding distortion component. S-ANP and SM-ANP both use F-response, so $\mathcal{A}$ can directly submit the forged response to the system through the short range wireless channel between $\mathcal{A}$ and $\mathcal{V}$.

The generation of the forged response is essentially the inverse process of fingerprint extraction, so the fingerprint extracted from the forged response is identical to $\mathcal{P}$'s fingerprint. The system identifies $\mathcal{A}$ as $\mathcal{P}$ and is thus compromised.

5.4 Pros and Cons of ANP Authentication

Accuracy. Our above experimental results show that the device can be identified using its ANP M-Print and ANP SM-Print with accuracy of 96.4% and 95.3%, respectively. Therefore, ANP fingerprint schemes are sufficiently accurate for mobile authentication.

Deployability. SM-ANP is naturally verifier-agnostic because $\mathcal{V}$'s acoustic elements have no impact on the response. M-ANP can also be verifier-agnostic by adopting the simple calibration as we demonstrated previously.

Security. M-ANP and SM-ANP are both resilient to random impersonation due to the high accuracy of ANP fingerprints. Since both M-ANP and SM-ANP adopt F-responses (i.e., audio files as responses), they are vulnerable to replay and fingerprint-emulation attacks.

5.5 Dynamic Challenge-Response for M-ANP

A straightforward defense against both replay and fingerprint-emulation attacks is to let the authentication server issue unique challenge audios for different authentication sessions to prevent possibly exposed fingerprints from being used for launching fingerprint-emulation attack. Specifically, the authentication server randomly selects κ tone pairs as the challenge for each authentication session and never reuses the same set of κ tone pairs in the future. However, even a subset of reused tone pairs can be used to launch the fingerprint-emulation attack. In what follows, we first quantify the amount of distinct tone pairs and then analyze the resilience of the dynamic challenge-response M-ANP.

5.5.1 ANP M-Print space

To estimate the fingerprint space of M-Print, we first examine the impact of tone frequencies and amplitudes on the distortion components (or equivalently fingerprint elements). Consider two arbitrary challenge tones $A_1 \cos(2\pi f_1 t)$ and $A_2 \cos(2\pi f_2 t)$. Ideally, this tone pair would result in a distortion component at frequency $f_2 - f_1$ with amplitude $a_{i,j} = g_2 A_1 A_2$. However, we observe from experiments that g_2 is not a constant but depends on A_1, A_2, f_1, and f_2. Due to ambient noise and measurement errors, this distortion component may appear at a slightly different frequency and with a slightly different amplitude. Even worse, it may be very similar to the distortion component induced by a different tone pair, say $A_1' \cos(2\pi f_1' t)$ and $A_2' \cos(2\pi f_2' t)$.

To guarantee sufficient distinguishably among different distortion components, it is necessary to ensure that no two tone pairs are very similar in both frequencies and amplitudes. We thus have the following criteria: (1) $\max\{|f_1 - f_1'|, |f_2 - f_2'|\} \geq h_f$; (2) $\max\{|A_1 - A_1'|, |A_2 - A_2'|\} \geq h_a$. As long as at least one criterion is satisfied, the two resulting distortion components can be distinguished with overwhelming probability.

We meet the above requirements by choosing the frequency and amplitude of each challenge tone from a set of predetermined candidate tones with sufficient gaps. In particular, let $f_{\max}$ and $f_{\min}$ denote the highest and lowest acoustic frequencies that can induce significant nonlinear distortion of the microphone, respectively. For example, we can set $f_{\min} = 18$ kHz and $f_{\max} = 50$ kHz according to [12]. The number of possible tone frequencies is then $N_f = \lceil (f_{\max} - f_{\min})/h_f \rceil$. In addition, let $A_{\max}$ and $A_{\min}$ denote the highest and lowest possible tone amplitudes, respectively, leading to $N_a = \lceil (A_{\max} - A_{\min})/h_a \rceil$ possible amplitudes for each tone. $A_{\max}$ depends on the maximum working voltage of the transducer, and $A_{\min}$ must be sufficiently large to induce nontrivial nonlinear distortion and can be obtained through experiments.

Given that $f_2 - f_1$ must be smaller than 18 kHz, we estimate the size of the fingerprint space as follows. For simplicity, assume that 18 kHz can be divided by h_f such that $\lambda = \frac{18\text{kHz}}{h_f}$. When f_1 is smaller than $f_{\max} - 18$ kHz, all the λ frequencies within the range $[f_1, f_1 + 18 \text{ kHz}]$ can be used as f_2. When f_1 is larger than $f_{\max} - 18$ kHz, there are $\frac{f_{\max} - f_1}{h_f}$ frequencies that can be used as f_2. Therefore, there are total $\lambda N_f - \lambda(\lambda + 1)/2$ tone-frequency pairs. Since each tone has N_a possible amplitudes, there are $\psi_\mathrm{M} = (\lambda N_f - \lambda(\lambda+1)/2)N_a^2$ distinct tone pairs, each leading to a unique distortion component (or fingerprint element). Prover $\mathcal{P}$ may have $N_{\mathrm{mic}} \geq 1$ microphones. For example, iPhone models starting from 6s and 6s+ all have four microphones. So we can have $N_\mathrm{M} = N_{\mathrm{mic}}\psi_\mathrm{M}$ unique fingerprint elements of $\mathcal{P}$, leading to $\binom{N_\mathrm{M}}{\kappa}$ distinct fingerprints in total for a challenge with κ audio.

5.5.2 System parameters

Now we discuss how we obtained $A_{\min}$, h_f, and h_a through experiments involving the same set of 20 devices shown in Table 1. Two Prowave 250ST160 transducers were used to generate the challenge audio, and two Agilent 33220A signal generators were used as the power supply.

We obtained the transducer's minimum input voltage $V_{\min}$ instead of $A_{\min}$. We tried 17 voltages ranging from 2 V to 10 V with a step length of 0.5 V as the transducer's input voltage V_{in}. For each V_{in} value, we generated 10 challenges. The tone frequencies of each challenge were randomly chosen, and the amplitudes of all the challenges tones were fixed to $g_1 V_{\text{in}}$, where g_1 denotes the gain factor of the transducer. We extracted the 20 devices' fingerprints corresponding to each of those challenges 20 times and obtained 4,000 testing samples. Then the 20 devices were chosen as $\mathcal{P}$ one by one. Given a chosen $\mathcal{P}$, we extracted its fingerprints corresponding to those 10 challenges one more time as the fingerprint profile stored in $\mathcal{S}$ which were used to classify the 4,000 testing samples. Based on the classification results, we calculated the F_1 score. We totally obtained 20 F_1 scores for each V_{in} value and calculated the average F_1 score. The results show that the average F_1 score increases as V_{in} increases and exceeds 0.95 when V_{in} is larger than 6 V. When V_{in} is lower than 5.5 V, the average F_1 score is below 0.78. Therefore, we chose $V_{\min}$ =6 V and $A_{\min} = g_1 V_{\min}$.

The choice of h_f and h_a should guarantee that a microphone's fingerprints with respect to different challenges are distinguishable, i.e., the distance between two fingerprints is larger than the threshold τ_M. Since the transducer is powered by the signal generator, the amplitude of the challenge tone is determined by the voltage of the signal generator. We denote the voltage corresponding to amplitude A_i by V_i. Since the second requirement for the amplitudes of challenge tones is equivalent to $\max\{|V_1 - V_1'|, |V_2 - V_2'|\} \geq h_v$, we obtained h_v instead of h_a. In M-ANP, the two most similar fingerprints, denoted by Θ and Θ', differ in only one element. Without loss of generality, we assume they differ in the first element and model the distance between the two fingerprints as

$$\mathsf{diff}(\Theta, \Theta') = \frac{\theta_1 - \theta_1'}{\theta_1 + \theta_1'}. \tag{11}$$

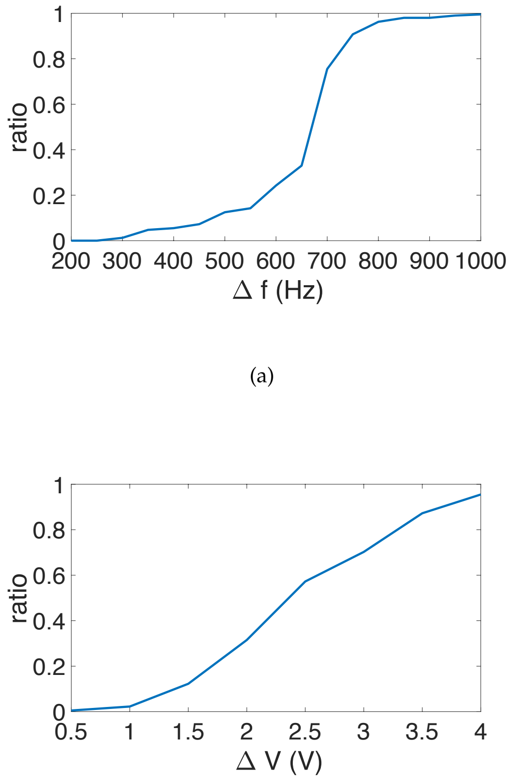
Therefore, two fingerprint elements should be distinguishable if the scaled distance between them is larger than τ_M. We seek to find the minimum values for Δf and ΔV so that the two elements corresponding to $\langle f_1,\ A_1,\ f_2,\ A_2\rangle$ and $\langle f_1,\ A_1,\ f_2 + \Delta f,\ A_2\rangle$ or the two elements corresponding to $\langle f_1,\ A_1,\ f_2,\ A_2\rangle$ and $\langle f_1,\ A_1,\ f_2,\ A_2 + g_1\Delta V\rangle$ are distinguishable.

We conducted an experiments with 20 mobile devices. We selected a base tone pair by randomly selecting a tone frequency pair and fixing the amplitudes of each tone to $A_{\min}$. The base tone pair is denoted by $\langle f_1,\ A_{\min},\ f_2,\ A_{\min}\rangle$. We extracted the fingerprint elements of the 20 mobile devices corresponding to the base tone pair as the reference elements. Then we increased Δf from 200 Hz to 1 kHz with a step length of 50 Hz and increased ΔV from 0.5 V to 4 V with a step length of 0.5 V. For each Δf and ΔV, we extracted the fingerprint elements corresponding to $\langle f_1,\ A_{\min},\ f_2{+}\Delta f,\ A_{\min}\rangle$ and $\langle f_1,\ A_{\min},\ f_2,\ A_{\min}{+}g_1\Delta V\rangle$ of each device for 20 times, where g_1 is the gain factor of the transducer. Totally 400 testing element samples for each individual tone pair were obtained. We repeated the whole process 10 times with different base tone pairs and calculated the scaled distances between each extracted element and the corresponding reference element. If the distance is larger than τ_M, the extracted element is considered distinguishable, and vice versa. Fig. 4(a) and Fig. 4(b) show the ratios of distinguishable elements corresponding to each Δf and ΔV, respectively. We can see that more than 96.3% of fingerprint elements are distinguishable when Δf is no less than 800 Hz, and more than 95.5% of fingerprint elements are distinguishable when ΔV is 4 V. So we adopt $h_f = 800$ Hz and $h_v = 4$ V.

Fingerprint space. Based on our experiment results, we estimate the fingerprint space as follows. There are total $N_f = \lceil (50 \text{ kHz} - 18 \text{ kHz})/800 \text{ Hz} \rceil = 40$ feasible tone frequencies. Since the maximum working voltage of the transducer is 20 V, there are total $N_a = \lceil (20 \text{ V} - 4 \text{ V})/5 \text{ V} \rceil = 4$ feasible tone amplitudes. Therefore, a mobile device with two microphones, like Samsung S7, has about N_M =20,000 distinct fingerprint elements and around 8×10^{42} distinct fingerprints.

5.5.3 Security analysis

Now we analyze the resilience of dynamic challenge-response M-ANP to the fingerprint-emulation attack. Assume that $\mathcal{A}$ has acquired ϵ fingerprints of $\mathcal{P}$ through the



(a)

(b)

Fig. 4. The ratios of distinguishable elements.

three methods illustrated in Section 3.2. $\mathcal{A}$ tries to impersonate $\mathcal{P}$ and starts an authentication instance at $\mathcal{V}$. $\mathcal{S}$ randomly selects a fingerprint $\mathcal{F}_s$ from the fingerprint pool, returns the corresponding challenge, and asks for the response. Since $\mathcal{P}$ fingerprints may contain common fingerprint elements, an element of $\mathcal{F}_s$ is known to $\mathcal{A}$ if another fingerprint containing the element has been exposed to $\mathcal{A}$. The probability P_e that a specific fingerprint element in $\mathcal{F}_s$ has been exposed can be estimated as

$$P_e = 1 - (1 - \frac{\kappa}{N_{\text{M}}})^{\epsilon} \,. \tag{12}$$

As shown in Section 5.5.1, $\mathcal{A}$ can successfully emulate $\mathcal{F}_s$ only if all the κ elements in $\mathcal{F}_s$ are exposed. So the probability for this to occur is given by

$$P_{\text{success}} = P_e^{\kappa} = (1 - (1 - \frac{\kappa}{N_{\text{M}}})^{\epsilon})^{\kappa}. \tag{13}$$

To achieve an attack success rate of 0.5, the attacker need obtain more than 5,000 fingerprints, which would take quite a long time and may not be feasible in practice. Therefore, the dynamic challenge-response scheme for M-ANP is resilient to the fingerprint-emulation attack.

5.6 Dynamic Challenge-Response for SM-ANP

Similar to M-ANP, SM-ANP can adopt the random challenge-response method to withstand the fingerprint-emulation attack. For this purpose, $\mathcal{S}$ maintains a set of fingerprints, referred to as a fingerprint pool, for each prover $\mathcal{P}$. For each authentication request concerning $\mathcal{P}$, the server randomly chooses one fingerprint from the pool and issues the corresponding challenge to $\mathcal{P}$.

5.6.1 ANP SM-Print space.

Now we discuss the ANP SM-Print space, i.e., the number of possible fingerprints for a single prover $\mathcal{P}$ in SM-ANP. Obviously, the scaled Euclidean distance between any two fingerprints should be larger than the threshold τ_{SM}. Since each challenge corresponds to a unique fingerprint, we can just estimate how many distinct challenges in Eq. (8) there can be. We first consider the impact of f_b. Although ideally all the distortion frequencies should be multiples of $f_b \in [f_{b,\min}, f_{b,\max})$ below 22 kHz, they may vary slightly due to noise and measurement errors. We thus require a minimum gap h_{f_b} between different f_bs to ensure that their corresponding distortion frequencies can be distinguished. This means that f_b can take $l_{f_b} = \lceil \frac{f_b^{\max} - f_b^{\min}}{h_{f_b}} \rceil$ values. Moreover, the carrier frequency f_c, A_{f_b}, and A_{f_c} all affect the amplitudes of distortion components. For each given f_b, we require f_c to be a multiple of f_b in $[18, 24)$ kHz, so f_c can take $l_{f_c} = \lceil \frac{6\,\text{kHz}}{f_b} \rceil$ possible values. A_{f_b} and A_{f_c} are determined once the modulation depth $\zeta = A_{f_b}/A_{f_c}$ is chosen from $[\zeta_{\min}, \zeta_{\max})$. We thus introduce a minimum gap h_ζ between different modulation depths so that the corresponding distortion components at the same frequencies can have sufficiently different amplitudes. This means that ζ can take $\lceil \frac{f_b^{\max} - f_b^{\min}}{h_{f_b}} \rceil$ values. Finally, we estimate the number of distinct SM-Print fingerprints for each speaker-microphone pair on $\mathcal{P}$ as

$$\psi_{\text{SM}} = \lceil \frac{f_b^{\max} - f_b^{\min}}{h_{f_b}} \rceil \times \sum_{i=1}^{l_{f_b}} \lceil \frac{6\text{ kHz}}{f_{b,\min} + i h_{f_b}} \rceil \,. \tag{14}$$

As in latest smartphones or smartwatches, prover $\mathcal{P}$ may have $m \geq 1$ speaker-microphone pairs, leading to $N_{\text{SM}} = m\psi_{\text{SM}}$ distinct fingerprints in total.

5.6.2 System parameters

We conducted experiments to obtain $\zeta_{\min}$, $\zeta_{\max}$, h_{f_b}, and h_ζ. 12 Android devices shown in Table 1 were used in experiments.

$\zeta_{\min}$ and $\zeta_{\max}$. Since the impact of ζ on the distortion component's amplitude is more significant when the distortion components' amplitudes are small, we chose f_b and f_c to be 700 Hz and 18.2 kHz, respectively. We increased ζ from 50% to 150 % with a step length of 5% and obtained the corresponding F_1 score for each ζ value. The F_1 score is above 0.95 when M_d is between 75% and 110%. As a result, we chose $\zeta_{\min} = 75\%$ and $\zeta_{\max} = 110\%$.

h_{f_b} and h_ζ. We obtained the minimum values of Δf_b and $\Delta\zeta$ so that the a device's fingerprints corresponding to $\langle f_b + \Delta f_b,\ \zeta \rangle$ and $\langle f_b,\ \zeta + \Delta\zeta \rangle$ are distinguishable from the fingerprint corresponding to $\langle f_b,\ \zeta \rangle$. Since f_b is within [700 Hz, 2 kHz], the length of the fingerprint, i.e., the number of distortion components, is between 11 and 31. We did not consider the length of 11 because 2 kHz is the only available f_b for this length. We first selected 20 challenges whose corresponding fingerprints are of different length. The f_b for the ith challenge is $\lfloor \frac{22\text{kHz}}{11+i} \rfloor$ and denoted by f_b^i, and the corresponding f_c^i is $\lceil \frac{18\text{ kHz}}{f_b} \rceil f_b$. The length of the fingerprint corresponding to the ith challenge is $11 + i$. The modulation depths of all the challenges are fixed as $\zeta_{\min}$. We iteratively chose one of the 12 devices as prover $\mathcal{P}$ and extracted its fingerprints corresponding to those challenges as the reference fingerprints. We then extracted $\mathcal{P}$'s fingerprints corresponding to $\langle f_b^i + \Delta f,\ \zeta_{\min} \rangle$. We tested 20 values of Δf ranging from 20 Hz to 400 Hz with a step length of 20 Hz. For each challenge, we extracted the $\mathcal{P}$'s fingerprints 20 times. If the extracted fingerprint has a different length from the reference fingerprint or the distance between the extracted fingerprint and the reference fingerprint is larger than τ_{SM}, the extracted fingerprint is considered distinguishable from the reference fingerprint. Next, we extracted the $\mathcal{P}$'s fingerprints corresponding to $\langle f_b^i,\ \zeta_{\min + \Delta\zeta} \rangle$ 20 times. We tested 20 values of $\Delta\zeta$ ranging from 1% to 20% with a step length of 1%. The results show that 97% fingerprints are distinguishable when Δf is larger than 120 Hz, and 95% fingerprints are distinguishable when $\Delta\zeta$ is larger than 6%. Therefore, we choose $h_{f_b} = 120$ Hz and $h_\zeta = 6\%$.

Fingerprint space. Based on the obtained parameters, we estimate that a speaker-microphone pair has approximately 580 distinguishable fingerprints. A mobile device with two speakers and two microphones has approximately 2,320 fingerprints. The fingerprint space is much smaller than that of ANP M-Print and may not be sufficiently large for long-term mobile authentication. The main reason is that the mobile device's speaker has limited power and frequency ranges, leading to a relatively small number of distinguishable challenges.

5.7 Summary of ANP System Performance

We summarize the pros and cons of ANP authentication systems in Table 5.

TABLE 5
Pros & Cons of ANP authentication systems.

Authentication system		M-ANP	SM-ANP
Accuracy		96.4%	95.3%
Deployability		ultrasound transducers	none
Security	Impersonation	yes	
	Replay attack	no	
	Fingerprint-emulation	no	
	Dynamic C-R	adoptable	not adoptable

6 Discussion

S-Print and SM-Print authentication systems should adopt FRC fingerprints to identify the prover. Since the nonlinearity of a microphone can hardly be eliminated, it is hard to extract a verifier-agnostic ANP fingerprint of a speaker, leaving FRC as the only choice for S-Print authentication systems. Due to the limited fingerprint space, SM-ANP cannot adopt the dynamic challenge-response countermeasure, so it is vulnerable to fingerprint-emulation attacks. SM-ANP thus has no advantage over FRC in terms of security. The distortion components in SM-ANP may be within the audible frequency range, while the responses and challenges of FRC are all within the inaudible frequency range. So SM-ANP is more disturbing to the user and less resilient to the ambient noise which is also in the audible frequency range.

M-ANP systems are more secure than FRC-based M-Print systems, but the latter are more deployable in some application scenarios. Particularly, the verifier in an FRC-based M-Print system must has a speaker and a microphone. In some application scenarios, speakers and microphones are already installed in the device which can act as the verifier. For example, many commercial smart lockers have speakers and microphones for the communication purpose. In this case, acoustic authentication can be integrated to the existing authentication system without any hardware modification. In contrast, M-ANP requires several ultrasound transducers to be installed on the verifier. Ultrasound transducers are less common compared with commercial speakers and microphone. Hardware modification is almost unavoidable to integrate M-ANP to an existing authentication system. Besides, M-ANP is more disturbing and less resilient to noise compared with FRC-based M-Print systems since it also involves audible distortion components.

7 Related Work

Fingerprinting a mobile device with the unique features of its hardware components has been a hot topic in recent years. The features of motion sensors are used to identify the mobile device in [18], [19], [20]. Researcher leveraged the imperfection of the WiFi chipset to identify the mobile device in [22], [23], [24]. Ba *et al.* proposed to use the Photo-Response Non-Uniformity of the camera as the mobile device's fingerprint [21]. Acoustic elements are more prevalent than the aforementioned components, and it is thus promising to identify a mobile device with its microphone or speaker. There have been many studies on fingerprinting the acoustic elements. Zhou *et al.*, Chen *et al.*, and Han *et al.* all proposed to used the frequency response as the fingerprint of the acoustic element [1], [3], [4]. Das *et al.* proposed to use the mel-frequency cepstral coefficients to identify an acoustic element [5]. In this paper, we investigate identifying the mobile device with the nonlinear feature of its acoustic elements and study the security of our and all the existing acoustic fingerprinting schemes in the context of mobile authentication. All the existing acoustic fingerprinting schemes are vulnerable to the powerful fingerprint-emulation attack, but our scheme can defeat this attack by adopting the dynamic challenge-response mechanism.

The nonlinearity of the acoustic element has been used for different purposes in previous studies. Roy *et al.* studied the feasibility of leveraging the nonlinear distortion of the microphone to record ultrasonic sounds [12]. Zhang *et al.* and Roy *et al.* utilized the nonlinearity of microphones to issue inaudible commands to the voice control system [13], [25]. Lin *et al.* proposed an ultrasonic positioning system for mobile devices using the nonlinearity of the microphone

[26]. The most related work to this paper is NAuth [6]. Zhou *et al.* proposed using ANP to verify the consistency of the audio source in the device-to-device authentication context. NAuth is quite efficient in the targeted context, but it does not fulfill the verifier-agnostic requirement of the distributed authentication system.

8 Conclusion

In this paper, we investigated the suitability of existing acoustic fingerprinting schemes for mobile authentication in terms of accuracy, deployability, and security. While we found that all the schemes achieve sufficiently high identification accuracy for mobile authentication, MFCC acoustic fingerprint schemes incur a prohibitive deployment cost due to the need for expensive acoustic elements. In contrast, FRC and ANP authentication systems are both low-cost and verifier-agnostic but are both vulnerable to the fingerprint-emulation attack. To address these limitations, we proposed a dynamic challenge-response mechanism as a strong defense. The proposed system can thwart the fingerprint-emulation attack by not reusing acoustic fingerprints across different authentication sessions. To evaluate whether the proposed mechanism can be integrated into FRC and ANP authentication systems, we quantify the space of FRC and ANP fingerprints of the speaker, microphone, and speaker-microphone pair on the prover device. Our experiment results show that ANP M-Print is the only scheme with a sufficiently large fingerprint space to support dynamic challenge-response to withstand the fingerprint-emulation attack. In the overall consideration of accuracy, deployability, and security, ANP M-Print is the best choice for the acoustic mobile authentication system.

References

[1] Z. Zhou, W. Diao, X. Liu, and K. Zhang, “Acoustic fingerprinting revisited: Generate stable device ID stealthy with inaudible sound,” in *ACM CCS*, Scottsdale, AZ, Nov. 2014.

[2] B. Hristo, M. Yan, N. Gabi, and B. Dan, “Mobile device identification via sensor fingerprinting,” in *arXiv preprint arXiv:1408.1416*, 2014.

[3] 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,” in *IEEE Internet of Things Journal*, vol. 4, no. 1, pp. 88–100, 2017.

[4] D. Han, Y. Chen, T. Li, R. Zhang, Y. Zhang, and T. Hedgpeth, “Proximity-proof: Secure and usable mobile two-factor authentication,” in *ACM MobiCom*, New Delhi, India, Oct.-Nov. 2018.

[5] A. Das, N. Borisov, and M. Caesar, “Do you hear what i hear?: Fingerprinting smart devices through embedded acoustic components,” in *ACM CCS*, Scottsdale, AZ, Nov. 2014.

[6] X. Zhou, X. Ji, C. Yan, J. Deng, and W. Xu, “Nauth: Secure face-to-face device authentication via nonlinearity,” in *IEEE INFOCOM*, Paris, France, Apr. - May 2019.

[7] N. Aurelle, D. Guyomar, C. Richard, P. Gonnard, and L. Eyraud, “Nonlinear behavior of an ultrasonic transducer,” in *Ultrasonics*, vol. 34, no. 2-5, pp. 187–191, 1996.

[8] C. Peng, G. Shen, Y. Zhang, Y. Li, and K. Tan, “BeepBeep: A High Accuracy Acoustic Ranging System using COTS Mobile Devices," in *ACM SenSys'07*, Sydney, Australia, Nov. 2007.

[9] W. Chen and Q Wu, “A proof of MITM vulnerability in public WLANs guarded by captive portal", in *Asia-Pacific Advanced Network*, Vol. 30, 2010, pp. 66-70.

[10] “Petterson L400 ultrasonic speaker,” 2019. https://batsound.com/product/l400-ultrasonic-speaker/

[11] “Audacity Software,” 2020. https://www.audacityteam.org/

[12] N. Roy, H. Hassanieh, and R. R. Choudhury, “Backdoor: Making microphones hear inaudible sounds,” in *ACM MobiSys*, Niagara Falls, NY, Jun. 2017.

[13] G. Zhang, C.Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, “Dolphinattack: Inaudible voice commands,” in *ACM CCS*, Dallas,TX, Oct.-Nov. 2017.

[14] L. Beranek, and M. Tim,“Acoustics Sound Fields and Transducers“, Oxford: Academic Press, 2012. Print.

[15] S. Gupta, A. Mauro, and A. DeJaco, “Method and apparatus for automatically adjusting speaker and microphone gains within a mobile telephone,” Jun. 2004, US Patent 6,744,882.

[16] (2010) 250st180. https://goo.gl/Z21CsP

[17] “Agilent 33220a,” 2014. https://goo.gl/2ZxkuA

[18] S. Dey, N. Roy, W. Xu, R. Choudhury, and S. Nelakuditi, “Accelprint: Imperfections of accelerometners maks smartphones trackable,” in *NDSS*, San Diego, CA, February 2014.

[19] A. Das, N. Borisov, and M. Caesar, “Tracking mobile web users through motion sensors: Attacks and defenses.” in *NDSS*, February, San Diego, CA 2016.

[20] H. Bojinov, Y. Michalevsky, G. Nakibly, and D. Boneh, “Mobile device identification via sensor fingerprinting.” *arXiv preprint arXiv:1408.1416*, 2014.

[21] Z. Ba, S. Piao, X. Fu, D. Koutsonikolas, A. Mohaisen, and K. Ren, “Abc: Enabling smartphone authentication with built-in camera,” in *NDSS*, San Diego, CA, February 2018.

[22] V. Brik, S. Banerjee, M. Gruteser, and S. Oh, “Wireless device identification with radiometric signatures,” in *ACM MobiCom*, San Francisco, CA, September 2008, pp. 116–127.

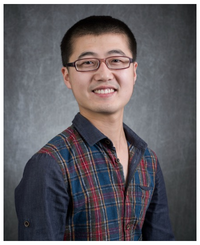
[23] A. Polak, S. Dolatshahi, and D. Goeckel, “Identifying wireless users via transmitter imperfections.” *IEEE Journal on selected areas in communications*, vol. 29, no. 7, pp. 1469–1479, 2011.

[24] K. Remley, C. Grosvenor, R. Johnk, D. Novotny, P. Hale, M. McKinley, A. Karygiannis, and E. Antonakakis, “Electromagnetic signatures of wlan cards and network security.” pp. 484–488, 2005.

[25] N. Roy, S. Shen, H. Hassanieh, and R. Choudhury, “Inaudible voice commands: the long-range attack and defense,” in *USENIX*, Baltimore, MD, August 2018.

[26] Q. Lin, Z. An, and L. Yang, “Rebooting ultrasonic positioning systems for ultrasound-incapable smart devices,” in *ACM MobiCom*, Los Cabos, Mexico, October 2019.

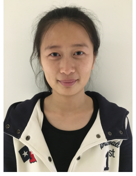
Dianqi Han received a B.S. degree in Information Security from University of Science and Technology of China and a M.E. degree in Computer and Electrical Engineering from the University of California, Davis. He is currently a Ph.D. student in Computer Engineering at Arizona State University. His research interests include indoor navigation, security and privacy issues in mobile systems, and machine learning in wireless networks.



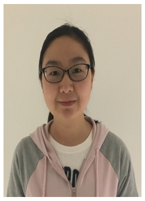
Tao Li received a Ph.D. in Computer Engineering from Arizona State University in 2020, a M.S. in Computer Science & Technology from Xi'an Jiaotong University in 2015, and a B.E. in Software Engineering from Hangzhou Dianzi University in 2012. His primary research is on security and privacy issues in networked/mobile/distributed systems, smart sensing, and wireless networks. He is an Assistant Professor in the Department of Computer and Information Technology at Indiana University-Purdue University Indianapolis (IUPUI).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2021.3053282, IEEE Transactions on Mobile Computing

Ang Li received the B.E. in Network Engineering from Guangxi University, China, in 2010, the M.S. in Computer Science from Beihang University, China, in 2014. Currently, he is a Ph.D. student in Computer Engineering from Arizona State University. His research interest is about security and privacy in social networks, machine learning, wireless networks, and mobile computing.



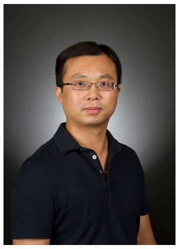
Lili Zhang received the B.S. in Information Security from University of Science and Technology of China, in 2016. Currently, she is a Ph.D. student in Computer Engineering from Arizona State University. Her research interest is about cyber security and privacy issues in mobile systems.



Yan Zhang received the B.S. in Information and Computing Science from Xi'an Jiaotong University, China, in 2014, the M.S. in Communication and Information System from Beijing Normal University, in 2017. Currently, she is a Ph.D. student in Computer Engineering from Arizona State University. Her research interest is about cyber security and privacy issues in mobile systems.

Jiawei Li is a Ph.D. student in Computer Engineering at Arizona State University. He received the B.E. in Telecommunication Engineering from Nanjing University of Posts and Telecommunications at 2013. His research interest is on security and privacy issues in wireless network and wireless sensing.

Rui Zhang received a B.E. degree in communication engineering and a M.E. degree in communication and information system from the Huazhong University of Science and Technology in 2001 and 2005, respectively, and a Ph.D. degree in electrical engineering from Arizona State University in 2013. He was an Assistant Professor with the Department of Electrical Engineering at the University of Hawaii from 2013 to 2016. He has been an Assistant Professor with the Department of Computer and Information Sciences at the University of Delaware since 2016. His current research interests include network and distributed system security, wireless networking, and mobile computing. He received the NSF CAREER Award in 2017 and is a member of the IEEE.



Yanchao Zhang received a B.E. in Computer Science and Technology from Nanjing University of Posts and Telecommunications in 1999, a M.E. in Computer Science and Technology from Beijing University of Posts and Telecommunications in 2002, and a Ph.D. in Electrical and Computer Engineering from the University of Florida in 2006. He is an Professor in School of Electrical, Computer and Energy Engineering at Arizona State University. His primary research interests are network and distributed system security, wireless networking, and mobile computing. He is/was on the editorial boards of IEEE Transactions on Mobile Computing, IEEE Wireless Communications, IEEE Transactions on Control of Network Systems, and IEEE Transactions on Vehicular Technology. He received the US NSF CAREER Award in 2009 and is an IEEE Fellow for contributions to wireless and mobile security. He also chaired the 2017 IEEE Conference on Communications and Network Security (CNS), the 2016 ARO-funded Workshop on Trustworthy Human-Centric Social Networking, the 2015 NSF Workshop on Wireless Security, and the 2010 IEEE GLOBECOM Communication and Information System Security Symposium.