VocalLock: Sensing Vocal Tract for Passphrase-Independent User Authentication Leveraging Acoustic Signals on Smartphones

LI LU, Shanghai Jiao Tong University, China
JIADI YU*, Shanghai Jiao Tong University, China
YINGYING CHEN, Rutgers University, USA
YAN WANG, Temple University, USA

Recent years have witnessed the surge of biometric-based user authentication for mobile devices due to its promising security and convenience. As a natural and widely-existed behavior, human speaking has been exploited for user authentication. Existing voice-based user authentication explores the unique characteristics from either the voiceprint or mouth movements, which is vulnerable to replay attacks and mimic attacks. During speaking, the vocal tract, including the static shape and dynamic movements, also exhibits the individual uniqueness, and they are hardly eavesdropped and imitated by adversaries. Hence, our work aims to employ the individual uniqueness of vocal tract to realize user authentication on mobile devices. Moreover, most voice-based user authentications are passphrase-dependent, which significantly degrade the user experience. Thus, such user authentications are pressed to be implemented in a passphrase-independent manner while being able to resist various attacks. In this paper, we propose a user authentication system, *VocalLock*, which senses the whole vocal tract during speaking to identify different individuals in a passphrase-independent manner on smartphones leveraging acoustic signals. *VocalLock* first utilizes FMCW on acoustic signals to characterize both the static shape and dynamic movements of the vocal tract during speaking, and then constructs a passphrase-independent user authentication model based on the unique characteristics of vocal tract through GMM-UBM. The proposed *VocalLock* can resist various spoofing attacks, while achieving a satisfactory user experience. Extensive experiments in real environments demonstrate *VocalLock* can accurately authenticate user identity in a passphrase-independent manner and successfully resist various attacks.

CCS Concepts: • **Security and privacy** → **Authentication**; • **Human-centered computing** → **Ubiquitous and mobile computing**.

Additional Key Words and Phrases: User authentication; acoustic signal; vocal-tract behavior; FMCW; passphrase-independent

ACM Reference Format:
Li Lu, Jiadi Yu, Yingying Chen, and Yan Wang. 2020. *VocalLock*: Sensing Vocal Tract for Passphrase-Independent User Authentication Leveraging Acoustic Signals on Smartphones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 2, Article 51 (June 2020), 24 pages. https://doi.org/10.1145/3397320

*Jiadi Yu is the corresponding author, Email: jiadiyu@sjtu.edu.cn

Authors' addresses: Li Lu, Shanghai Jiao Tong University, Department of Computer Science and Engineering, 800 Dongchuan Rd, Shanghai, China, 200240, luli_jtu@sjtu.edu.cn; Jiadi Yu, Shanghai Jiao Tong University, Department of Computer Science and Engineering, 800 Dongchuan Rd, Shanghai, China, 200240, jiadiyu@sjtu.edu.cn; Yingying Chen, Rutgers University, WINLAB and Department of Electrical and Computer Engineering, New Brunswick, NJ, USA, 08854, yingche@scarletmail.rutgers.edu; Yan Wang, Temple University, Department of Computer and Information Sciences, Philadelphia, PA, USA, 19122, y.wang@temple.edu.

2474-9567/2020/6-ART51 $15.00
https://doi.org/10.1145/3397320

1 INTRODUCTION

Recent years have witnessed the surge of biometric-based user authentication for mobile devices as it is a promising alternative to classic passwords for user authentication. Among various biometric modalities (e.g., fingerprint [18] and facial [41]), voice has wide applicability due to that the speaking is one of the primary behaviors widely existed in daily work and life. Different from other biometrics, voice-based user authentication can be implemented in a convenient and low-cost manner on mobile devices. Such user authentications are thus commercially available in industrial products like Google Trusted Voice [17], and WeChat Voiceprint Lock [50]. Most voice-based authentications rely on physiological voiceprint to identify different individuals. However, such approaches have been demonstrated to be vulnerable to replay attacks [21, 58, 59], due to the lack of liveness verification.

Taking a close look, when an individual speaks, the audible voice with specific frequencies is modulated and filtered by the movements of multiple components in the vocal tract. Similar to the distinct voiceprint for different individuals, the vocal tract, including the static shape and dynamic movements, also embeds the unique characteristics during speaking [6]. Along this direction, new researches turn to explore the physiological and behavioral characteristics of speaking for realizing secure voice-based user authentication. The whole vocal tract involves the static shape of vocal tract and the highly-coordinated dynamic movements of all the organs in the vocal tract. Existing solutions [27, 46, 58] can only capture the dynamic movements of partial vocal tract (i.e., the lip), due to the limited capability of Doppler-based approaches. Hence, the recordable movements of vocal tract are probable to be imitated by adversaries, thus leading to the vulnerability of [27, 46, 58] to mimic attacks. Another work [59] localizes the phoneme sound inside the vocal tract during speaking to realize the user authentication. Although this work is immune to mimic attacks, the localization approach limits users to fix the smartphone in the same position every time, which is almost impossible for mobile users to follow in practice. Thus, this work is motivated to explore the possibility to utilize both the static shape and dynamic movements of the vocal tract to characterize the individual uniqueness for user authentication.

Additionally, most current voice-based user authentications are passphrase-dependent, i.e., users are required to speak the same passphrase in the register and login, which leads to poor user experience [12]. Although some voiceprint-based solutions [14, 38] explore the in-depth individual uniqueness without specific passphrase contents to realize the passphrase-independent voice-based user authentication, the vulnerability of voiceprint-based approaches to replay attacks still remain unsolved. Despite voiceprint-based solutions, speaking behavior-based approaches (e.g., mouth movements and phoneme localization) can only identify individuals in a passphrase-dependent manner. Therefore, in this work, we not only capture the individual uniqueness of the whole vocal tract for user authentication, but also aim to authenticate user identity in a passphrase-independent manner.

To achieve the passphrase-independent user authentication system leveraging the vocal tract, we consider utilizing acoustic signals to characterize the unique speaking behavior of the whole vocal tract, because the acoustic signals are robust to various environments without the requirement of additional infrastructures. To realize such user authentication, we face a number of challenges in practice. First, since the speaking behavior involves both the static shape and dynamic movements of the vocal tract, we need to accurately characterize the whole vocal tract during speaking with the acoustic signals. Second, since the acoustic signals generated by mobile devices are easily eavesdropped in some physically-insecure spaces, we should well-design the signals from the mobile devices to resist the replay attacks. Finally, since extracted unique features from the vocal tract involve passphrase contents, we need to eliminate such contents to realize the passphrase-independent user authentication model for user-friendly experience.

In this paper, we first characterize the whole vocal tract during speaking leveraging FMCW (Frequency Modulated Continuous Wave) technique on acoustic signals. Since FMCW modulates the signals to capture features of the vocal tract, the features from demodulated signals are difficult to be eavesdropped. Through

analyzing the features of vocal tract from demodulated acoustic signals, we find there are unique patterns of the vocal tract and such patterns could be exhibited in a passphrase-independent manner through statistical methods. Inspired by the observations, we propose a user authentication system, *VocalLock*, which identifies different individuals in a passphrase-independent manner through sensing the whole vocal tract during speaking leveraging acoustic signals on smartphones. In *VocalLock*, a smartphone's speaker continuously transmits acoustic signals modulated by FMCW, and the microphone receives the acoustic signals. Then, *VocalLock* extracts the unique features of the vocal tract from the demodulated signals by FMCW to construct a user authentication model. To improve the user experience, we propose an EDNN (Encoder-Decoder Neural Network) to transfer the features of the vocal tract to that of speech voices, and then employ the speech voice-based GMM-UBM (Gaussian Mixture Model-Uniform Background Model) [38] to construct a passphrase-independent user authentication model. Such an authentication model can not only resist the replay attack and mimic attack, but also authenticate individuals in a passphrase-independent manner.

We highlight our contributions as follows.

- We characterize the whole vocal tract during speaking by FMCW on acoustic signals for user authentication that are resilient to both mimic and replay attacks.
- We extract the unique features from the characterized vocal tract in the modulated acoustic signals, and further propose a user authentication approach by sensing unique characteristics of vocal tract during the speaking.
- We construct a passphrase-independent authentication model based on GMM-UBM to achieve security and maintain the user-friendly experience simultaneously.
- We conduct experiments in different real environments. The results show that *VocalLock* can achieve 91.1% accuracy on average in user authentication and 5.1% false accept rate in attack resistance.

The rest of this paper is organized as follows. We first review related works in Section 2. Then, we present the attack scenario and investigate the feasibility of sensing vocal tract characteristics for passphrase-independent user authentication leveraging acoustic signals in Section 3. Next, Section 4 presents the system overview of *VocalLock*. We further illustrate design details of *VocalLock* in Section 5 and 6. The evaluation results of *VocalLock* are shown in Section 7. We further discuss several limitations in Section 8. Finally, we make a conclusion in Section 9.

2 RELATED WORKS

In this section, we review existing researches related to this work.

Acoustic Sensing Background. Previous studies explored acoustic signals for activity recognition [52], gesture recognition [8], tracking [56], indoor localization [35], and even lip reading-based speaking recovery [45], etc., which supports various practical applications. Recent researches also utilized acoustic signals to replace specific sensors, such as replacing specialized sensors to monitor heart beats [37], breath rates [53], and even replacing cameras for imaging [31].

Classic Approaches. The most prevalent and widely-deployed user authentication is the password [55], which derives PIN and pattern lock [47]. However, such user authentication is knowledge-based and suffers from inconspicuous stealing attacks [25, 29, 60]. To overcome the vulnerability, many biometric-based user authentications are developed for mobile devices, such as fingerprint [18], face recognition [41], iris recognition [40], etc. However, these approaches either require specialized expensive equipments (e.g., infrared camera for Apple FaceID [3]), or are vulnerable to replay attacks due to lack of liveness verification [59] and sensitivity to environments [27].

Voiceprint-based User Authentication. Among various voice-based user authentications, voiceprint-based approach is the most prevalent one [10, 24, 32]. However, voiceprint only measures the physiological characteristics underlying the speech voices during speaking without achieving the liveness verification. This indicates such

approaches are vulnerable to replay attacks, in which an adversary attempts to attack the authentication by using a pre-recorded voice sample collected from a legitimate user. Existing studies [42, 49, 51] validated that replay attacks to the voiceprint-based user authentication can achieve over 50% successful attack rate, which reveals the severe vulnerability of voiceprint-based user authentication. Although recent work [36] designed an end-to-end attack detection system to protect voiceprint-based user authentication, the involvement of WiFi infrastructures may hinder its employment in ubiquitous mobile scenarios.

Mouth Movement-based User Authentication. To defend voice-based user authentication against attacks, recent researches [26, 27, 46, 58, 59] explored the behavioral characteristics of speaking to achieve the liveness verification for user authentication. VoiceLive [59] utilized Time-Difference-of-Arrival (TDoA) of pervasive acoustic signals transmitted from smartphones to localize the phoneme sound for user authentication. Due to the absolute distance measurement through TDoA, users are required to place the smartphone in the same relative position every time, which is almost impossible for mobile device users to follow such strict constraints in practice. Other studies [26, 27, 46, 58] adopted Doppler effect to sense the relative position-independent behavioral characteristics of mouth movements during speaking for user authentication. However, these Doppler-based solutions can only sense the dynamic movements of the vocal tract. This is because the Doppler-based approach measures the moving velocity of a targeted object theoretically, indicating that such solutions could not sense the static characteristics intrinsically. Such dynamic movements of the vocal tract are easily imitated by adversary, which leads to probable mimic attacks. Also, these mouth movement-based solutions are exposed to probable replay attacks, due to the un-modulated signals used for capturing mouth movements. However, all the aforementioned studies only provide passphrase-dependent solutions, which requires users to remember the passphrase in the register for subsequent logins. This strong restriction induces a similar user experience with typical password-based authentication [12]. Hence, it significantly degrades the experience when using such user authentications. A recent work [54] proposed to extract fieldprint from the physical field of acoustic energy as the propagation of the signals, and explored the underlying uniqueness of different individuals for the passphrase-independent user authentication. But this work could not realize the liveness verification of speech, leading to probable failure in resisting well-designed replay attacks.

Table 1. Comparison of voiceprint-based and mouth movement-based user authentication studies.

Work	Biometric	Targeted Task	Performance	Dataset
Dehak et al.[10]	Voiceprint	Passphrase-dependent	1.12% EER	NIST SRE'06 & 08
Matejka et al.[32]	Voiceprint	Passphrase-dependent	2.94% EER	NIST SRE'10
Lei et al.[24]	Voiceprint	Passphrase-independent	1.66% EER	NIST SRE'12
LipPass[26, 27]	Mouth Movements	Passphrase-dependent	90.2% Accuracy	Collected under 12 participants in 4 environments
SilentKey[46]	Mouth Movements	Passphrase-dependent	76.7% TPR	Collected under 50 participants in 2 environments
VoiceGesture[58]	Mouth Movements	Passphrase-dependent	99% Accuracy	Collected under 21 participants in 3 environments
VoiceLive[59]	Mouth Movements	Passphrase-dependent	99% Accuracy	Collected under 12 participants in 2 environments
CaField[54]	Fieldprint	Passphrase-independent	98.4% Accuracy	Collected under 20 participants in 1 environment
Our work	Vocal Tract	Passphrase-independent	91.1% Accuracy	Collected under 25 participants in 3 environments

Table 1 summarizes existing studies of voiceprint-based and mouth movement-based user authentications. Compared with voiceprint-based user authentications, our work employs both the physiological and behavioral characteristics instead of the voiceprint of speaking, which enables the user authentication to resist replay attacks. On the other hand, existing mouth movement-based solutions, including TDoA-based, Doppler-based, and fieldprint-based user authentications, contributed to enhancing voice-based user authentication with the capability of attack resistance. However, TDoA-based solution restricts users to hold the smartphone in the same place during every use, and Doppler-based approaches could not characterize the whole vocal tract for user authentication. Even worse, most existing mouth movement-based solutions except CaField [54] can only realize passphrase-dependent user authentication, which degrades user experiences [12].

Different from the aforementioned researches, our work turns to explore the individual uniqueness embedded underlying both the static shape and dynamic movements of vocal tract for authentication. We creatively employ FMCW (Frequency Modulated Continuous Wave) techniques to sense the individual uniqueness of vocal tract in an attack-resilient manner, and propose a transfer learning-based approach to realize the passphrase-independent user authentication for user-friendly experience. The proposed authentication system outperforms existing approaches on achieving both the security and user friendliness simultaneously.

3 MOTIVATION AND FEASIBILITY STUDIES

In this section, we first describe the attack scenarios of voice-based user authentication. Then, we further characterize the whole vocal tract during speaking for user authentication leveraging acoustic signals, and analyze the security of the vocal tract-based user authentication.

3.1 Attack Scenarios

Speaking is one of the most natural and common activities in human daily life and work, which embeds the unique characteristics of different individuals. Hence, it is feasible to utilize the speaking for user authentication. Voice-based user authentication approaches either employ the voiceprint underlying the speech voices or sense the behavior of mouth movements. All of them suffer from the replay attack and mimic attack.

1) *Eavesdropping and Replay Attack*. An adversary deliberately eavesdrops a legitimate user's speaking including the speech voices and mouth movements in an inconspicuous distance. Then, the adversary attacks the user authentication system through replaying the eavesdropped speaking.

2) *Recording and Mimic Attack*. Before attacking the user authentication, an adversary records a video of a legitimate user's speaking including both mouth movements and speech voices without arousing the user's awareness. Based on the recorded video, the adversary imitates the user's mouth movement or speech voice during speaking to attack the user authentication.

Under these two kinds of attacks, we assume the adversary could only obtain these knowledge (e.g., the eavesdropped speaking or recorded video) in an indirect manner, i.e., without compromising the legitimate users' devices to provide the knowledge for attacks. And we assume the legitimate users are hardly aware of such an indirect eavesdropping and recording, but are conspicuous of the direct eavesdropping and recording, such as a pre-implanted malicious APP in the users' devices.

3.2 Attack-Resisted User Authentication Through Sensing Vocal Tract

To resist various attacks, we explore the individual uniqueness from vocal tract during speaking for the user authentication.

3.2.1 Behavior of Vocal Tract during Speaking. When human speaks, the speech voice is generated by an airflow passing through the whole vocal tract, as shown in Fig. 1. The airflow is first generated from the human lung. Then, the pressure of airflow induces the vibration of vocal cords (including pharynx and larynx) to generate

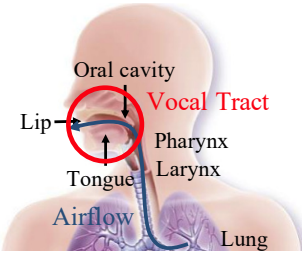


Fig. 1. Illustration of vocal-tract behavior during speaking.

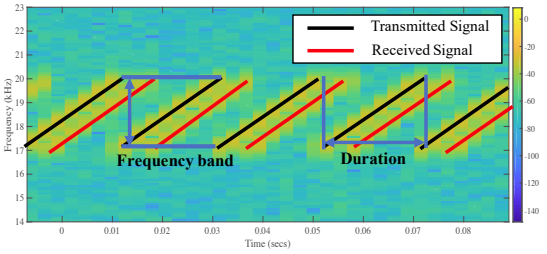


Fig. 2. Illustration of FMCW technique.

an audible voice with specific frequencies. Finally, the airflow carries the voice to pass through the vocal tract (including oral cavity, lip, etc.), in which the voice is further modulated and filtered for calibration.

Usually, the generated speech voice during speaking embeds the uniqueness for different individuals. This is because each individual is endowed with the unique static shape of vocal tract [4]. When an individual speaks, the audible voice with specific frequencies is modulated and filtered by highly-coordinated dynamic movements of all components in the vocal tract, which is different from other individuals [6]. Hence, except for the speech voice is distinct, both the static shape and dynamic movements of vocal tract during speaking embed the individual uniqueness, which exhibits the possibility of employing them in user authentication.

3.2.2 Characterizing Vocal Tract for User Authentication Leveraging FMCW. Since FMCW modulates the signals for the sensing, the demodulated signals are difficult to be eavesdropped. Hence, we consider to characterize a user's vocal tract during speaking leveraging FMCW on acoustic signals.

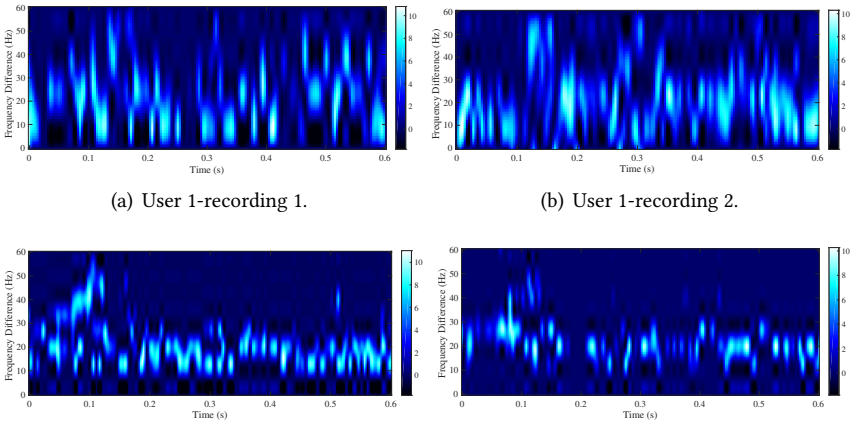
FMCW is a widely-used distance measurement technique for radar. The technique first modulates the signals for transmission, and then demodulates the signals for measuring the distance between the signal source and a target object. Fig. 2 illustrates the basic principle of FMCW. For acoustic signals, the speaker of a smartphone continuously transmits a modulated chirp signal, which sweeps across a bandwidth B ($B = 20kHz - 17kHz = 3kHz$ in the figure) with a duration τ ($\tau = 0.02s$ in the figure). After reflected from the object, the chirp signal is received by the microphone of smartphone. Then, FMCW demodulates the acoustic signal, i.e., performs the dechirp operation [44] on both the transmitted and received signals to measure the frequency difference Δf between transmitted and received signals for Time-Of-Flight (TOF) estimation. Based on the geometry similarity principle of a triangle, the TOF T is derived as

$$T = \frac{\Delta f \times \tau}{B}. \tag{1}$$

Based on Eq. (1), the distance d between the smartphone and an object is derived as

$$d = \frac{c \times T}{2}, \tag{2}$$

where c is the speed of acoustic signals. Through FMCW, we can obtain the distance between an arbitrary point of the vocal tract and the smartphone. We further utilize the FMCW on acoustic signals to characterize a user's vocal tract during speaking. Specifically, a smartphone continuously transmits a modulated chirp signal (which sweeps a specific frequency band) by the speaker during the speaking of a user, and receives the acoustic signal by the microphone. To demodulate the received signal, the system performs the dechirp operation on both transmitted and received signal to derive the frequency difference Δf. From the frequency differences, we can derive the absolute distance between a point in the vocal tract and smartphone based on Eq. (1) and (2). Combining all the absolute distances, we can characterize the vocal tract during speaking. Although there exist some studies [30, 48] employing FMCW on acoustic signals to monitor breathing and realize in-air hand tracking, little work adopts



(a) User 1-recording 1. (b) User 1-recording 2.

(c) User 2-recording 1. (d) User 2-recording 2.

Fig. 3. Examples of the frequency difference induced by speaking 'Hello' under two different users.

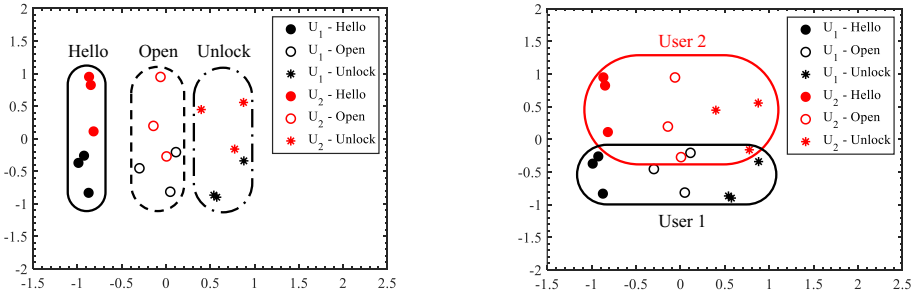
FMCW on acoustic signals to sense mouth movements and further realize the user authentication. To the best of our knowledge, this work is the first research employing FMCW techniques on acoustic signals for user authentication.

To investigate the feasibility of utilizing FMCW to characterize the vocal tract for user authentication, we further conduct an experiment, in which two different users are asked to speak 'Hello' for two times respectively. In each experiment, we fix the distance between smartphone and vocal tract as $5cm$ and place the smartphone directly towards the vocal tract, so as to limit the impact of relative position on FMCW measurements. Fig. 3 shows the examples of frequency difference induced by speaking 'Hello' under two different users respectively. We can observe that the two different speakings of the same user exhibit a similar frequency difference, as shown in Fig. 3(a) and 3(b), as well as Fig. 3(c) and 3(d). On the other hand, the frequency difference of speaking the word 'Hello' presents different variation trends between the two different users. These encouraging results demonstrate that there are unique patterns of the vocal tract during speaking, which can be used to identify different individuals.

3.2.3 Security Analysis. Through characterizing the whole vocal tract with FMCW technique, such user authentication could be immune to various attacks, such as the replay attack and mimic attack.

As aforementioned, we utilize FMCW technique to characterize the whole vocal tract. To extract the features, FMCW requires that the transmitter and receiver are synchronized, i.e., the transmitted and received signals should be obtained under the same clock for feature extraction [30]. For a legitimate user, the transmitter and receiver are both integrated into a smartphone, thus achieving the requirement. However, for an adversary, the transmitter is located in the user's smartphone while the receiver is integrated into the adversary's smartphone. Hence, these two separated devices hardly synchronize, which leads to the adversary not able to decode the correct information from modulated signals. Based on this, user authentications employing FMCW technique can resist replay attacks.

Moreover, during a speaking, both the static shape and dynamic movements of the whole vocal tract exhibit the uniqueness of different individuals. Different from the recordable dynamic movements (such as mouth movements), the static shape of the vocal tract always remains distinct for different individuals and thus is hardly imitated, which helps to resist the mimic attacks for user authentication.



(a) Perspective of passphrases. (b) Perspective of volunteers.

Fig. 4. Distribution of principle components from acoustic patterns induced by two different volunteers speaking three passphrases.

According to the analysis above, we characterize the individual uniqueness of vocal tract leveraging FMCW technique on acoustic signals, which contributes to realizing the attack-resilient user authentication.

3.3 Passphrase-Independent Vocal Tract-based User Authentication

Although both static shape and dynamic movements of the vocal tract could be utilized for user authentication, such an authentication remains passphrase-dependent, i.e., it requires users to remember the passphrase during the register for subsequent logins. Such an authentication manner significantly degrades the user experience [12]. To improve the user experience, we are motivated to construct a passphrase-independent user authentication system. In this section, we investigate the feasibility of distinguishing different individuals through sensing vocal tract in the passphrase-independent manner.

3.3.1 Individual Uniqueness Underlying Sensed Vocal Tract. As mentioned in Section 3.2.1, each individual's vocal tract is endowed with the inborn uniqueness, which could be utilized to distinguish different individuals. During speaking, such differences of the physiological shape of vocal tract further induce the distinct uniqueness of vocal-tract behaviors. Through FMCW technique, both the static shape and dynamic movements of vocal tract could be captured. However, different from the static shape of vocal tract, the vocal-tract behaviors embed not only the inborn uniqueness, but also the passphrase contents. Hence, the mixture of both physiological and behavioral characteristics of vocal tract results in the passphrase-dependence, i.e., such features change following different passphrase contents, instead of remaining fixed for a specific individual.

Different from the individual uniqueness under a specific passphrase, the passphrase-independent uniqueness is hardly observed from the signal patterns directly. Taking a close look at another domain, i.e., the text-independent speaker identification [14], we are inspired to explore the statistical features underlying reflected acoustic signals instead of temporal signal patterns. Existing studies of text-independent speaker identification reveal that individual uniquenesses could be separated from speech contents through extracting the statistics from speech voice sequences. Along this direction, we investigate the feasibility of exhibiting the passphrase-independent uniqueness from signal patterns induced by the whole vocal tract leveraging statistical methods.

3.3.2 Feasibility Study of Passphrase-Independent User Authentication. We conduct a feasibility study to validate whether the passphrase-independent uniqueness could be exhibited through statistical methods. In the experiment, we recruit two volunteers to speak three different passphrases, i.e., 'Hello', 'Open', 'Unlock'. Usually, users tend to select simple passphrases during user authentication for a stronger usability [22], which supports such a

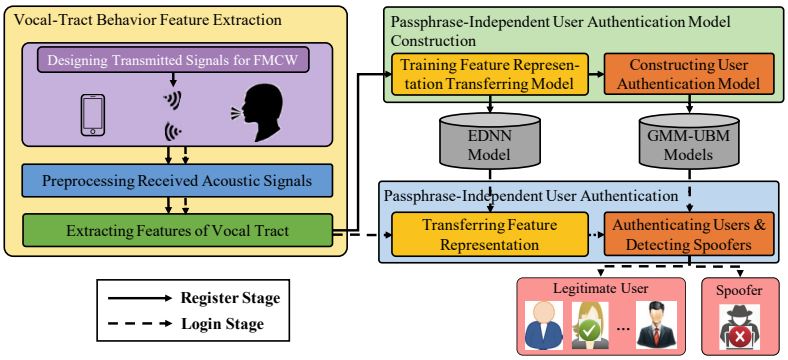


Fig. 5. System architecture of *VocalLock*.

passphrase selection in the feasibility study. To capture the whole vocal tract through acoustic sensing, we implement the FMCW on acoustic signals as illustrated in Section 3.2.2 and deploy it on a Galaxy S6. Each volunteer is required to speak each passphrase three times.

To extract statistics underlying the frequency differences, we employ Principle Component Analysis (PCA) method, which derives correlations between different frequency differences to extract statistical features. Fig. 4 shows the distribution of principal components from acoustic patterns induced by two different volunteers speaking three passphrases. In this figure, the x-axis is the first dominant component of PCA results, while the y-axis means a specially-screened component of PCA results. We can see from Fig. 4(a) that the three different passphrases are distinctly separated. Furthermore, from the other perspective shown in Fig. 4(b), it can be observed that although speaking different passphrases, the two volunteers could still be roughly distinguished. This result supports the feasibility of utilizing statistical features to uncover the passphrase-independent individual uniqueness underlying the vocal tract.

According to the study above, we validate the feasibility of using statistical methods to extract the passphrase-independent uniqueness, which contributes to realizing the user-friendly user authentication.

4 SYSTEM OVERVIEW

To provide both secure and user-friendly mobile authentication experiences, we propose a passphrase-independent user authentication, *VocalLock*, which identifies different individuals through sensing the whole vocal tract during speaking leveraging FMCW on acoustic signals. Fig. 5 shows the architecture of *VocalLock*, which includes two stages, i.e., the register and login stages.

In the register stage, a user speaks a passphrase several times. Meanwhile, a smartphone's speaker continuously transmits the designed acoustic signals that are modulated by FMCW with satisfactory resolution, and then the microphone receives the acoustic signals. *VocalLock* segments the received signals into several episodes in signal preprocessing, each of which represents the behavior of speaking a word. Next, *VocalLock* extracts unique features through FMCW technique, which characterize the user's unique vocal tract during speaking. Given the extracted features, *VocalLock* is able to construct a user authentication model to identify different individuals. In particular, an Encoder-Decoder Neural Network (EDNN) model is first constructed to transfer the feature representation from the vocal tract to speech voice. Based on transferred features, *VocalLock* further constructs a user authentication model through the speech voice-based speaker recognition algorithm, i.e., Gaussian Mixture Model-Uniform Background Model (GMM-UBM), which is able to authenticate individuals in a passphrase-independent manner.

In the login stage, a login user speaks an arbitrary passphrase, which can be either the same or different with that in the register stage, to request the access to the system. During the speaking, *VocalLock* first transmits the

modulated signals and then captures the reflected signals by vocal tract when a user speaks the passphrase, and then preprocesses the signal for signal segmentation. Further, *VocalLock* extracts unique features to characterize the vocal tract through FMCW technique, and transfers the feature of vocal tract to that of speech voice leveraging the constructed EDNN model. Finally, to authenticate the user's identity, *VocalLock* further feeds the transferred feature to the trained GMM-UBM model for verifying the user whether a legitimate user or a spoofer.

5 VOCAL TRACT FEATURE EXTRACTION

Before constructing the user authentication model, *VocalLock* first extracts the unique features of vocal tract through FMCW on acoustic signals.

5.1 Designing Chirp Acoustic Signal for FMCW

VocalLock utilizes FMCW on acoustic signals to sense the vocal tract for user authentication on smartphones. The smartphone continuously transmits a chirp signal by the speaker and receives the acoustic signals by the microphone. To ensure the features of vocal tract can be extracted from the demodulated signals (i.e., the dechirp result of transmitted and received signals), we first need to design the transmitted signals for FMCW and preprocess of received signals.

Designing Transmitted Signals for FMCW. During speaking, the vocal tract involves relatively rapid and minute movements. Hence, to capture such movements, it is necessary to take both the prompt response and sensing resolution into consideration. Thus, we first design the transmitted chirp signal in FMCW to achieve a satisfactory response and resolution for sensing.

In FMCW, chirp signal includes the duration and bandwidth, which has a certain impact on the response and resolution of the acoustic-based sensing. The design of chirp duration is related to the speaking time. Usually, a user's speaking is considered as stationary on short time scales of around $20ms$ [43]. To capture unique characteristics of vocal tract under the short time scale of a stationary speaking, the duration of a chirp signal is set as $20ms$ in our system.

Moreover, for the bandwidth design of chirp signal, it is essential to ensure that the transmitted signal provides enough resolution for sensing the vocal tract while keeping inconspicuous to users for user-friendly experience. According to Fourier Transformation theory [20], two chirp signals reflected by the vocal tract can be resolved in frequency when the frequency difference satisfies $\Delta f > 1/\tau$, where Δf is the frequency difference between the two chirp signals and τ is the duration of the chirp signal. Based on Eq. (1) and (2), we can derive $\Delta f = \frac{2dB}{c\tau}$, where d is the distance between the smartphone and vocal tract, B is the bandwidth of chirp signal and c is the speed of acoustic signals. With the inequation and equation above, we can obtain

$$d > \frac{c}{2B}. \tag{3}$$

This indicates that the resolution of FMCW increases as the bandwidth increases. Hence, the bandwidth should be selected as large as possible. Combining the inconspicuous human auditory range (i.e., $\geq 16kHz$) with limited recoding capability of smartphones (i.e., $\leq 24kHz$) [23], we select the bandwidth of $[16, 24]$ kHz for the transmitted chirp signal.

With the designed chirp signal, *VocalLock* can sense the vocal tract during speaking with satisfactory response and resolution.

Preprocessing Received Acoustic Signals. The smartphone's speaker continuously transmits the acoustic signals modulated by FMCW, and the microphone receives the acoustic signals. The received acoustic signals need to be first segmented into signal episodes of speaking a word. The speaking is usually not a consistent process, in which there exists an inactive period (i.e., non-speaking state) between two active periods (i.e., speaking a word). Usually, the short interval of an inactive period is around $300ms$ [27]. As shown in the top part of Fig.

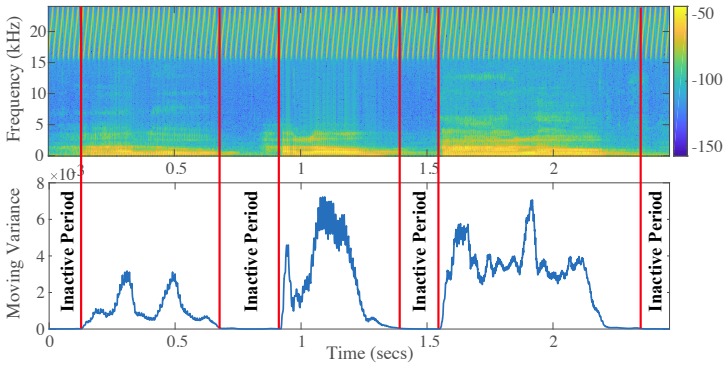


Fig. 6. Illustration of signal segmentation.

6, the speech voices lying in $[0.3, 5]kHz$ exhibit significant differences between speaking and non-speaking on the frequency response, which can be used to separate an active period from inactive periods. We utilize the moving variance on received signals to extract the difference between active and inactive periods for signal segmentation. Fig. 6 illustrates the signal segmentation through applying moving variance on received signals. It can be observed that the moving variance of received signals in an arbitrary inactive period is zero, while that in active periods is non-zero. Thus, we can use a sliding window to segment the received signals into signal episodes of active periods.

5.2 Extracting Features of Vocal Tract

After the signal is preprocessed, *VocalLock* further extracts features of vocal tract from the acoustic signals. A speaking process usually involves the vocal tract (including the static shape and dynamic movements) and speech voice. Since the received signals include both chirp signals and voice signals that characterize the vocal tract and speech voice respectively, we first separate these two kinds of signals in the received signals. Usually, the voice signal lies in the frequency range of $[0.3, 5]kHz$, which is different from the chirp signal of $[16, 24]kHz$. Hence, we use a highpass filter with cut-off frequency $16kHz$ and Equiripple window to extract the chirp signals for further extracting features of vocal tract. After the chirp signals are separated, *VocalLock* can extract the features from the signals.

Mitigating Multipath Effect. Due to the short distance between microphone and speaker, LOS signal is far more significant than reflected signals. Since LOS signal always exhibits stable frequency response in received signals, it is possible to separate LOS signal with reflected signals through comparing the frequency response. Hence, we propose a heuristic method based on Short Time Fourier Transformation (STFT) [1], which divides a long-time signal into short sliding windows and derives frequency responses through Fourier transform separately on each sliding window. Specifically, we perform STFT operation on the demodulated signal (i.e., $s_d = s_t \times \bar{s}_r$, where s_t and s_r are the transmitted and received signals respectively) and search the frequency difference Δf with the n^{th}-largest frequency response on each sliding window, i.e.,

$$\Delta f_n(t) = \arg_f \max^n FFT_f(s_d(t)), \tag{4}$$

where $FFT_f(\cdot)$ is the Fast Fourier Transformation (FFT), t is the index of the sliding window, $\max^n(\cdot)$ is the operation selecting the n^{th}-maximum value. After that, a Frequency Difference Series (FDS) with the n^{th}-largest frequency response $\Delta f_n(t)$ $(t = 1, \cdots, N)$ is extracted (i.e., n^{th}-FDS), which represents a specific signal component (i.e., LOS signal or reflected signals) in the received signal. To match the short time scale of a stationary speaking,

we use a sliding window of $20ms$ with a 50% overlap for FFT operation of each sliding window in the heuristic method.

After separating reflected signals with LOS signal, we further extract the reflected signals by the user's vocal tract from the FDS feature Δf_n. Usually, the distance between a user's vocal tract and smartphone is far less than that between other ambient objects and the smartphone. Hence, TOF of signal reflected by vocal tract is far smaller than that by ambient objects, which leads to a smaller value of frequency difference. Based on the analysis, we set a threshold on the FDS features to separate reflected signal by the user's vocal tract from that by ambient objects. After that, we can extract the FDS features of reflected signals by the user's vocal tract from received signals. Based on the FDS features, we can derive the absolute distance between a point of the vocal tract and the smartphone based on Eq. (1) and (2). However, such absolute distances limit users to fix the smartphone in the same place every time, which is almost impossible to follow.

Extracting Position-Irrelative Features. To release the strict restriction, we first need to convert the absolute distance to relative distance, which is irrelative with the relative position between the smartphone and vocal tract. Based on the absolute distance between the smartphone and a point of the user's vocal tract, we can derive the relative position between two arbitrary points in the vocal tract. Fig. 7 illustrates the conversion from the absolute distance to the relative distance. There are two points A and B in the vocal tract, whose distance is Δd. During speaking, the smartphone measures the absolute distances between the smartphone and two points A as well as B in the vocal tract as d_A and d_B through FMCW respectively. According to the cosine law, we can derive

$$\Delta d^2 = d_A^2 + d_B^2 - 2d_A d_B \cos\alpha, \tag{5}$$

where α is the angle between the sides of d_A and d_B in the triangle. Usually, due to the limited space of the vocal tract, the value of α is relatively small. For example, we assume the distance between the lip and smartphone is $3cm$, so the values of d_A and d_B are set as $2.51cm$ and $3cm$ respectively. Considering the limited size of the vocal tract, the value of Δd is assumed as $0.5cm$. Following Eq. (5), the value of α can be derived as $2.08°$, which is relatively small. Hence, Eq. (5) can be approximately derived as

$$\Delta d^2 \approx d_A^2 + d_B^2 - 2d_A d_B = (d_A - d_B)^2. \tag{6}$$

From Eq. (6), we can find that there is the one-to-one correspondence between the relative position of two points in the vocal tract and difference between the two distances of the smartphone and vocal tract (i.e., the relative distance). Thus, the vocal tract can be characterized by the relative distance independent of relative position between smartphone and vocal tract. We apply moving variance method to convert the absolute distance to relative distance. Specifically, we first derive the moving variance of frequency differences in each time slot. Then, the frequency difference with non-zero moving variances is extracted, which corresponds to both the static shape and dynamic movements of vocal tract during speaking. After that, all frequency differences are converted from the absolute distances to relative distances through subtracting the minimum frequency difference with non-zero moving variance, which extract the features of vocal tract in the position-irrelative manner.

Through the approach above, *VocalLock* extracts FDS features characterizing the vocal tract during speaking based on FMCW technique, which are extracted in an attack-resilient manner.

6 PASSPHRASE-INDEPENDENT USER AUTHENTICATION MODEL CONSTRUCTION

Based on the extracted FDS features, *VocalLock* can construct a user authentication model on smartphones. Since FDS features are extracted from the acoustic signal variation induced by the vocal tract, especially the dynamic movements during speaking, the features embed the knowledge about the content of passphrase that is predefined in the register. This indicates that the user authentication directly employing FDS feature is passphrase-dependent. To improve the user experience, we consider to construct a user authentication model based on extracted FDS features of the vocal tract for smartphones in a passphrase-independent manner.

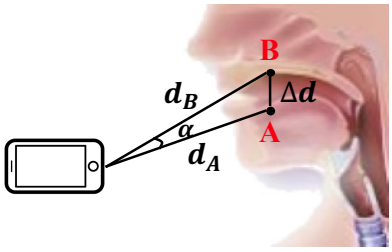


Fig. 7. Illustration of conversion from absolute distance to relative distance.

As mentioned in Section 3.3, the passphrase-independent features could be exhibited through the statistical approaches. Along this direction, we employ Gaussian Mixture Model-Uniform Background Model (GMM-UBM) [38], which explores the statistical characteristics that merely related with the individual uniqueness to achieve the passphrase-independent user authentication. Typically, GMM-UBM authenticates users with Mel Frequency Cepstral Coefficient (MFCC) [10, 24, 32] features, which characterize the speech voice during speaking. Although FDS feature and MFCC feature are extracted from the same speaking process, they exhibit significant difference in characterizing the speaking.

To bridge the difference for the passphrase-independent authentication, we first design a feature-based transfer learning model to transfer the feature representation from FDS features to MFCC features. Then, we construct the passphrase-independent model based on GMM-UBM.

6.1 Building Feature-based Transfer Learning Model for Feature Representation Transferring

VocalLock extracts the FDS features based on FMCW technique, which characterize both the static shape and dynamic movements of vocal tract during speaking. However, such FDS features are not directly related with MFCC features, which characterizes the speech voice based on the auditory properties of human ear. To build a connection between the two kinds of features, we design a feature-based transfer learning method for feature representation transferring from FDS features to MFCC features.

6.1.1 Encoder-Decoder Neural Network Design for Feature-based Transfer Learning. During a speaking process, FDS features can be extracted from the vocal tract, while MFCC features can be generated by the speech voice. Hence, we consider there is a uniform representation between the FDS feature and MFCC feature. To model the connection between the two kinds of features, we propose an Encoder-Decoder Neural Network (EDNN) based on the feature-based transfer learning, which maps the FDS features to MFCC features.

Fig. 8 shows the architecture of EDNN based on the feature-based transfer learning. The input FDS feature could be regarded as a frequency-domain feature during a specific period. To suppress the passphrase information from the FDS feature as much as possible for the passphrase-independent user authentication, we realize the EDNN network using a convolution-based neural network, because of its strong capability to exploit in-depth spatial characteristics underlying the original input [28]. The proposed EDNN consists of two convolutional encoders and two deconvolutional decoders. Each convolutional encoder is based on Convolutional Neural Network (CNN), which consists of three layers, i.e., a *convolutional layer*, a *pooling layer* and a *normalization layer* [19]. The convolutional layer first abstracts the input features as several blocks of compressed representations, and then the pooling layer reduces the dimension of extracted representation in each block through the average pooling operation. The normalization layer finally normalizes the representation to accelerate the convergence during the training. On the other hand, each deconvolutional decoder is based on Deconvolutional Network (DeconvNet) [57], which consists of three layers, i.e., an *unpooling layer*, a *deconvolutional layer* and a *normalization layer*.

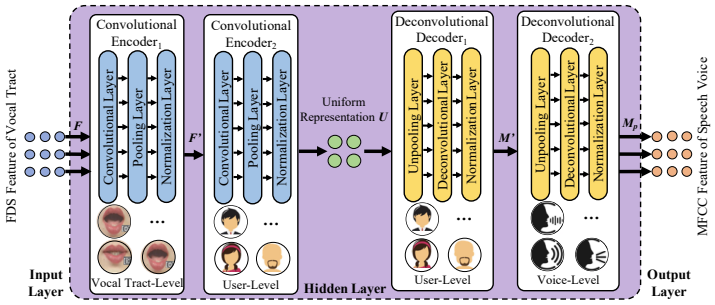


Fig. 8. Architecture of the EDNN for transferring feature representation.

The first two layers perform the inverse operations of the pooling layer and convolutional layer respectively to reconstruct the feature, and the normalization layer finally normalizes the representation.

Given FDS feature F of vocal tract during a user's speaking, the two convolutional encoders in the EDNN first compress the feature to uniform representation of the speaking behavior. Specifically, the first encoder $e_1(F)$ abstracts FDS feature F to vocal tract-level characteristics F' of the speaking with 32 convolutional kernels of 3×3-dimension. Then, the second encoder $e_2(F')$ further compresses the vocal tract-level feature F' to the user-level feature U (i.e., the uniform representation of speaking for a user) with 64 convolutional kernels of 3×3-dimension. Both encoders employ ReLU (i.e., Rectified Linear Unit) activation function [15] and 2×2-dimension average pooling filter. After that, the two deconvolutional decoders can generate the MFCC features of voice based on the uniform representation. The first decoder $d_1(U)$ uncompresses the uniform representation U to the user-level feature M' with 64 deconvolutional kernels of 3×3-dimension. Then, the second decoder $d_2(M')$ further regenerates the voice-level characteristics M_p, i.e., the transferred MFCC feature, which is called as the Vocal-Tract-transferred MFCC (VT-transferred MFCC) with 32 deconvolutional kernels of 3×3-dimension. Also, all of the two encoders and two decoders employ the batch normalization method in the normalization layer to accelerate convergence velocity of training. To enable the EDNN with the capability of transferring FDS feature to MFCC feature, the EDNN is trained with the objective as follows:

$$\min DIF(M, M_p) = \min ||M - M_p||_2 + \lambda \Omega_{weights}, \tag{7}$$

where M_p is the VT-transferred MFCC feature from the EDNN, i.e., $M_p = d_2(d_1(e_2(e_1(F))))$, M is the input MFCC feature of users' speech voice, $|| \cdot ||_2$ is the L_2 norm, $\Omega_{weights}$ is L_2 regulariser for the parameters and λ is the coefficient of $\Omega_{weights}$.

6.1.2 Model Training. To train the EDNN model, *VocalLock* requires both FDS features of vocal tract and MFCC features of speech voice. The received acoustic signals consist of the chirp signal and voice signal. With the chirp signal, the FDS features are extracted as mentioned in Section 5.2. On the other hand, the MFCC features can be extracted from the voice signal. Specifically, we first use a lowpass filter with cut-off frequency of $5kHz$ and Equiripple window to extract the voice signal in received signals. Then, we derive the frequency response of speech voices with a Hanning sliding window of $20ms$ and 50% overlap, which matches parameters in FDS extraction. After that, the frequency response is filtered by the filterbanks to derive the ear-sensitive components in speech voices. The number of filterbank channels is set to 35 that corresponds to $[133, 4860]Hz$, which matches the frequency passband of speech voices. Finally, the MFCC can be derived through performing discrete cosine transformation on the filtered frequency response. Combining the extracted MFCC of speech voice with FDS of vocal tract, the EDNN model can be trained through optimizing Eq. (7) for transferring the feature representation from FDS feature to MFCC feature, i.e., the VT-transferred MFCC features.

Since ENDD model is an individual-independent model for feature representation transferring, the model can be pre-trained with the data collected from other individuals, instead of the legitimate users registered in the system. Hence, the training of EDNN does not require a large number of data samples in the register stage to ensure user-friendly experience. Moreover, instead of directly utilizing MFCC feature for user authentication, *VocalLock* actually exploits the FDS features, which embed the liveness information, for user authentication. Thus, such a FDS feature-based user authentication can resist the replay attacks.

6.2 Constructing User Authentication Model

To implement the passphrase-independent user authentication, we construct the user authentication model based on the transferred feature (i.e., the VT-transferred MFCC feature) utilizing the GMM-UBM.

GMM-UBM exploits the statistical characteristics that are merely related with the individual uniqueness to achieve the passphrase-independent user authentication. GMM-UBM utilizes a UBM (i.e., a special GMM in the model), which is pre-trained with a large number of individuals, to build an individual-independent model. Specifically, GMM uses multiple Gaussian distributions to fit the characteristics of a specific user, i.e.,

$$p(x|\lambda) = \sum_{i=1}^{K} \alpha_i \phi(x|\mu_i, \Sigma_i), \tag{8}$$

where x is voice feature of a user, $\phi(x|\mu_i, \Sigma_i)$ is the i^{th} Gaussian distribution, μ_i and Σ_i are the mean and variance of the distribution respectively, α_i is the weight of i^{th} Gaussian distribution, K is the number of Gaussian distributions and λ is the parameter set of GMM (i.e., $\lambda = \{\alpha_i, \mu_i, \Sigma_i | i = 1, \cdots, K\}$). The pre-trained data could be from an existing open user identification dataset that includes a large number of different individuals, such as VoxCeleb [34]. Specifically, in the register stage, an individual-independent GMM is first trained as the UBM through the Expectation-Maximization (EM) method [11] with MFCC features of existing open-source user identification dataset [34], which includes a large number of individuals. Then, *VocalLock* calibrates the pre-trained UBM to GMM with the VT-transferred MFCC feature of the user through Bayesian adaptation [13] for constructing a unique GMM for each user. Given the trained UBM and VT-transferred MFCC feature of the user, *VocalLock* estimates a parameter set λ to best match the user's features, i.e., maximizing the posteriori probability,

$$\arg\max_{\lambda} p(\lambda|M_p) = \arg\max_{\lambda} p(M_p|\lambda)p(\lambda), \tag{9}$$

where M_p is the VT-transferred MFCC feature of the user, $p(\lambda)$ is the prior probability of the parameter set λ in the pre-trained UBM, and $p(M_p|\lambda)$ is the likelihood. The user's GMM is trained with the objective of Eq. (9) to authenticate the identity of the user. Therefore, after registering for the model training, *VocalLock* can build a GMM that is calibrated from pre-trained UBM for each user, which is used for the passphrase-independent user authentication and spoofer detection.

6.3 Authenticating Users & Detecting Spoofers

When a user requests the login access, the user can speak an arbitrary passphrase (i.e., can be either the same or different with the passphrase in the register stage) to *VocalLock* for login. *VocalLock* first preprocesses the received signal and extracts FDS features from the signal. Then, the FDS feature of vocal tract is transferred to VT-transferred MFCC feature through the trained EDNN. Next, based on the VT-transferred MFCC feature, *VocalLock* needs to determine which legitimate user best matches the feature for user authentication. Specifically, there are n users registered in the system, which constructs n GMM-UBM models for the legitimate users, i.e., $GMM_1, \cdots, GMM_n$. The login user L logins the system with the VT-transferred MFCC feature m. To authenticate the user's identity, *VocalLock* derives the log-likelihood ratio [39] with the model of each legitimate user, i.e.,

$$\Lambda_i(m) = \log p(m|\lambda_{GMM_i}) - \log p(m|\lambda_{UBM}), \tag{10}$$

where λ_{GMM_i} and λ_{UBM} are the parameter sets of the i^{th} user's GMM-UBM and the UBM models respectively. The login user is identified as the i^{th} user with the maximum log-likelihood ratio, i.e., $L = \arg\max_i \Lambda_i(m)$.

A login user can be a curious or malicious spoofer without legitimate identity in the system. Hence, *VocalLock* should detect the spoofer in the login stage to prevent the smartphone from malicious use. Assume a spoofer intends to login the system. Since the FDS feature of vocal tract can characterize the user uniqueness, the VT-transferred MFCC feature has the same capability of exhibiting the uniqueness. This induces a quite low value of $\Lambda_i(m), i \in [1, n]$, i.e., the log-likelihood ratio for any legitimate user under a spoofer's feature. Hence, if the maximum log-likelihood ratio of the login user is lower than empirical results, *VocalLock* regards the login user as an illegal spoofer.

Through the EDNN-based feature transferring and GMM-UBM-based authentication model, *VocalLock* could authenticate individual identity in a passphrase-independent manner, realizing the user-friendly authentication system.

7 PERFORMANCE EVALUATION

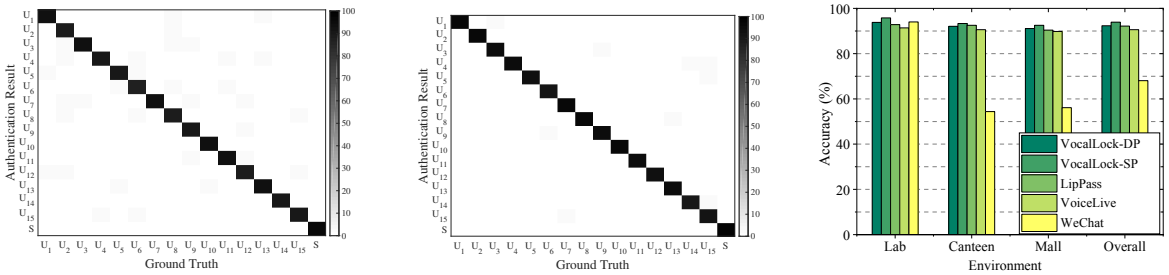
In this section, we evaluate the performance of *VocalLock* under the collected data from 75 volunteers in three real environments.

7.1 Experimental Setup & Methodology

We implement *VocalLock* on three commercial smartphones, i.e., a Galaxy S6, a Xiaomi 6 and a Huawei P10. The transmitted acoustic signal for FMCW is designed as mentioned in Section 5.1. The sampling rate of microphone in the smartphones is set as $48kHz$. Our experiments are conducted in three different environments, i.e., a lab (quiet and few people walking around), a canteen (noisy but few people walking around), and a mall (noisy and many people walking around). In each environment, we randomly recruit 25 volunteers including 14 males and 11 females with the ages ranging in $[20, 45]$ to conduct the experiments, so there are 75 volunteers for the experiments in total. Among the 25 volunteers, 15 of them register the system as legitimate users, while the rest 10 as spoofers. In each experiment, each volunteer randomly selects a smartphone and holds the smartphone with the microphone directed towards the vocal tract. The distance between microphone and volunteer's vocal tract ranges in $[3, 20]cm$. Each of the 15 legitimate users are required to speak a passphrase 3 times in the register stage. Each user repeats the register with 20 different passphrases respectively to comprehensively take the impact of passphrase content into consideration. The passphrases used in the experiments are selected from Word Frequency [5]. Each passphrase contains 1-4 words, and each word in the passphrase is with the phonemes more than 3. But during the login stage, the volunteers (including legitimates users and spoofers) can speak an arbitrary passphrase that can be either the same or different with that in the register stage for the login. The experiments require each volunteer to perform 20 times passphrase speaking in the login stage. The UBM model of *VocalLock* is trained by VoxCeleb [34] dataset involving 1,251 different persons. Then, the GMM-UBM for each legitimate user is trained by calibrating the UBM with each legitimate user's collected data.

To evaluate the performance of *VocalLock*, we define several metrics as follows.

- *Confusion Matrix.* Each row and each column of the matrix denotes the ground truth and the authentication result of *VocalLock* respectively. The i^{th}-row and j^{th}-column entry of the matrix shows the percentage of samples that are authenticated as the j^{th} user while actually are the i^{th} user.
- *Accuracy.* The probability that a user who is A is exactly authenticated as A.
- *False Accept Rate.* The probability that a user not a legitimate user is authenticated as a legitimate user.
- *False Reject Rate.* The probability that a user not a spoofer is authenticated as a spoofer.
- *Response Time.* Assume the end time of a user's speaking is t_{speak}, and the time of a user successfully logins the system is t_{login}. The response time is defined as $t = t_{login} - t_{speak}$.



(a) Confusion matrix of *VocalLock* under different passphrases in register and login stages.

(b) Confusion matrix of *VocalLock* under the same passphrase in register and login stages.

(c) Accuracy of *VocalLock* under different passphrases (DP), and the same passphrase (SP), *LipPass*, *VoiceLive*, WeChat.

Fig. 9. Performance of *VocalLock* on user authentication in three different environments.

• *Energy Consumption.* Assume the battery power of a mobile device is P. When a user activates *VocalLock* for register or login to the mobile device, the system consumes $w\%$ of the power. The energy consumption is defined as $p = P \times w$.

7.2 Performance on User Authentication

We first evaluate the overall performance of *VocalLock* on user authentication. Fig. 9(a) shows the confusion matrix of *VocalLock*, each entry of which is the average accuracy of that in three different environments. We can see that *VocalLock* can achieve an average accuracy of 90.4% in authenticating a legitimate user's identity, and an average accuracy of 96.7% in detecting a spoofer. Overall, the average accuracy (including legitimate user identification and spoofer detection) of *VocalLock* on user authentication is 91.0% with a standard derivation of 3.1%. We also evaluate the performance of *VocalLock* under the same passphrase in the register and login stages, and the results are shown in Fig. 9(b). It can be observed that *VocalLock* can achieve an average accuracy of 94.0% in user authentication and 97.8% in spoofer detection under the same passphrase, which are both higher than that of *VocalLock* under different passphrases. This is because the knowledge of passphrase can be exploited to improve the performance [14]. But we also find that the difference of *VocalLock*'s accuracy between different passphrases and the same passphrase is not significant, i.e., only 2.2%.

We further perform an experiment to compare the performance of *VocalLock* under different passphrases (DP) and the same passphrase (SP) with the mouth movement-based user authentications (i.e., *LipPass* [26, 27] and *VoiceLive* [59]) and voiceprint-based user authentication (i.e., WeChat [50]) in three different environments. In each experiment, the 10 legitimate users are required to speak the 6 predefined passphrases 3 times to the four user authentication systems respectively in the register stage. During the login stage, the volunteers (including legitimates users and spoofers) perform 20 times to the four systems respectively for the login. Note that *VocalLock*-SP, *LipPass*, *VoiceLive* and WeChat require volunteers to speak the same passphrase with that in the register stage for login, while *VocalLock*-DP does not have such a constraint. Fig. 9(c) shows the accuracies of *VocalLock*-DP, *VocalLock*-SP, *LipPass*, *VoiceLive* and WeChat in the three environments respectively. It can be observed that the accuracy of *VocalLock* under different passphrases is 93.8% in the lab, which is similar to 92.8% of *LipPass* and 94.0% of WeChat. This indicates *VocalLock* can achieve satisfactory performance on user authentication. But the accuracy of *VoiceLive* is a little lower than the other three methods, i.e., 90.6%. This is because *VoiceLive* requires users to keep the same relative position between smartphone and their mouth, which is difficult for users to follow. Hence, slight position shift induces performance degradation. Also, the *VocalLock* under the same passphrase can achieve a higher accuracy than *VocalLock* under different passphrases, i.e., 95.8% in the lab. But

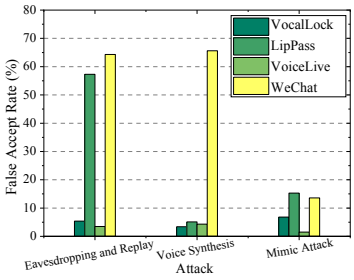


Fig. 10. False accept rate of *VocalLock*, *LipPass*, *VoiceLive* and WeChat under three environments.

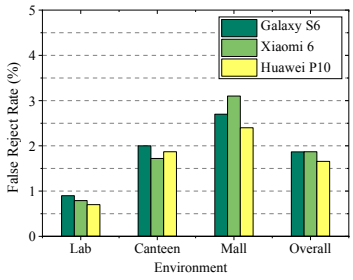


Fig. 11. False reject rate of *VocalLock* under three environments.

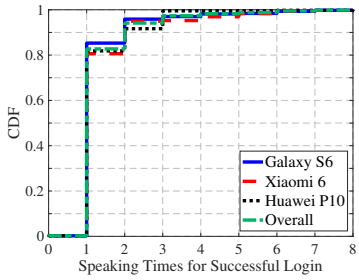


Fig. 12. CDF of speaking times for successful login under three smartphones.

the difference between them is still not significant, which is consistent with the results of the two confusion matrices. Moreover, we can see that the accuracies of *VocalLock* under different passphrases are 93.8%, 92.1% and 91.1% in the three environments respectively, which exhibits insignificant difference of *VocalLock*'s accuracies in various environments. On the contrary, WeChat based on voiceprint suffers from significant performance degradation in some environments. Specifically, the accuracies of WeChat decrease to 54.4% and 56.1% in noisy environments respectively, i.e., the canteen and mall.

7.3 Performance on Attack Resistance

To demonstrate that *VocalLock* can resist various attacks in real authentication scenarios, we conduct an experiment under three kinds of attacks. The first attack is the eavesdropping and replay attack, in which an adversary places a smartphone close to a legitimate user and records the acoustic signals including reflected signals by the vocal tract and voice signals. The second attack is the voice synthesis attack, i.e., an adversary indirectly traces a few voice samples of a legitimate user by recording daily speech or phone conversation, and synthesizes the target voices of the legitimate user through a speech synthesizer [33] for attack. The third attack is the mimic attack, in which an adversary mimics the movement of a legitimate user's vocal-tract behavior during speaking for the attack.

We recruit 24 volunteers for the experiment. The 1^{st}-6^{th} volunteers register to the system as legitimate users, and the rest 18 volunteers are divided into 3 groups equally to attack the system as adversaries through eavesdropping and replay (i.e., 7^{th}-12^{th} volunteers), voice synthesis (i.e., 13^{th}-18^{th} users), and mimic attack (i.e., 19^{th}-24^{th} users) respectively. The experiments are repeated in the three real environments respectively, i.e., lab, canteen and mall. In each experiment, the 6 legitimate users register to the system through speaking a passphrase 3 times, and the 18 adversaries try to login the system with the three kinds of attacks respectively, in which each adversary performs the attack 12 times. We also repeat the experiments on *LipPass*, *VoiceLive* and WeChat to compare the performance on attack resistance among the four user authentication systems.

Fig. 10 shows the false accept rates of *VocalLock*, *LipPass*, *VoiceLive* and WeChat under the three attacks respectively. We find that the false accept rates of *VocalLock* under the three kinds of attacks are all less than 10%, which indicates *VocalLock* can resist various attacks. For mouth movement-based user authentication (i.e., *LipPass*), the false accept rates under voice synthesis and mimic attacks are both below 15%, while that under eavesdropping and replay attack is 57.3%. This result demonstrates the vulnerability of mouth movement-based user authentication to the eavesdropping and replay attack. Also, *VocalLock* outperforms *LipPass* in resisting mimic attacks with FAR approaching 5%. This is because the adversary can only mimic the dynamic movements of vocal tract, instead of the static shape of multiple organs in the vocal tract. Another mouth movement-based solution,

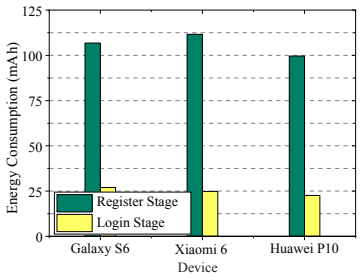


Fig. 13. Energy consumption of *VocalLock* in register and login stages.

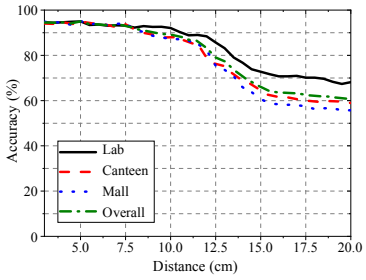


Fig. 14. Accuracy under different distances between smartphones and vocal tract.

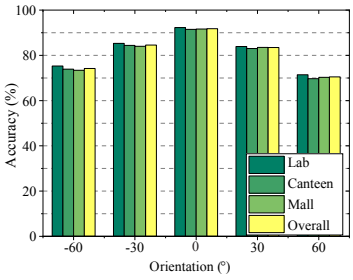


Fig. 15. Accuracy under different orientations between smartphones and vocal tract.

i.e., TDoA-based *VoiceLive*, achieves better performance in attack resistance. For the three attacks, *VoiceLive* can achieves 3.5%, 4.3% and 1.5% false accept rates, which even outperforms the proposed *VocalLock*. However, considering its strict requirements, *VocalLock* could be a more user-friendly option for user authentication. As for voiceprint-based user authentication (i.e., WeChat), the false accept rates under eavesdropping and replay as well as voice synthesis attacks are both above 60%, which are significantly higher than that of *VocalLock*. This is because WeChat only utilizes the physiological characteristics underlying the voices during speaking without liveness verification. Hence, a successful attack could be performed as long as an adversary obtains the voice samples of a legitimate user. All the results demonstrate that *VocalLock* outperforms existing mouth movement-based and voiceprint-based user authentications on attack resistance.

7.4 Performance on User Experience

A large false reject rate indicates a high probability that a legitimate user is rejected by the system in the login stage, which significantly degrades the user experience. Fig. 11 shows the false reject rate of *VocalLock* under three different smartphones in three different environments. We can see that the overall false reject rates are all less than 2% under the three smartphones. In the lab environment, *VocalLock* can further achieve a false reject rate below 1% under the three smartphones. Moreover, it can be observed the false reject rates in the canteen and mall are a little higher than that in the lab. But the difference between false reject rates in the lab and other two environments is less than 3%, which is not significant. These results demonstrate *VocalLock* seldom rejects the login request from a legitimate user and thus achieves a user-friendly experience.

We also evaluate the user experience through the times that a user speaks a passphrase to the system for a successful login, i.e., the speaking times for a successful login. During the login stage, a user may fail to have access to the system within speaking a passphrase once. Hence, the speaking times for successful login exhibits the user experience during using *VocalLock*. Fig. 12 shows the CDF of speaking times for successful login under three different smartphones. We can see that 82.7% volunteers can successfully login the system with only speaking the passphrase once. Overall, over 95% volunteers could successfully login the system within speaking a passphrase 3 times. Such a speaking time for successful login is acceptable for users. Also, CDF of speaking times for successful login under different smartphones exhibits subtle differences. All these results indicate that *VocalLock* can achieve a user-friendly experience.

As a service for mobile devices, the energy consumption of *VocalLock* also affects the user experience. Fig. 13 shows the energy consumption of *VocalLock* in the register and login stages. We can see that in the login stage, the average energy consumption of *VocalLock* on the three devices is 24.8mAh, which is close to that under the medium brightness of screens [7]. This result indicates *VocalLock* would not induce significant power

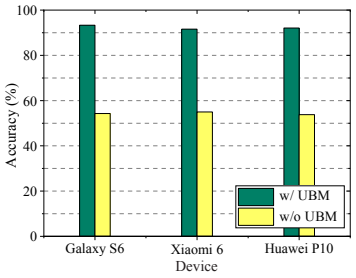


Fig. 16. Accuracy of *VocalLock* with and without UBM.

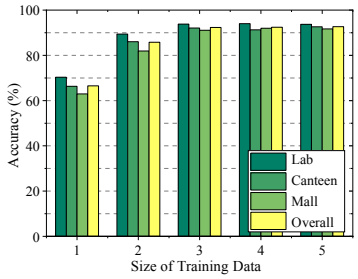


Fig. 17. Accuracy under different sizes of training data.

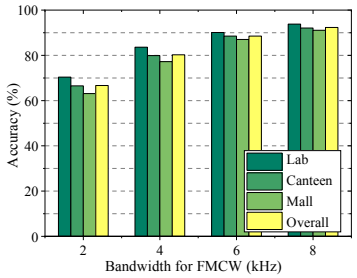


Fig. 18. Accuracy under different bandwidths for FMCW.

overhead for mobile devices during user authentication. But in the register stage, the average energy consumption dramatically increases to 106mAh. Fortunately, the register stage only runs once when a new user registers to the system. Therefore, *VocalLock* achieves a user-friendly experience in terms of energy consumption.

7.5 Impact of Relative Position between Smartphone and Vocal Tract

Acoustic signals attenuate as propagating through a distance. Hence, the distance between a smartphone and a user's vocal tract has a certain impact on the performance of *VocalLock*. We enable smartphones to measure the distance between users' vocal tract and the microphone through Time of Arrival (ToA). Fig. 14 shows the accuracy of *VocalLock* on user authentication under different distances between smartphones and vocal tract in the three environments. We can observe that the accuracy decreases with the increase of distance between smartphone and vocal tract. This is because as the distance increases, the signal strength of received acoustic signals significantly decreases, which induces a low Signal-Noise-Ratio (SNR) and reduces the resolution of extracted features for user authentication. However, when the distance is within 8cm, *VocalLock* achieves an accuracy over 90% in all three different environments. The overall accuracy of *VocalLock* is 91.8% under the distance of 8cm. If a user is in a quiet environment (i.e., lab in our experiments), the distance can be extended to 10cm for an accuracy over 90% of *VocalLock*. Such distances are usually natural for a user to use *VocalLock* on smartphones [9].

Except for the distance, the orientation (i.e., the angle) between the smartphone and vocal tract also affects the performance of *VocalLock*. The angle when the smartphone is directly towards the lip is defined as the orientation of 0°. Along with this, the angle θ between the smartphone-lip connection and 0° line is defined as the orientation between the smartphone and vocal tract. And the angle is positive when the smartphone-lip connection is above the 0° line, and vice versus. In this experiment, we fix the distance between the smartphone and vocal tract as 10cm for variable control. Fig. 15 shows the accuracy of *VocalLock* under different orientations between smartphones and the vocal tract. We can find that the accuracy decreases when the orientation is away from 0°. This result indicates the extracted features under various orientations are different from each other, inducing performance degradation on user authentication. But when the orientation is within $[-30°, 30°]$, the accuracies are still above 80%. Considering the slight shift in orientation in most cases, such performance is acceptable for users.

7.6 Impact of Universal Background Model (UBM)

VocalLock employs the GMM-UBM model to construct the user authentication model, which intrinsically belongs to GMM-based approaches. Hence, we perform an experiment to evaluate the performance of *VocalLock* with GMM-UBM and typical GMM (i.e., without UBM). Fig. 16 shows the accuracy of *VocalLock* with and without UBM under different devices. It can be seen that *VocalLock* with UBM achieves an average accuracy of 92.3%,

while that without UBM only achieves 54.4%. This is because the UBM is pre-trained using a large number of samples from different persons, e.g., trained with VoxCeleb dataset in this experiment, which enables UBM with the basic knowledge of individual uniqueness under the vocal tract. Calibrating the GMM for each legitimate user on the basis such knowledge significantly helps to improve the authentication performance.

7.7 Impact of Training Data Size

The size of training data is the number of a user's speaking times in the register stage. More times of a user's speaking provide more data for model training, which leads to higher accuracy on user authentication. However, too many speaking times significantly degrade user experience in the register stage. Fig. 17 shows the accuracy of *VocalLock* under different sizes of training data in the three different environments. We can see that as the size of training data increases, the accuracy first increases and then remains stable. When a user speaks over 3 times in the register stage, *VocalLock* can achieve an accuracy over 90% in all the three environments. More speaking times do not contribute to an improvement in accuracy of *VocalLock*. Even speaking two times in the register stage, *VocalLock* achieves an accuracy of 85.8%. This is because *VocalLock* constructs the user authentication model based on GMM-UBM, which only requires users to provide a few training data samples [38]. Usually, existing voiceprint-based user authentication requires a user to speak a passphrase two times [50]. Therefore, *VocalLock* can achieve a user-friendly experience, while remaining secure as a user authentication for mobile devices.

7.8 Impact of Bandwidth for FMCW

As mentioned in Section 5.1, the bandwidth of transmitted chirp signals for FMCW affects the resolution of sensing the vocal tract. A fine-grained sensing of vocal tract leads to accurate user authentication of *VocalLock*. To validate it, we evaluate the performance of *VocalLock* on user authentication under different bandwidths for FMCW, i.e., $18kHz{-}16kHz{=}2kHz$, $20kHz{-}16kHz{=}4kHz$, $22kHz{-}16kHz{=}6kHz$, and $24kHz{-}16kHz{=}8kHz$. Fig. 18 shows the accuracy of *VocalLock* on user authentication under different bandwidths in the three environments. We can see that the overall accuracy of *VocalLock* increases from 66.7% to 92.3% with the increase of bandwidth from $2kHz$ to $8kHz$, which is consistent with the analysis in Section 5.1. It can be also observed that the overall accuracy is still over 80% when the bandwidth is larger than $4kHz$ (i.e., $[16, 20]kHz$). This result indicates that *VocalLock* can achieve an acceptable performance for the smartphones with a speaker that may be not capable to transmit an acoustic signal as high as $24kHz$.

8 LIMITATIONS AND FUTURE DIRECTIONS

VocalLock is only a research prototype, instead of a mature industry product, thus remaining several limitations:

i). The authentication task of *VocalLock* is to serve as the auxiliary channel for two-factor authentications. *VocalLock* explores both the static shape and dynamic movements of vocal tract as biometrics to realize the user authentication, which has the potential to serve as a primary authentication factor theoretically. However, similar to other behavioral characteristics-based solutions, the vocal tract, especially its dynamic movement, is easily affected by habit change, disease infection, etc., which induces unstable biometrics and leads to authentication performance degradation. Hence, it is appropriate to treat *VocalLock* as auxiliary channel for two-factor authentications. Moreover, considering the same speaking behavior with voiceprint-based authentication, *VocalLock* is qualified as a natural and secure auxiliary channel. We will explore the connection between voiceprint and vocal tract to realize a two-factor authentication as one of our future directions.

ii). The relative position between a smartphone and vocal tract is limited in the authentication scenario. To ensure *VocalLock* can accurately sense the vocal tract for user authentication, a user should keep the relative distance less than $10cm$, and the orientation within angles of $[-30, 30]^\circ$, as demonstrated in Section 7.5. The distance restriction may hinder the extension of *VocalLock* on gradually-prevalent smart speakers (e.g., Amazon

Echo [2] and Google Home [16]) in smart homes. However, for the user authentication in smartphones, such a distance still remains satisfactory [9]. On the other hand, the orientation restriction is intrinsically caused by the strong directionality from the FMCW technique employed in *VocalLock*. But compared with VoiceLive, *VocalLock* partially releases the strong relative position restriction. Potential further solutions may lie on utilizing multiple speakers widely integrated on smartphones to extend the sensing orientation. We leave the extension of *VocalLock* for far-field and wide-angle scenarios in our future work.

iii). The register stage of *VocalLock* requires users to provide data samples in quiet environments for training the EDNN model. This is because the EDNN model requires both the knowledge of vocal tract and speech voices to construct the correspondence between these two features for feature-based transfer learning. Hence, the clean speech voices (i.e., without inferred noises) need to be captured for the model training. Fortunately, the register stage is a one-off data collection process. In the more frequent login stage, *VocalLock* can be used in various environments regardless of the noises, which does not degrade the user experience.

iv). The proposed *VocalLock* may be vulnerable to direct attacks. An adversary can pre-implant a malicious APP in a user's smartphone. The malicious APP compromises the user's smartphone to invoke the microphone for eavesdropping FDS features of the user's vocal tract in an inconspicuous manner. Then, based on the eavesdropped features, the adversary can directly launch the replay attacks to the system. The straightforward countermeasure is that users should be aware of and carefully control the microphone permission, i.e., the permission of microphones should not be granted for APPs from unknown sources.

v). The stability of *VocalLock* over time remains unknown. This paper has demonstrated that *VocalLock* can accurately distinguish 25 different persons in real authentication scenarios. However, the data is collected within 1 month for each environment only. Different from the inborn biometrics (e.g., fingerprint), the individual uniqueness derived from vocal-tract behaviors may vary due to the habit change as time goes on. This open issue also remains to be verified for existing mouth movement-based user authentication studies [26, 27, 46, 58, 59]. It is necessary to perform long-term experiments in real authentication scenarios for further validations. Potential solutions may lie on involving temporal feature-based techniques, such as Hidden Markov Model (HMM) and Recurrent Neural Network (RNN), to enable the capability of long-term feature learning. We leave the long-term validation and relative solution design in our future work.

9 CONCLUSIONS

In this paper, we propose a user authentication system, *VocalLock*, which characterizes the whole vocal tract leveraging acoustic signals to identify different individuals. *VocalLock* first extracts unique features of the whole vocal tract through FMCW technique, which is immune to mimic and replay attacks. Then, *VocalLock* constructs a passphrase-independent model to authenticate user identities. To construct such a model, we first propose an EDNN to transfer the features of vocal tract to that of speech voices. Based on transferred features, *VocalLock* further employs the speech voice-based GMM-UBM to construct the passphrase-independent authentication model. Experiments under 75 volunteers in three real environments demonstrate that *VocalLock* can accurately authenticate user identity in a passphrase-independent manner, and resist replay attack as well as mimic attack.

ACKNOWLEDGMENTS

This research is supported in part by NSFC (No. 61772338) and National Key R&D Program of China (No. 2018YFC1900700). We would like to sincerely thank the anonymous editors and reviewer for their helpful suggestions and comments to improve the quality of this paper.

REFERENCES

[1] Jont B Allen and Lawrence R Rabiner. 1977. A unified approach to short-time Fourier analysis and synthesis. *Proc. IEEE* 65, 11 (1977), 1558–1564.

[2] Amazon. 2019. Echo & Alexa - Amazon Device. [Online]. Available: https://www.amazon.com. (2019).
[3] Apple. 2019. iPhone XS - FaceID - Apple. [Online]. Available: https://www.apple.com/iphone-xs/face-id/. (2019).
[4] L. Benedikt, D. Cosker, P. L. Rosin, and D. Marshall. 2010. Assessing the Uniqueness and Permanence of Facial Actions for Use in Biometric Applications. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 40, 3 (2010), 449–460.
[5] C. BYU. 2020. Word frequency: based on 450 million word coca corpus. [Online]. Available: https://www.wordfrequency.info. (2020).
[6] J. P. Campbell. 1997. Speaker recognition: a tutorial. *Proc. IEEE* 85, 9 (1997), 1437–1462.
[7] Aaron Carroll and Gernot Heiser. 2010. An Analysis of Power Consumption in a Smartphone. In *Proc. USENIX ATC*. Boston, MA, USA, 21:1–21:14.
[8] Mingshi Chen, Panlong Yang, Jie Xiong, Maotian Zhang, Youngki Lee, Chaocan Xiang, and Chang Tian. 2019. Your Table Can Be an Input Panel: Acoustic-based Device-Free Interaction Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 1 (2019), 3:1–3:21.
[9] S. Chen, K. Ren, S. Piao, C. Wang, Q. Wang, J. Weng, L. Su, and A. Mohaisen. 2017. You Can Hear But You Cannot Steal: Defending Against Voice Impersonation Attacks on Smartphones. In *Proc. IEEE ICDCS*. 183–195.
[10] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. 2011. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing* 19, 4 (2011), 788–798.
[11] Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* 39, 1 (1977), 1–22.
[12] G. R. Doddington. 1985. Speaker recognition—Identifying people by their voices. *Proc. IEEE* 73, 11 (1985), 1651–1664.
[13] J.-L. Gauvain and Chin-Hui Lee. 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing* 2, 2 (1994), 291–298.
[14] H. Gish and M. Schmidt. 1994. Text-independent speaker identification. *IEEE Signal Processing Magazine* 11, 4 (Oct 1994), 18–32.
[15] Xavier Glorot, Antoine Bordes, Yoshua Bengio, Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2012. Deep Sparse Rectifier Neural Networks. In *Proc. AISTATS'12*. La Palma, Canary Islands, 315–323.
[16] Google. 2019. Google Home - Smart Speaker & Home Assistant. [Online]. Available: https://store.google.com/us/product/google_home. (2019).
[17] Google. 2019. Google Smart Lock. [Online]. Available: https://get.google.com/smartlock/. (2019).
[18] Diego Gragnaniello, Giovanni Poggi, Carlo Sansone, and Luisa Verdoliva. 2015. Local contrast phase descriptor for fingerprint liveness detection. *Pattern Recognition* 48, 4 (2015), 1050–1058.
[19] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).
[20] Cesar Iovescu and Sandeep Rao. 2017. *The fundamentals of millimeter wave sensors.* Technical Report. Texas Instruments. http://www.ti.com/lit/wp/spyy005/spyy005.pdf
[21] Artur Janicki, Federico Alegre, and Nicholas Evans. 2016. An assessment of automatic speaker verification vulnerabilities to replay spoofing attacks. *Security and Communication Networks* 9, 15 (2016), 3030–3044.
[22] Mark Keith, Benjamin Shao, and Paul John Steinbart. 2007. The usability of passphrases for authentication: An empirical field study. *International journal of human-computer studies* 65, 1 (2007), 17–28.
[23] HJ Landau. 1967. Sampling, data transmission, and the Nyquist rate. *Proc. IEEE* 55, 10 (1967), 1701–1706.
[24] Yun Lei, Nicolas Scheffer, Luciana Ferrer, and Mitchell McLaren. 2014. A novel scheme for speaker recognition using a phonetically-aware deep neural network. In *Proc. IEEE ICASSP*. Florence, Italy, 1695–1699.
[25] Mengyuan Li, Yan Meng, Junyi Liu, Haojin Zhu, Xiaohui Liang, Yao Liu, and Na Ruan. 2016. When CSI Meets Public WiFi: Inferring Your Mobile Phone Password via WiFi Signals. In *Proc. ACM CCS*. Vienna, Austria, 1068–1079.
[26] Li Lu, Jiadi Yu, Yingying Chen, Hongbo Liu, Yanmin Zhu, Linghe Kong, and Minglu Li. 2019. Lip Reading-Based User Authentication Through Acoustic Sensing on Smartphones. *IEEE/ACM Transactions on Networking* 27, 1 (2019), 447–460.
[27] Li Lu, Jiadi Yu, Yingying Chen, Hongbo Liu, Yanmin Zhu, Yunfei Liu, and Minglu Li. 2018. LipPass: Lip Reading-based User Authentication on Smartphones Leveraging Acoustic Signals. In *Proc. IEEE INFOCOM*. Honolulu, HI, USA, 1466–1474.
[28] Li Lu, Jiadi Yu, Yingying Chen, Yanmin Zhu, Minglu Li, and Xiangyu Xu. 2019. I3: Sensing Scrolling Human-Computer Interactions for Intelligent Interest Inference on Smartphones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3 (2019), 97:1–97:22.
[29] Li Lu, Jiadi Yu, Yingying Chen, Yanmin Zhu, Xiangyu Xu, Guangtao Xue, and Minglu Li. 2019. KeyListener: Inferring Keystrokes on QWERTY Keyboard of Touch Screen through Acoustic Signals. In *Proc. IEEE INFOCOM*. Paris, France, 1–9.
[30] Wenguang Mao, Jian He, and Lili Qiu. 2016. CAT: high-precision acoustic motion tracking. In *Proc. ACM MobiCom*. New York City, NY, USA, 69–81.
[31] Wenguang Mao, Mei Wang, and Lili Qiu. 2018. AIM: Acoustic Imaging on a Mobile. In *Proc. ACM MobiSys*. Munich, Germany, 468–481.
[32] Pavel Matějka, Ondřej Glembek, Fabio Castaldo, Md Jahangir Alam, Oldřich Plchot, Patrick Kenny, Lukáš Burget, and Jan Černocky. 2011. Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification. In *Proc. IEEE ICASSP*. Prague, Czech Republic, 4828–4831.

[33] Dibya Mukhopadhyay, Maliheh Shirvanian, and Nitesh Saxena. 2015. All Your Voices are Belong to Us: Stealing Voices to Fool Humans and Machines. In *Proc. ESORICS*. Springer, Vienna, Austria, 599–621.
[34] A. Nagrani, J. S. Chung, and A. Zisserman. 2017. VoxCeleb: a large-scale speaker identification dataset. In *Proc. ISCA INTERSPEECH*. Stockholm, Sweden, 2616–2620.
[35] Swadhin Pradhan, Ghufran Baig, Wenguang Mao, Lili Qiu, Guohai Chen, and Bo Yang. 2018. Smartphone-based Acoustic Indoor Space Mapping. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 2, Article 75 (2018), 26 pages.
[36] Swadhin Pradhan, Wei Sun, Ghufran Baig, and Lili Qiu. 2019. Combating Replay Attacks Against Voice Assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3 (2019), 100:1–100:26.
[37] K. Qian, C. Wu, F. Xiao, Y. Zheng, Y. Zhang, Z. Yang, and Y. Liu. 2018. Acousticcardiogram: Monitoring Heartbeats using Acoustic Signals on Smart Devices. In *Proc. IEEE INFOCOM*. Honolulu, HI, USA, 1574–1582.
[38] Douglas A. Reynolds. 1997. Comparison of Background Normalization Methods for Text-Independent Speaker Verification. In *Proc. ISCA EUROSPEECH*. Rhodes, Greece, 963–966.
[39] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. 2000. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing* 10, 1 (2000), 19–41.
[40] Samsung. 2017. Iris recognition on Galaxy S8. [Online]. Available: https://www.samsung.com/au/iris/. (2017).
[41] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proc. IEEE CVPR*. Boston, MA, USA, 815–823.
[42] Wei Shang and Maryhelen Stevenson. 2010. Score normalization in playback attack detection. In *Proc. IEEE ICASSP*. Dallas, Texas, USA, 1678–1681.
[43] Sigurdur Sigurdsson, Kaare Brandt Petersen, and Tue Lehn-Schiøler. 2006. Mel Frequency Cepstral Coefficients: An Evaluation of Robustness of MP3 Encoded Music. In *Proc. ISMIR*. Victoria, Canada, 286–289.
[44] Merrill Ivan Skolnik. 1970. *Radar handbook*. McGraw-Hill, Incorporated, New York, NY, USA.
[45] Jiayao Tan, Cam-Tu Nguyen, and Xiaoliang Wang. 2017. SilentTalk: Lip reading through ultrasonic sensing on mobile phones. In *Proceedings of IEEE INFOCOM*. IEEE, Atlanta, GA, USA, 1–9.
[46] Jiayao Tan, Xiaoliang Wang, Cam-Tu Nguyen, and Yu Shi. 2018. SilentKey: A New Authentication Framework Through Ultrasonic-based Lip Reading. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1 (2018), 36:1–36:18.
[47] Emanuel von Zezschwitz, Paul Dunphy, and Alexander De Luca. 2013. Patterns in the Wild: A Field Study of the Usability of Pattern and Pin-based Authentication on Mobile Devices. In *Proc. ACM MobileHCI*. Munich, Germany, 261–270.
[48] Tianben Wang, Daqing Zhang, Yuanqing Zheng, Tao Gu, Xingshe Zhou, and Bernadette Dorizzi. 2018. C-FMCW Based Contactless Respiration Detection Using Acoustic Signal. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4 (2018), 170:1–170:20.
[49] Zhi-Feng Wang, Gang Wei, and Qian-Hua He. 2011. Channel pattern noise based playback attack detection algorithm for speaker recognition. In *Proc. IEEE ICMLC*. Guilin, China, 1708–1713.
[50] Wechat. 2015. Voiceprint: The New Wechat Password. [Online]. Available: https://blog.wechat.com/2015/05/21/voiceprint-the-new-wechat-password/. (2015).
[51] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li. 2015. Spoofing and countermeasures for speaker verification: A survey. *Speech Communication* 66 (2015), 130–153.
[52] Xiangyu Xu, Hang Gao, Jiadi Yu, Yingying Chen, Yanmin Zhu, Guangtao Xue, and Minglu Li. 2017. ER: Early recognition of inattentive driving leveraging audio devices on smartphones. In *Proc. IEEE INFOCOM*. Atlanta, GA, USA, 1–9.
[53] Xiangyu Xu, Jiadi Yu, Yingying Chen, Yanmin Zhu, Linghe Kong, and Minglu Li. 2019. BreathListener: Fine-grained Breathing Monitoring in Driving Environments Utilizing Acoustic Signals. In *Proc. ACM MobiSys*. Seoul, South Korea, 1–13.
[54] Chen Yan, Yan Long, Xiaoyu Ji, and Wenyuan Xu. 2019. The Catcher in the Field: A Fieldprint Based Spoofing Detection for Text-Independent Speaker Verification. In *Proc. ACM CCS*. London, United Kingdom, 1215–1229.
[55] J. Yan, A. Blackwell, R. Anderson, and A. Grant. 2004. Password memorability and security: empirical results. *IEEE Security Privacy* 2, 5 (2004), 25–31.
[56] Sangki Yun, Yi-Chao Chen, Huihuang Zheng, Lili Qiu, and Wenguang Mao. 2017. Strata: Fine-grained acoustic-based device-free tracking. In *Proc. ACM MobiSys*. Niagara Falls, NY, USA, 15–28.
[57] Matthew D Zeiler, Graham W Taylor, Rob Fergus, et al. 2011. Adaptive deconvolutional networks for mid and high level feature learning. In *Proc. IEEE ICCV*. Barcelona, Spain, 2018–2025.
[58] Linghan Zhang, Sheng Tan, and Jie Yang. 2017. Hearing Your Voice is Not Enough: An Articulatory Gesture Based Liveness Detection for Voice Authentication. In *Proc. ACM CCS*. Dallas, TX, USA, 57–71.
[59] Linghan Zhang, Sheng Tan, Jie Yang, and Yingying Chen. 2016. Voicelive: A phoneme localization based liveness detection for voice authentication on smartphones. In *Proc. ACM CCS*. Vienna, Austria, 1080–1091.
[60] Man Zhou, Qian Wang, Jingxiao Yang, Qi Li, Feng Xiao, Zhibo Wang, and Xiaofeng Chen. 2018. PatternListener: Cracking Android Pattern Lock Using Acoustic Signals. In *Proc. ACM CCS*. Toronto, Canada, 1775–1787.