LVID: A Multimodal Biometrics Authentication System on Smartphones

Libing Wu, *Member, IEEE*, Jingxiao Yang, Man Zhou, Yanjiao Chen, *Member, IEEE*, and Qian Wang, *Senior Member, IEEE*

***Abstract*— Voice authentication is becoming increasingly popular, which offers potential benefits over knowledge and possession based authentication methods. Meanwhile, the unique features of lip movements during speaking have been proved to be useful for authentication. However, the unimodal biometric authentication systems based on either voice or lip movements have certain limitations. Voice authentication systems are prone to spoofing attacks and suffer from serious performance degradation in noisy environments. Lip movements authentication systems are unstable and are sensitive to the user's physical and psychological conditions. In this paper, we propose and implement LVID, a multimodal biometrics authentication system on smartphones, which resolves the defects of the original systems by combining the advantages of lip movements and voice. LVID simultaneously captures these two biometrics with the built-in audio devices on smartphones and fuses them at the data level. The reliable and effective features are then extracted from the fused data for authentication. LVID is practical as it requires neither cumbersome operations nor additional hardwares but only a speaker and a microphone that are commonly available on smartphones. Our experimental results with 104 participants show that LVID can achieve 95% accuracy for user authentication, and 93.47% of the attacks can be detected. It is also verified that LVID works well with different smartphones and is robust to different smartphone positions.**

***Index Terms*— Authentication, multimodal biometrics, acoustic sensing.**

Manuscript received January 22, 2019; revised May 20, 2019 and August 17, 2019; accepted September 16, 2019. Date of publication September 27, 2019; date of current version January 16, 2020. The work of Q. Wang was supported in part by the NSFC under Grant 61822207 and Grant U1636219, in part by the Equipment Pre-Research Joint Fund of Ministry of Education of China (Youth Talent) under Grant 6141A02033327, and in part by the Outstanding Youth Foundation of Hubei Province under Grant 2017CFA047. The work of Y. Chen was supported in part by the National Natural Science Foundation of China under Grant 61972296 and Grant 61702380, in part by the Wuhan Advanced Application Project under Grant 2019010701011419, and in part by the Hubei Provincial Technological Innovation Special Funding Major Projects under Grant 2017AAA125. The work of L. Wu was supported in part by the National Natural Science Foundation of China under Grant 61772377 and Grant 61572370, in part by the Natural Science Foundation of Hubei Province of China under Grant 2017CFA007, and in part by the Science and Technology Planning Project of Shenzhen under Grant JCYJ20170818112550194. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. William R. Schwartz. *(Corresponding author: Yanjiao Chen.)*

L. Wu, J. Yang, and Y. Chen are with the School of Computer Science, Wuhan University, Wuhan 430072, China, and also with the State Key Laboratory of Cryptology, Beijing 100878, China (e-mail: wu@whu.edu.cn; yangjingxiao@whu.edu.cn; chenyanjiao@whu.edu.cn).

M. Zhou and Q. Wang are with the School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China, and also with the State Key Laboratory of Cryptology, Beijing 100878, China (e-mail: zhouman@whu.edu.cn; qianwang@whu.edu.cn).

Digital Object Identifier 10.1109/TIFS.2019.2944058

I. Introduction

BIOMETRICS is a promising way to replace knowledge and possession based methods for user authentication. Compared with other biometrics technologies (e.g., face [1], fingerprint [2] and gait [3]), voice biometrics is widely adopted due to the pervasiveness and accessibility of voice. In recent years, voice authentication has matured as a low-cost and reliable method for authentication in a wide range of applications such as access control, forensics and law enforcement [4], [5].

With the increasing popularity of mobile devices, voice authentication has been integrated into mobile systems and mobile applications. For example, Google provides voice biometrics for screen unlock of the Android operating system [6]. Baidu, one of China's largest web service providers, integrates voice-unlock in the operating system of their smartphones [7]. WeChat allows users to log in through their "Voiceprint" [8] generated from their voice passphrase. It is foreseeable that voice authentication will be more popular in future mobile markets.

However, a growing body of researches have revealed that voice authenticating systems are susceptible to spoofing attacks [9], [10], where an imposter attempts to spoof the authentication system via a prerecorded or synthetic voice sample of victims, some attacks can also be conducted without the victim's awareness [11]. Several technologies have been proposed to defend against spoofing attacks and liveness detection is the most common way. For example, Wang *et al.* [12] proposed a practical and effective anti-spoofing system for voice authentication based on pop noise. Zhang *et al.* [13] designed a liveness detection system based on articulatory gesture for voice authentication. Chen *et al.* [14] explored the magnetic field emitted from loudspeakers to detect machine-based voice impersonation attacks. Meng *et al.* [15] proposed a two-factor liveness detection system by analyzing the correlation between voice samples and wireless signals. Unfortunately, besides spoofing attacks, voice authentication systems are also vulnerable in noisy environments [16], such as factories and markets.

Numerous studies have shown that lip movements can be used for user authentication [17]–[19]. Recently, three state-of-art acoustic-based lip movements sensing schemes [20]–[22] were proposed to use the smartphone as an ultrasonic sonar to sense the unique lip movements patterns for user authentication, which can defend against spoofing attacks and are robust to noisy environments. LipPass [20], [22] chose time

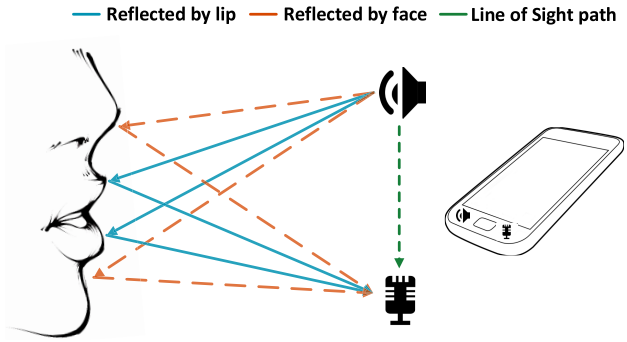


Fig. 1. Sensing lip movements with acoustic signals generated by smartphones.

sequential Doppler shift as the features of lip movements and leveraged Short-Time Fourier Transform (STFT) to measure the Doppler shift. However, only a coarse-grained estimation of lip movements is provided. SilentKey [21] conquers this limitation with the signal envelope, but failed to eliminate the interference of multipath and unpredictable phase, thus only reaches 86.7% True Negative Rate (TNR) and 70% True Positive Rate (TPR) with 5 training samples. Moreover, the unimodal authentication based on lip movements is not stable enough since lip movements are sensitive to the user's physical and psychological conditions.

Inspired by the fact that lip movements can be captured with the built-in audio devices on smartphones as the voice, we deem it promising to build a multimodal biometrics authentication system by incorporating the characteristics of both lip movements and voice, which benefits from the advantages of these two biometrics (e.g., voice authentication is matured and stable, and lip movements authentication is inimitable and free from ambient noise) and avoids the shortcomings of the original unimodal systems. However, we are facing several challenges. First, we need to achieve fine-grained estimation of the subtle lip movements with acoustic signals. Second, the coordination of these two biometrics is necessary for accurate authentication based on the fused biometrics. Finally, the solution should be lightweight and computationally efficient for smartphones.

In this paper, we propose LVID, a multimodal biometrics authentication system based on lip movements and voice. Users can hold the smartphones in their habitual ways and speak to the smartphones when authenticating with LVID. LVID is able to achieve the best of both worlds: it is highly accurate in user authentication and can effectively defend against different types of spoofing attacks without requiring cumbersome user operations or additional hardwares. As shown in Fig. 1, lip movements can be sensed with acoustic signals generated by smartphones. High-frequency acoustic signals are transmitted by the speaker, and the signals reflected by moving lips are received by the microphone to be processed to estimate the lip movements. LVID exploits the mobile audio hardware advances to sense the lip movement patterns when a user speaks the passphrase to the smartphone. More specifically, the high definition audio capability supported by smartphones allows the built-in audio devices to play and record acoustic signals at frequencies higher than 18 kHz. Such high frequencies are inaudible to most human [23] and can be easily distinguished from the user's voice. LVID leverages the speaker to emit high-frequency acoustic tones higher than 18 kHz and the microphone to listen to the signals reflected by the moving lip when a user conducts authentication. In the mean time, the user's voice is captured by the microphone as well. The recorded acoustic signals will be processed by a remote server to authenticate users. LVID attains a fine-grained lip movements estimation from the reflected signals and fuses it with the preprocessed voice sample at the data level [24]. Finally, LVID extracts user-specific features from the fused data to compare with the features stored in the user model for authentication. Our extensive experiments demonstrate that LVID is reliable and efficient for user authentication even in noisy environments and can defend against different types of spoofing attacks with a very high accuracy. The main contributions of our work are summarized as follows:

- We propose LVID, a multimodal biometrics system on smartphones which leverages lip movements and voice for authentication. This system is practical without requiring cumbersome user operations or additional hardwares and can defend against two kinds of attacks.
- We develop a novel mechanism to obtain a fine-grained estimation of the lip movements by eliminating the interference of multipath and system delay. The estimation characterizes both the absolute movement patterns of user's lips and the relative movement patterns of the upper lip and the lower lip.
- We conquer the incompatibility issue by capturing the biometrics of both lip movements and voice with acoustic signals and processing them at the time domain, which makes it possible to fuse the two biometrics at the data level to realize an effective multimodal biometrics system.
- We build the prototype of LVID on smartphones and conduct comprehensive evaluations in real environments. The results with 104 participants show that LVID can achieve 95% accuracy for user authentication and 93.47% of the attacks can be detected. It is also demonstrated that LVID works well with different smartphones and is robust to different smartphone positions.
- We publish a free and open dataset with a total size of 55GB on Google Drive[1], which consists of more than 140,000 pieces of raw data from the 104 volunteers with their consent. These acoustic samples contain both voice information and lip movement information, captured with different smartphones in different environments, distances and positions. The dataset may enable the research community to further explore potential interesting topics in related fields.

The rest of the paper is organized as follows. We review related works in Section II and introduce the preliminaries in Section III. The approach of sensing lip movements with acoustic signals is depicted in Section IV. Section V presents the details of LVID and the performance evaluation is given

[1]https://drive.google.com/open?id=1sTXkBpbu1SkaaXNKmilS5bsyVvQ3f-upi

in Section VI. We finally discuss this work in Section VII and conclude the paper in Section VIII.

II. Related Work

A. Unimodal Authentication System Based on Lip Movements

Existing lip movement sensing approaches mainly exploit video for sample capturing [17]–[19], which first collect facial images and then extract the lip movement characteristics from the image sequences for authentication. Unfortunately, vision-based sensing methods require a specific viewing angle and are sensitive to lighting conditions. In addition, they are also susceptible to spoofing attacks. There are three recent works that use acoustic signals instead of videos for lip movement sensing. Lu *et al.* [20], [22] used the smartphone as a Doppler sonar to sense lip movements, but it is difficult to obtain the fine-grained estimation of lip movements with Doppler shift. Tan *et al.* [21] design the signal envelope to extract lip movement characteristics. Nevertheless, the signal envelope suffers from the interference of multipath and system delay. Moreover, the unimodal authentication based on lip movements is unstable since lip movements are sensitive to the user's physical and psychological conditions. Apart from sensing lip movements, acoustic sensing has been applied to two other main scenarios, position tracking [25], [26] and gesture recognition [27]. However, different from other objects, lips are flexible and lip movements are non-rigid where different parts of lips exhibit different motion patterns during speaking. Moreover, lip movements are mostly less than one centimeter. Therefore, existing acoustic sensing technologies are not suitable for fine-grained lip movements sensing.

B. Liveness Detection in Voice Authentication

Voice authentication systems are vulnerable to spoofing attacks [9]–[11] and liveness detection is a common way to defend against spoofing attacks in voice authentication systems. For instance, Wang *et al.* [12] use pop noise produced by the user breathing while speaking close to the microphones to build a practical and efficient anti-spoofing system for voice authentication. Chen *et al.* [14] develop a liveness detection system by measuring the magnetic field emitted from loudspeakers. However, it requires the user to speak the passphrase while moving the smartphone with predefined trajectory around the sound source. VoiceLive [28] proposes to capture the time-difference-of-arrival (TDoA) changes from a sequence of phoneme sounds to the two microphones of the smartphone which does not exist under replay attacks, this requires users to hold the smartphone at a specific position. Zhang *et al.* [13] design a liveness detection system based on articulatory gestures for voice authentication. Unfortunately, the extent of articulatory movements affects the effectiveness of this countermeasure. Meng *et al.* [15] use the correlation between voice samples and wireless signals for liveness detection, but the WiFi signals are easily influenced by other moving objects, e.g., people nearby.

C. Multimodal Biometrics Authentication

Multimodal biometric authentication systems can achieve significant accuracy gains over unimodal biometric authentication systems. Ichino *et al.* [29] exploite lip movements and voice to build a multimodal biometric system. While unimodal system based on lip movements or voice can achieve 89.9% and 68.2% accuracy respectively, the accuracy of the multimodal system can reach 93.6%. Çetingül *et al.* [30] propose a multimodal biometrics system using lip motion, lip texture and audio. Compared to the unimodal systems with the error rate of 2.4%, 1.7% and 5.2% respectively, the multimodal system's error rate is only 0.4%. Other multimodal biometric systems [31], [32] exhibit the same benefits.

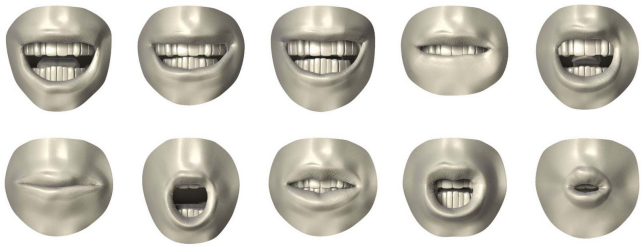


Fig. 2. The typical lip shapes when pronounciation [33].

III. Preliminaries

A. Lip Movements for Authentication

Lip movements during speaking involve complicated coordination of muscles and bones. The upper and lower lips will form particular shapes when producing different phonemes in speech as shown in Fig. 2. The shape of lip changes with pronouncing, resulting in lip movements. The structure of the muscles and bones related to lips varies from person to person, which leads to subtle individual differences in lip shapes. Moreover, the unique speaking ways of individuals also yield different lip movements patterns. As a result, it is possible to employ the special features of lip movements as the biometrics for user authentication.

B. The Production of Human Voice

When we exhale, the air is pressed from the lung through the trachea to create an airstream. The airstream provides the energy for the vocal cords to produce sounds. The larynx which contains two vocal cords sits on top of the trachea. When we produce voice, the airstream passes between the two vocal cords that have come together. The vocal cords are set into vibration by the passing airstream. The frequency of the vibration depends on the pitch of the sound, and the pitch is determined by the length and tension of the vocal cords, which features individual difference. The audio produced by the vocal cords themselves sounds like simple buzzing. The organs above the cords, including the throat, nose, and mouth, constitute the resonator system. The buzzing sound created by vocal cord vibration is shaped by the resonator to produce the unique human voice, which also suggests that voice production is closely related to lip movements.

C. System and Attack Model

LVID uses the biometrics of the pronunciation to authenticate users. Both lip movement patterns and voice features are associated with the user's speech, thus LVID is a text-dependent system. A typical text-dependent system usually consists of two phases: enrolment and login. During

the enrolment phase, a user-defined or system-recommended passphrase is repeatedly by the user to construct the user model. The user can then login with the passphrase.

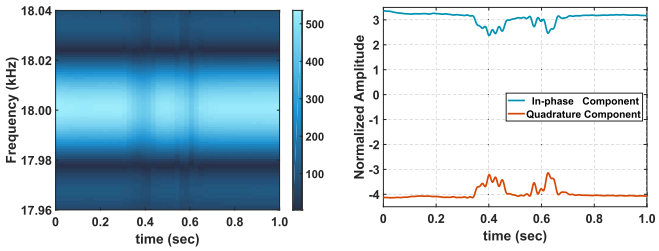
Due to the openness of acoustic signals and the popularity of social networks, we assume that the spoofers can easily acquire the users' voice samples corresponding to the passphrase by either placing a recording device surreptitiously nearby or utilizing the users' public speeches. In particular, to spoof LVID, the spoofers also need to provide the samples of the user's lip movements, which is assumed to be conducted by imitating the lip movements of the victim or playing the LVID signals. Therefore, we consider two types of spoofing attacks: lip-voice attacks and replay attacks. In lip-voice attacks, a spoofer plays the recording of the victim saying the passphrase and mimics lip movements of the victim simultaneously. In replay attacks, a spoofer records the diffused acoustic signals (both of the voice and the lip movements) when the victim logs in LVID, then plays the recordings to the smartphone.

IV. Sensing Lip Movements With Acoustic Signals

The applications of acoustic sensing have been investigated in position tracking [25], [26], gesture recognition [27] and pattern inference [34]. The targeted objects in these systems (e.g., smartphone [26], human body [25] and hand [27], [34]) have two main characteristics: (1) every part of the object has similar motion patterns (e.g., moving speed and direction), thus the object can be viewed as a whole; (2) the trajectories of the objects range from a few centimeters to dozens of centimeters. However, unlike these objects, lip is flexible and lip motion is a non-rigid process where different parts of the lip exhibit different motion patterns when speaking. Moreover, lip motion is limited to less than one centimeter in normal speech. Therefore, these acoustic sensing technologies are not suitable for lip movement sensing. There are two main approaches for sensing lip movements with acoustic signal: Doppler shift based approach, used by most existing acoustic-based lip movements sensing systems [13], [20], [22], [35] to estimate lip movements based on the frequency shift caused by the moving lips; Tan *et al.* [21] propose an envelope based approach for lip movements estimation, while we estimate lip movements based on the changes of phase shift. Compared with existing works, our approach is able to obtain fine-grained estimation of lip movements. Before we give detailed descriptions of our system design, we analyze the limitations of existing approaches for lip movements sensing and present the superiorities of our approach based on the changes of phase shift.

A. Limitations of Existing Approaches

1) Doppler Shift Based Approach: The moving speed of the lip can be calculated by measuring the Doppler frequency shift of the signals with STFT. However, The basic restrictions of time-frequency analysis limit the resolution of STFT, as a result, it can only provide a coarse-grained estimation of moving objects [27]. It is obvious that Doppler shift based approach is not suitable for estimating subtle lip movements that are usually less than one centimeter. Fig. 3(a) shows the STFT result of a moving lip when speaking the word "open",



(a) Doppler shift of lip movements. (b) I/Q waveforms.

Fig. 3. Sensing lip movements with acoustic signals.

where each audio frame contains 2,048 samples. We can see that most of the small frequency variations are buried in the wide frequency band around 18 kHz and we can only roughly recognize the lip movements from 0.32 s to 0.69 s.

2) Envelope Based Approach: SilentKey [21] extracts the user-specific features from the received signal based on the shape of the signal envelope (e.g., variation of the envelope, time interval between spike and duration of spike). However, the shape of the signal envelope is sensitive to multipath interference and is easily influenced by the unpredictable interference caused by the system delay.

B. The Changes of Phase Shift Based Approach

To overcome the limitations of the existing approaches, we propose to leverage the changes of phase shift to obtain the fine-grained estimation of lip movements. The basic idea is to treat the reflected acoustic signal as phase modulated signal whose phase shift changes with the lip movements. Compared to the Doppler frequency shift, we can easily measure the changes of phase shift of received signals in the time domain. As the acoustic signal transmitted at time t_1 is received at time t_2, there is a phase shift of $2\pi f(t_2 - t_1)$ between the transmitted and received signals, where f is the frequency of the transmitted signal. The propagation delay $t_2 - t_1$ equals $\frac{d}{c}$, where c is the speed of sound, d is the propagation distance between the transmitter and the receiver. It is obvious that the phase shift is proportional to d. Suppose that the lip moves 1 mm, then the change of phase shift is 0.21π, where f =18 kHz and c =343 m/s. Such a large phase shift change is very easy to detect. Fig. 3(b) shows the In-phase (I) and the Quadrature (Q) components of the baseband signal obtained from the same signal as in Fig. 3(a). Note that the I/Q components reflect the changes of phase shift from different aspects, and the method for acquiring I/Q components will be described in detail in Section V-C. We can learn from Fig. 3(b) that I/Q waveforms remain stationary when there is no lip motions and exhibit clear fluctuations as the lip moves when speaking the word "open" during 0.32 s $\sim$ 0.69 s. Compare to the envelope based approach, we propose a novel method to eliminate the interference in the received and obtain fine-grained lip movements estimation, which is described in details in Section V-C.

V. System Design

In this section, we first give an overview of LVID, then describe the detailed design.

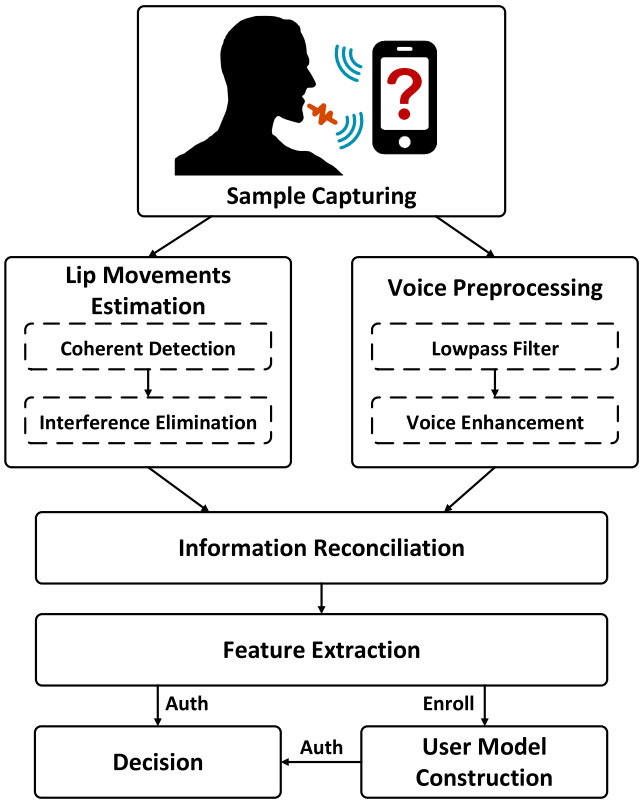


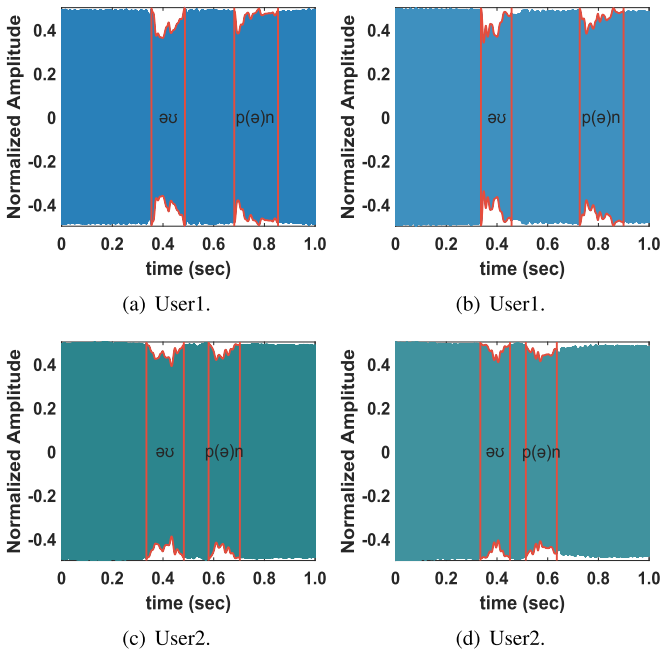
Fig. 4. System architecture.

A. System Overview

LVID consists of six major components: *Sample Capturing*, *Lip Movement Estimation*, *Voice Preprocessing*, *Information Reconciliation*, *Feature Extraction*, and *Decision*. As shown in Fig. 4, when the user conducts authentication, LVID emits high-frequency acoustic tones from the built-in speaker of the smartphone and captures the signals reflected by the user's lip with the microphone. Simultaneously, the user's voice is recorded by the microphone as well. The signal recorded by the microphone is stored as an audio file and will be uploaded to the server for further processing. The recorded signal is first split into two identical copies, one for obtaining fine-grained lip movements estimation and the other for generating a pure voice sample. The two signals are fused at the data level to generate the signal that contains both biometrics information. At last, LVID extracts reliable features from the fused signal and compares them to those stored in the user model (upon enrolment) to make a decision.

B. Sample Capturing

To sense the lip movements, LVID exploits the speaker of the user's smartphone to transmit the generated carrier signals. The generated sound is a high-frequency continuous wave acoustic signal $A\cos(2\pi ft)$, where A is the amplitude and f is the frequency of the acoustic signal. The frequency f is set within the range $18 \sim 19.75$ kHz. The reason that we choose this frequency range is that the response frequencies of most speakers and microphones on smartphones are from 50 Hz to 20 kHz and most people cannot hear sound with a frequency higher than 18 kHz [36]. In addition, the fact that the ambient noise becomes negligible at frequencies higher than 18 kHz makes LVID undisturbed by ambient noise.



(a) User1. (b) User1. (c) User2. (d) User2.

Fig. 5. The signals reflected by the moving lips of two different users when speaking the word 'open'.

LVID then exploits the microphone on the same device to capture the high-frequency signals reflected by the moving lips and the voice signals simultaneously. The interferences in recorded signals of lip movements and voice are different and non-negligible, which should be eliminated with different methods (refer to Section V-C and V-D). Since the frequency of human voice is usually lower than 3 kHz [37], which is far lower than the carrier frequency, the signals of lip movements and user voice can be easily divided by frequency analysis for further processing on the server.

Fig. 5 shows the signals reflected by the moving lips of two different users when speaking the word "open". Comparing Fig. 5(a), Fig. 5(b) with Fig. 5(c), Fig. 5(d), we can observe that the signals of different users demonstrate different patterns, while signals of the same user exhibit similar characteristics. These observations enable us to adopt the change of phase shift of the reflected signals of user lips for authentication.

C. Lip Movement Estimation

There may be multiple paths from the speaker to the microphone for the carrier signal, including the reflected path from the user's lip, the Line of Sight (LOS) path and the reflected paths from other surrounding objects (e.g., the user's face) as shown in Fig. 1. Suppose there are N different paths, the corresponding signal received by the microphone can be given as follows if we ignore the voice signal:

$$R(t) = \sum_{k=1}^{N} 2a_k(t)\cos\left(2\pi ft - 2\pi f\frac{d_k(t)}{c} - \theta_k(t)\right), \quad (1)$$

where t is the time slot, k represents the k-th path, $2a_k(t)$ is the amplitude, $2\pi f\frac{d_k(t)}{c}$ is the phase shift caused by the

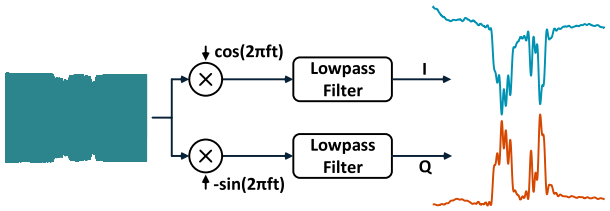


Fig. 6. The structure of the coherent detector.

propagation delay $\frac{d_k(t)}{c}$, and $\theta_k(t)$ is the unpredictable phase shift introduced by the system delay (e.g., hardware delay and software delay). The propagation delay stays constant when there is no moving object around. During speaking, the moving lips introduce changes to the propagation delay of reflected signals, which results in changes of phase shift of the corresponding signals in $R(t)$. Comparing terms in $R(t)$ with those in the original carrier signal, we can regard $R(t)$ as a mixture of several signals whose phases are modulated by the baseband signals, and the signals related to lip movements are part of the baseband signals. To obtain fine-grained lip movements estimation, we first demodulate the baseband signals from the received signal and then eliminate the interference.

1) Coherent Detection: The synchronization between the recorded signal and the transmitted signal makes it possible to demodulate the baseband signal from the recorded signal with a traditional coherent detector. The structure of the coherent detector is shown in Fig. 6, where the received signal is split into two identical copies to obtain the I-component and the Q-component of the baseband. To obtain the I-component, we multiply the $R(t)$ in Eq. (1) with the generated carrier signal $\cos(2\pi f t)$:

$$R(t) \times \cos(2\pi f t) = \sum_{k=1}^{N} a_k(t)\{\cos(2\pi f \frac{d_k(t)}{c} + \theta_k(t)) + \cos(4\pi f t - 2\pi f \frac{d_k(t)}{c} - \theta_k(t))\}, \quad (2)$$

the resultant is then passed through a lowpass filter to remove the high-frequency term with a frequency of $2f$ in Eq. (2). For simplicity, we use $\phi_k(t)$ to represent $2\pi f \frac{d_k(t)}{c} + \theta_k(t)$. Therefore, we have the I-component of the baseband as $I_m(t) = \sum_{k=1}^{N} a_k(t)\cos(\phi_k(t))$. Similarly, we can obtain the Q-component of the baseband as $Q_m(t) = -\sum_{k=1}^{N} a_k(t)\sin(\phi_k(t))$ by multiplying the signal with the phase shifted version of the carrier signal, i.e., $-\sin(2\pi f t)$.

2) Interference Elimination: It is obvious that the demodulated signals $I_m(t)$ and $Q_m(t)$ are still the mixtures of multipath signals. To obtain the fine-grained estimation of lip movements, it is necessary to eliminate the multipath interference introduced by other objects and LOS path. We category the multipath interference into two types: multipath interference from moving objects and static objects. Apart from the multipath interference, I/Q components also contain unpredictable phase interference $\theta_i(t)$ caused by system delay, which should be eliminated. An example of interference elimination is shown in Fig. 7, where the user speaks the word "open".

a) Multipath interference from moving objects: The first type of multipath interference is introduced by the other moving objects except lips. In most authentication scenarios, the other moving objects are the human body, which usually leads to a frequency shift ranging in [50,200] Hz [38] (e.g., waving, walking). Fortunately, the maximum frequency shift caused by the subtle lip movements is usually no more than 40 Hz, which means that a low-pass filter can be used to eliminate the multipath interference caused by the other moving objects. We set the cut-off frequency of the low-pass filter of the coherent detector at 40 Hz to achieve this goal together with baseband demodulation. After multiplying with the carrier signal, the voice signal has a high frequency ranging from $f - f_v$ to $f + f_v$ (f_v is the fundamental frequency of voice), which can be removed by the lowpass filter.

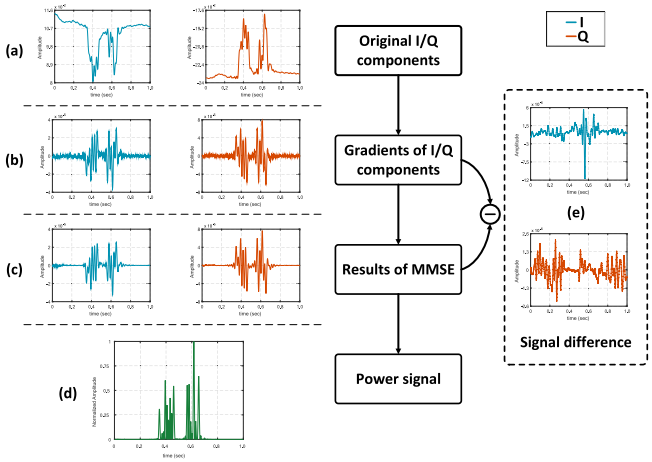


Fig. 7. An example of interference elimination.

b) Multipath interference from static objects: The second type of multipath interference is from static objects (e.g., the user's face). The interference of LOS path belongs to this type since the distance between the built-in speaker and microphone of the smartphone is constant, which is the same as static objects reflection. To eliminate this interference, we regard I-component as the sum of two components: static component $I_s(t)$ and dynamic component $I_d(t)$, which represent the superposition of signals reflected by the static objects and the moving lips respectively. Thus we can rewrite the I-component as follows:

$$I_m(t) = I_s(t) + I_d(t) = I_s(t) + \sum_{i \in P_d} a_i(t)\cos(\phi_i(t)), \quad (3)$$

where P_d is the set of paths corresponding to the lip movements. In ideal conditions, the phase shift corresponding to the static objects is constant, and $I_s(t)$ is constant as well. In practice, this static component $I_s(t)$ may also vary slowly as shown in Fig. 7 (a). Therefore, the multipath interference introduced by the static objects cannot be easily eliminated with a DC (Direct Current) removal filter. We observe that the static component fluctuates around a certain value. Therefore, the static component can be regarded as a sum of a constant term and a slowly changing stochastic term. We first calculate the gradient of the I-component to remove the constant term and then adopt a *minimum mean square error* (MMSE) [39] algorithm, which is capable of dealing with the random

additive error, to eliminate the slowly changing term. To demonstrate the effects of MMSE, we show the difference between the signals before and after MMSE processing in Fig. 7 (e). From Fig. 7 (b) and (c), we can see that the signal amplitude is approximately 0 in the absence of lip movements, which confirms that constant term and the slowly changing stochastic term are almost completely eliminated.

The I-component after multipath interference elimination is calculated as:

$$I(t) = \sum_{i \in P_d} \{a_i'(t)\cos(\phi_i(t)) - a_i(t)\phi_i'(t)\sin(\phi_i(t))\}, \quad (4)$$

where $a_i'(t)$ and $\phi_i'(t)$ are the derivative of $a_i(t)$ and $\phi_i(t)$ respectively. $a_i(t)$ is a coefficient associated with the propagation distance. Since lip movements are mostly less than one centimeter, the value of $a_i(t)$ hardly changes with lip movements [40]. As mentioned in Section IV-B, 1mm of lip movements leads to a change of 0.21π in $\phi_i(t)$. Therefore, the I-component after interference elimination $I(t)$ is approximately equivalent to $-\sum_{i \in P_d} a_i(t)\phi_i'(t)\sin(\phi_i(t))$. For simplicity, we use $A_i(t)$ to represent $-a_i(t)\phi_i'(t)$, then $I(t) = \sum_{i \in P_d} A_i(t)\sin(\phi_i(t))$. Similarly, we get the Q-component after multipath interference elimination as $Q(t) = \sum_{i \in P_d} A_i(t)\cos(\phi_i(t))$.

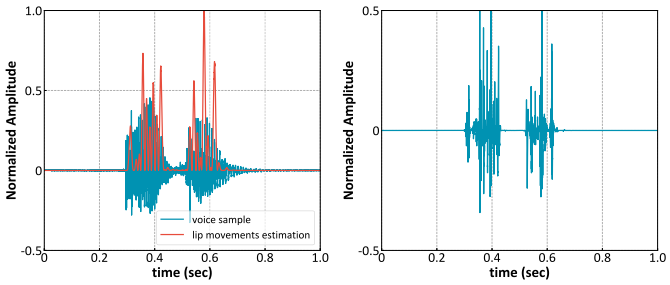
c) Unpredictable phase interference: Note that system delay has the same influence on all signals recorded by the microphone, which means $\theta_i(t)$ is the same for all paths. To address $\theta_i(t)$ in the remainder, we first combine Q-component and I-component as the real and imaginary part of a complex signal as follow, where $j^2 = -1$:

$$\begin{aligned} y(t) = Q(t) + I(t)j &= \sum_{i \in P_d} A_i(t)e^{j\phi_i(t)} \\ &= e^{j\theta(t)} \cdot \sum_{i \in P_d} A_i(t)e^{j2\pi f \frac{d_i(t)}{c}}. \end{aligned} \quad (5)$$

Then, we measure the power of $y(t)$ as follow (detailed derivations are omitted due to space constraints):

$$\begin{aligned} |y(t)|^2 &= |e^{j\theta(t)} \cdot \sum_{i \in P_d} A_i(t)e^{j2\pi f \frac{d_i(t)}{c}}|^2 \\ &= \sum_{\substack{l,m \in P_d \\ l \neq m}} 2|A_l(t)A_m(t)|\cos[2\pi f \frac{d_l(t) - d_m(t)}{c}] \\ &\quad + \sum_{i \in P_d} |A_i(t)|^2. \end{aligned} \quad (6)$$

In Eq. (6), the component $\sum_{i \in P_d} |A_i(t)|^2$ describes the absolute movement characteristics of the user's lip, and the other component shows the relative movement characteristics between the upper and the lower lips. By measuring the power of $y(t)$, we can not only eliminate the interference introduced by the unpredictable phase shift, but also obtain the fine-grained lip movement estimation. An illustration of the power signal is shown in Fig. 7(d), based on which we can recognize lip movements corresponding to each syllable.



(a) Voice sample and corresponding lip movements estimation.

(b) The fused signal.

Fig. 8. An example of information reconciliation.

D. Voice Pre-processing

As mentioned in Section V-B, the fundamental frequency of human voice is usually lower than 3 kHz, while the frequencies of the carrier signals are higher than 18 kHz. Therefore, the user's voice can be separated from the recorded mixed signal with a lowpass filter whose cut-off frequency is 10 kHz. As the energy of ambient noise is mainly distributed in the low-frequency band ($0 \sim 8$ kHz), the filtered signal may also contain ambient noise. Many methods have been proposed to reduce the noise in voice samples and it has been proved that it is hard to reduce the noise entirely without damaging the integrity of the voice signal [41]. In contrast, LVID maintains the integrity of the voice, and will not be affected by the residual noise since LVID extracts features from the fusion of lip movement estimation and the voice sample, where the lip movement features are free from ambient noise.

E. Information Reconciliation

Information reconciliation can be deployed in any stage of a multimodal biometrics system, and there are four main stages where fusion can take place.

1) The data level. Either the raw data or the processed data are fused.
2) The feature level. The features from various sources are merged.
3) The score level. The match score of classifiers for different sources are combined.
4) The decision level. The decisions of classifiers for different sources are pulled for the final decision via techniques such as majority voting.

The systems that consolidate information at an earlier stage are considered to be more effective since information loss increases along the process. However, the spatiotemporal difference in signals (e.g., one-dimensional audio data with 44,100 samples per second, two-dimensional video data with 30 frames per second) hinder the early fusion.

In this work, the information of lip movements and voice is captured by microphone simultaneously and saved as identical audio files, which leads to spatiotemporal consistency. LVID fuses these two biometrics at the data level by multiplying the fine-grained lip movement estimation with the pure voice sample in the time domain, which is equivalent to making a convolution with these two signals in the frequency domain. As the frequency range of lip movement estimation is much

Algorithm 1 Separate the Presence of Human Speech

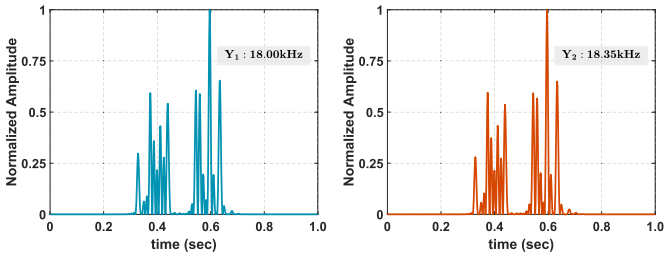
Input: Fused signal X
Output: The set of true positive segments P
Description: $P = \{P_1, P_2, \ldots\}$ *denotes the segment set;* $Start_i$ *and* End_i *denote the start and end time of* i*-th segment in* P *respectively.*
$P \leftarrow VAD(X)$;
$N \leftarrow$ The number of segments in P;
$i \leftarrow 1$;
while $i \leq N$ **do**
  $temp \leftarrow$ the standard deviation of $P(i)$;
  if $temp < Thr$ **then**
    Remove P_i from P;
    $N \leftarrow N - 1$;
  else if $Start_i - End_{i-1} < Interval$ **then**
    $P_{i-1} \leftarrow X[Start_{i-1} : End_i]$;
    $N \leftarrow N - 1$;
  else if $End_{i-1} - Start_{i-1} < Activity$ **then**
    Remove P_{i-1} from P;
    $N \leftarrow N - 1$;
  else
    $i \leftarrow i + 1$;
  end
end
return P;

smaller than that of the voice sample, performing a convolution in the frequency domain can be deemed as shaping the voice sample with the lip movement estimation. An example of information reconciliation is shown in Fig. 8.

F. Feature Extraction

In order to improve the quality of signals and facilitate feature extraction, the redundancy in the fused signal should be removed first. Specifically, we aim to remove segments that correspond to the absence of human speech and only keep segments that correspond to the syllables in passphrase. To achieve this goal, we first leverage the *Voice Activity Detection* (VAD) [42] algorithm to obtain several possible activity segments, and then filter out the segments whose standard deviations are smaller than an empirical threshold Thr. Thr is set as $3/10$ of the standard deviation of the fused signal. As shown in Fig. 8 (b), the signal is stable in the absence of lip movement, but fluctuates dramatically once the user starts to speak. Therefore, the standard deviation can be leveraged to infer whether a segment is active (contains speaking activity). As a syllable may consist of several segments, we splice segments together based on a pre-determined maximum interval time $Interval$. There may still be false positive activity segments with large variations but very short durations after splicing, which we eliminated using the minimum activity time $Activity$. Algorithm 1 summarizes details of redundancy elimination.

We choose *Mel-frequency cepstral coefficients* (MFCC) based features that are widely used in voice authentication systems. To obtain MFCC, a segment is first separated into



(a) The carrier frequency is 18 kHz (b) The carrier frequency is 18.35 kHz

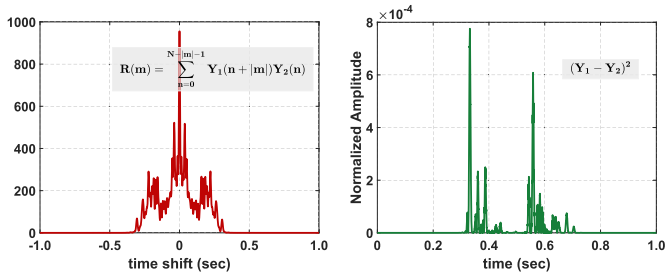
Fig. 9. Lip movement estimations obtained from different tones.

frames. The frame length is set to be 32ms and the overlap between frames is set to be 16ms such that the signal keeps stationary within and between frames. Then, mel-spectrum is calculated on each frame by FFT and Mel-filterbanks. Finally, we take the logarithm of the mel-spectrum and use *discrete cosine transformation* (DCT) to generate the 13-element MFCC feature vector. However, speech information is dynamic where the 13-element MFCC feature vector only describes the characteristics of a single frame. Therefore, we calculate the deltas and double deltas of the feature vectors to obtain 39-element feature vectors on the activity segment.

G. Decision

When first signing up the system, a user chooses the passphrase and repeats it three times to register in the authentication system. LVID uses the samples to construct the user model and store the model in the database for further authentication. In this paper, we adopt the *Gaussian Mixed Model* (GMM) to construct the user model. During the authentication process, LVID first extracts the features from the user input and then compare them with the user-specific GMM to make a decision.

To reduce the burden of data collection on a user and to improve the robustness of the constructed model, we propose a *data augmentation mechanism* based on frequency diversity. Since the wavelengths of sound with different frequencies are different, the phase shifts are also different. To implement this mechanism, we generate signal $A \sum_{i=1}^{NT} \cos[2\pi(f + \Delta f * i)t]$, which is the superposition of multiple tones instead of the original one. We treat each recorded tone separately as if the transmitted signal is a single tone, and the only difference during processing is the carrier frequency. We can acquire a lip movement estimation for each recorded tone. As the frequency shift caused by other body movements ranges within [50,200] Hz [38], we set the frequency interval Δf as 350 Hz to avoid the interference from adjacent frequencies, and the number of different tones NT is set as 5 to fit the high-frequency audio response of smartphones. We can also learn from Eq. (6) that different frequency f only changes the amplitude of the estimation but not the position of peaks. Fig. 9 shows lip movements estimations obtained from two different tones, which appear to be similar. To compare the two signals in Fig. 9, we demonstrate their *cross correlation coefficient* and *Eucliden distance* in Fig. 10. We can observe that the maximum cross correlation coefficient takes place at 0 and the Eucliden distance has significant



(a) Cross-correlation coefficient (b) Eucliden distance

Fig. 10. The similarities and differences between the estimation obtained from the same sample.

fluctuations, which indicates that the estimations for the same lip movements generated from tones of different frequencies are similar in shape though the magnitudes are slightly different. Therefore, we can obtain multiple estimations for identical lip movements from different frequencies and these estimations can be fused with voice samples to generate different user samples.

VI. Evaluation

In this section, we first evaluate the overall performance of LVID, then compare LVID with unimodal authentication systems (e.g., voice, lip movements) and examine the resistance of LVID to spoofing attacks. Moreover, we also demonstrate the robustness of LVID to the length of passphrase, the distance between the smartphone and users' mouthes, the position change of the smartphone and the smartphone types.

A. Experimental Setup

1) Implementation: We implement a prototype of LVID on off-the-shelf smartphones to evaluate and validate its performance and effectiveness. As discussed in Section V, any commercial smartphones equipped with speakers and microphones can be used to capture the audio samples by installing an APP that has access to the speaker, microphone and storage, high-frequency acoustic signal generation and communication with the server can also be performed. The generated high-frequency acoustic signals are at range of 18 $\sim$ 20 kHz. The server is a PC with 3.1 GHz CPU and 8 GB memory.

2) Default Setting: We evaluate LVID on three different smartphones: Samsung C9 (Android 6.0), Mi 5s (Android 8.0) and Samsung S4 (Android 5.1), which differ in audio chipsets and the position of the audio components. Samsung C9 is chosen as the default smartphone in most experiments since Android 6.0 has the largest market share among these versions [43]. The sampling rate of speakers and microphones is set as 48kHz. Our experiments are conducted under 3 different environments: a laboratory (relatively quiet), a market (slightly noisy) and a street (very noisy). The laboratory is the default environment unless claimed otherwise. We also investigate the impact of authentication distance and position on the performance of LVID. In most of our experiments, the participants perform authentication when the phone is horizontally held towards the mouth at a distance of 4cm.

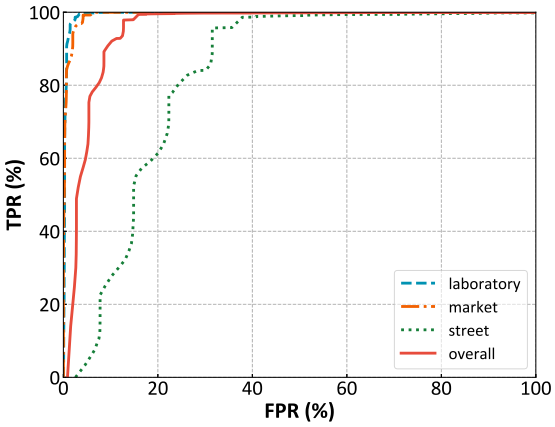


Fig. 11. ROC curves under different environments.

3) Data Collection: Our experiments involve 104 participants aged from 19 to 38, including 62 males and 42 females. For each scenario, the 104 volunteers talk to the smartphone as required and perform user authentication. As discussed in Section III-C, LVID is a text-dependent system, which means that a passphrase should be preset during enrolment. We predefine 9 passphrases for evaluations but in fact users can freely select their passphrases. The length of the passphrases ranges from 2 $\sim$ 10 syllables, and we choose 5 $\sim$ 7 syllables for passphrases in most experiments. Each participant is required to speak each passphrase for 3 times upon enrolment to build the user-specific model and then perform authentications 10 times for testing. With the consent of the participants, we made the raw data public without violating user privacy.

4) Attacks: We evaluate LVID under two types of spoofing attacks as mentioned in Section III-C. Each attack is implemented in the three environments. For each passphrase, the spoofer makes 10 trials. In lip-voice attacks, we asked the participants to randomly select another participant as the victim, and they play as the spoofers to mimic the characteristics of lip movements of the selected victim. The spoofers synchronize their lip movements with the voice samples of the victims, i.e., the spoofers mimic the lip movement patterns of the victims while playing the voice samples of the same victims. In replay attacks, the spoofer secretly records the signals of the victim performing authentication, then replays the signals with a loudspeaker that has a frequency response as high as 18 kHz.

5) Metrics: We evaluate our system with the following metrics: *True Positive Rate* (TPR) is the probability that LVID correctly recognizes legitimate users. *False Positive Rate* (FPR) is the likelihood that LVID incorrectly declares a attacker as an legitimate user. *Receiver Operating Characteristic* (ROC) curve describes the relationship between TPR and FPR when varying the detection threshold. *True Negative Rate* (TNR) is the probability that LVID correctly detects a spoofer. TPR and TNR measure the accuracy of LVID for user identification and spoofer detection respectively.

B. Overall Performance

We first present the overall performance of LVID under lip-voice attack and replay attack in Fig. 11, which depicts the

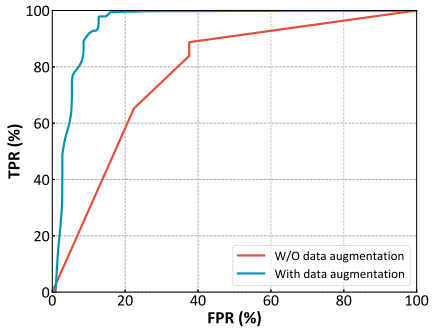


Fig. 12. The effect of the data augmentation mechanism.

TABLE I
EXECUTION TIME OF EACH COMPONENT IN LVID

	With data augmentation	W/O data augmentation
Data transmission (s)	0.005	0.005
Lip Movement Estimation (s)	0.85	0.45
Voice Preprocessing (s)	0.04	0.04
Information Reconciliation (s)	0.05	0.02
Feature Extraction (s)	0.22	0.07
Decision (s)	0.07	0.006
Total (s)	1.235	0.591

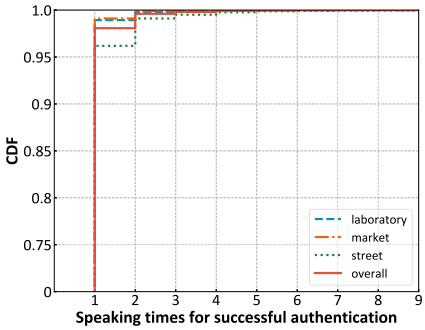


Fig. 13. CDF of the speaking times for successful authenticate.

ROC curves of LVID under different environments. We can observe that LVID works well in different environments, and performances the best in the quiet environment. The values of the *area under the curve* (AUC) of ROC in different environments are 0.9957, 0.9947 and 0.8263 respectively. The overall AUC is 0.9542. The "overall" category means that the value is estimated by all the data from different environments in all figures unless claimed otherwise. We also compare the performance of LVID with and without data augmentation mechanism in Fig. 12. We can observe that data augmentation mechanism significantly improves the AUC from 0.8976 to 0.9542.

We further evaluate the user experience by the authentication time and the number of times for successful authentication. Table I shows the execution time of five major components of LVID after the user speaks the passphrase. We can observe that the total authentication time of LVID with data augmentation is 1.235 s and that without data augmentation is only 0.591 s. The most time-consuming component in LVID is lip movement estimation which conducts a lot of computation to eliminate the interference of multipath and unpredictable phase. In spite of this, the total execution time is

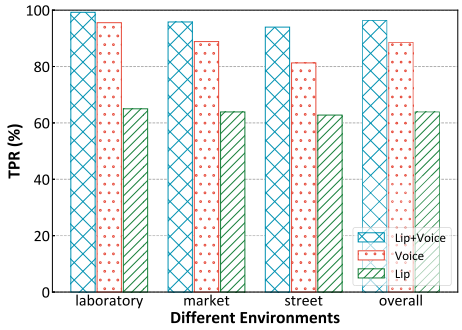


Fig. 14. Comparison of LVID with unimodal authentication system.

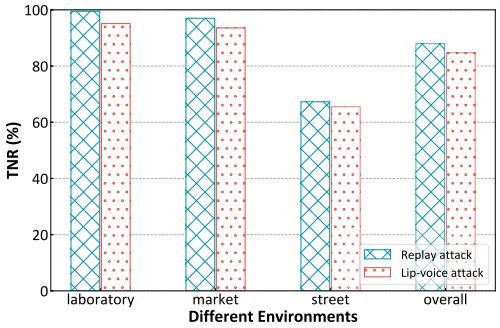


Fig. 15. The resistance to spoofing attacks.

acceptable. Since we conduct *Client-Server* model, it is necessary to take the time for data transmission into consideration, the average file size of each sample is 129.8 KB, it takes average 0.005 s for data transmission between client and server. Fig. 13 shows the *Cumulative Distribution Function* (CDF) of the number of times for legitimate users to be successfully authenticated in three different environments. We can observe that 98.07% legitimate users can be successfully authenticated in one attempt, and 99.6% legitimate users can be successfully authenticated within three attempts. The short authentication time and high probability for successful authentication indicate that LVID is very user-friendly and practical.

C. Comparison With Single Voice and Lip Biometrics

Next, we compare the performance of LVID with voice-based [44] and lip-based [21] unimodal authentication systems. Note that we select the same data set for all the systems. As shown in Fig. 14, LVID performs better than the voice authentication system in all three environments. The TPR of LVID is above 93% while that of the voice authentication system decreases with the increase of ambient noise. For lip authentication system, the TPR remains at slightly over 60%. This is because the energy of ambient noise is mainly distributed below 8 kHz [45] and the carrier signals used for lip movements sensing are higher than 18 kHz which is free of ambient noise, yet the voice signals will be seriously disturbed by ambient noise. This demonstrates that LVID is robust to ambient noise and has better performance than voice-based and lip-based authentication systems.

D. Performance Under Spoofing Attacks

Fig. 15 demonstrates that LVID performs well in spoofer detection. As shown in Fig. 15, the values of TNR for lip-voice

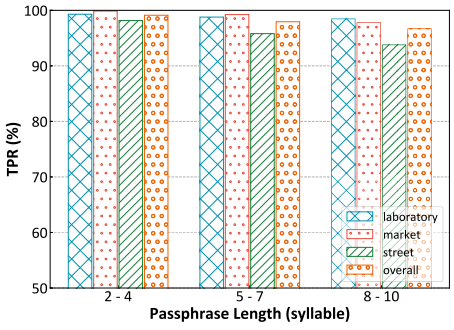


Fig. 16. Impact of passphrase length.

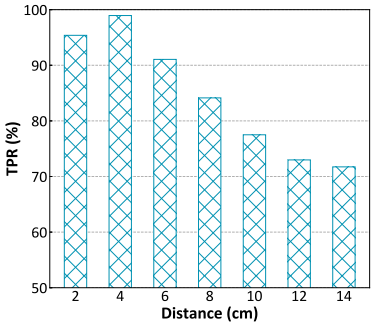


Fig. 17. Impact of authentication distance.

attacks in the laboratory and the market are 95.1% and 93.56% respectively. For replay attacks, the TPR is as high as 99.49% in the laboratory and 97% in the market. For the street scenario, the TNR drops to 65.48% and 67.28% for lip-voice and replay attacks respectively. The results reveal that it is quite difficult to spoof LVID in most scenarios. Note that the spoofer cannot spoof LVID with sole voice signals or sole lip movement signals.

E. Impact of Passphrase Length

In this section, we show how the length of passphrase affects the performance of LVID. A longer passphrase usually provides more information for authentication, however, it also brings more overhead. In our experiment, we classify the passphrases into three categories according to their lengths: $2 \sim 4$ syllables, $5 \sim 7$ syllables and $8 \sim 10$ syllables. Fig. 16 displays the TPR of LVID with different passphrase lengths in different environments. We can see that the TPR is over 93.78% for different passphrase lengths, and the TPR changes very little when the passphrase length changes.

F. Impact of Authentication Distance

Acoustic signals attenuate quickly as the propagation distance increases, thus a longer propagation distance between the smartphone and the user's mouth may damage the information integrity of the carrier signals of lip movements, which further leads to a performance degradation of the authentication system. Therefore, we evaluate the impact of the distance between the smartphone and the user's mouth on LVID. We consider a three-dimensional coordinate system (X, Y, Z), where the user's face is regarded as a surface aligning with the XOY plane; the mouth locates at $(x, y, 0)$; the smartphone locates

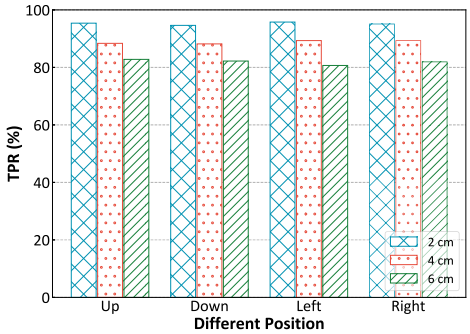


Fig. 18. Impact of different position.

at (x, y, z); and z is the distance between the smartphone and the user's mouth. Fig. 17 reveals the TPR of LVID when the distance between the smartphone and the user's mouth varies from 2 cm to 14 cm. Although the accuracy decreases with a longer distance, the TPR is over 84.13% when the distance is no more than 8cm. This suggests that users should better keep the distance within 8 cm when using LVID.

The experiment results show that the performance of LVID degrades as the authentication distance increases. The authentication distance is limited by the signal strength for lip movement sensing, i.e., the power of the transmitted signal limits the sensing range. Fortunately, the application scenarios of LVID, i.e., mobile device authentication, usually have a short authentication distance, which guarantees the security distance. Although voice authentication system may have a longer authentication distance, it is vulnerable to various well-known attacks, which endangers the system security.

G. Impact of Authentication Position

In practice, the relative position of the smartphone and the user's mouth changes in each authentication, which will also be different during the enrolment phase and the authentication phase. Hence, we investigate the performance of LVID when the relative position of the smartphone and the user's mouth changes. During enrolment, the smartphone is at (x, y, z) and the user's mouth is at $(x, y, 0)$. During authentication, the smartphone moves along the XOY plane in one of the four directions, i.e., up, down, left, right, in each experiment. Fig. 18 reveals the TPR of LVID when the smartphone moves 2 cm, 4 cm and 6 cm in each direction. We can observe that the TPR decreases when the smartphone moves away from the user, but the smallest TPR is still as high as 80.67% (Left, 6cm).

H. Impact of Smartphone Model

Finally, we examine the impact of different smartphones on LVID's performance. We use three Android smartphones, namely Samsung C9, Mi 5s and Samsung S4, which differ in size, audio chipsets and the location of speakers and microphones. All these smartphones are able to record and playback sound with a frequency over 18kHz. Fig. 19 shows the performance of LVID on different smartphones with different lengths of passphrases. The "overall" value in this figure is estimated by all the data from different smartphones. We can

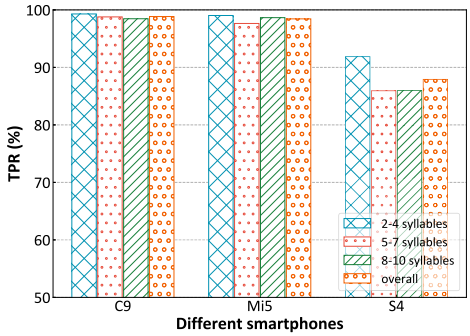


Fig. 19. Impact of different smartphones.

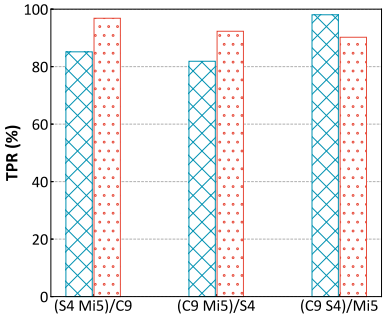


Fig. 20. Impact of cross-phone scenario.

observe that the overall TPR on C9 and Mi5 are close to 99%, and the overall TPR on S4 is 87.9%, which demonstrates that LVID is compatible with different smartphones. The reason that S4 performs poorly may be that S4 is an old phone model with relatively low audio system quality. We can also observe that the performance of LVID varies with different smartphones and it is likely that one person owns more than one smartphones, so we further examine the scenario where the enrolment and authentication take place on different smartphones. As shown in Fig. 20, the performance is associated with both of the smartphones for enrolment and authentication, and LVID can still works well in the cross-phone scenario.

VII. Discussion

In this section, we theoretically analyze the resistance of LVID to the attacks mentioned in Section III-C and discuss the limitations of LVID.

A. Security Analysis

Robustness to various attacks and high authentication accuracy are both important to an authentication system. The former realizes security and the latter guarantees user experience. We theoretically analyze why LVID is able to defend against the spoofing attacks mentioned in Section III-C.

After performing MMSE on I/Q components, we calculate the difference between signals before and after MMSE, as shown in Fig. 7 (e), then we compute the ratio of the signal power after MMSE to that of the signal difference. This ratio can accurately reveal whether there are moving objects near the microphone of the smartphone. Through extensive experiments, we set an empirical value to decide whether there are lip movements. If no lip movements are detected, LVID deems that the input is from a spoofer. This can defend against voice-only attacks but will fail to detect lip-only attacks as the spoofer mimics the lip movements of the victim. Fortunately, due to a lack of voice sample in lip-only attacks, fused signal cannot be reconstructed, thus LVID can identify such attacks. In lip-voice attacks, due to asynchronization and mismatch between lip movements and the voice sample, LVID can easily detect the spoofer since the fused signal is far different from the real one.

Replay attacks are the most difficult to defend against, since the spoofer possesses both the lip movement information and voice sample of the victim. The signal recorded by the spoofer is also a mixture of the signals that travel through multiple paths from the speaker of the victim's smartphone to the spoofer, thus the signal can be expressed similar to Eq. (1) as

$$R_s(t) = \sum_{i=1}^{N_1} 2a_i(t)\cos\left(2\pi f t - 2\pi f \frac{d_i(t)}{c} - \theta_s(t)\right). \quad (7)$$

The signal reflected by the victim's moving lips is a component of this signal. When the spoofer replays this signal to log in LVID, LVID also plays the carrier signal $A\cos(2\pi f t)$. The signal received by the microphone is the superposition of the two after multipath propagation. The component corresponding to the carrier signal is

$$R_c(t) = \sum_{m=1}^{N_3} 2a_j(t)\cos\left(2\pi f t - 2\pi f \frac{d_m(t)}{c} - \theta_r(t)\right), \quad (8)$$

and that corresponding to $R_s(t)$ is

$$\begin{aligned} R'_s(t) &= \sum_{k=1}^{N_2} \alpha_k(t) R_s\left(t - \frac{d_k(t)}{c} - \tau(t)\right) \\ &= \sum_{k=1}^{N_2}\sum_{i=1}^{N_1} 2a_{ki}(t)\cos\left(2\pi f t - 2\pi f \frac{d_{ki}(t)}{c} - \theta(t)\right), \end{aligned} \quad (9)$$

where $\alpha_k(t)$ is the attenuation coefficient and $a_{ki}(t) = \alpha_k(t)a_i(t), d_{ki}(t) = d_k(t) + d_i(t), \theta(t) = \theta_s(t) + 2\pi f t\tau(t)$. Ignore the phase shift caused by the victim's smartphone and the spoofer's recording device, we can simplify the received signal as

$$\begin{aligned} R(t) = &\sum_{k=1}^{N_2}\sum_{i=1}^{N_1} 2a_{ki}(t)\cos\left(2\pi f t - 2\pi f \frac{d_{ki}(t)}{c}\right) \\ &+ \sum_{m=1}^{N_3} 2a_j(t)\cos\left(2\pi f t - 2\pi f \frac{d_m(t)}{c}\right). \end{aligned} \quad (10)$$

We consider that the attack takes place in an ideal environment where all objects stay still including the playback device and the victim's smartphone. The signal corresponding to the carrier signal can be regarded as the static component in the I/Q components. The replay operation only propagates $R_s(t)$ again and will not affect the original signal composition in $R_s(t)$, i.e., the static component and the dynamic component. After

processing $R(t)$, we obtain the power of the complex signal as

$$|y(t)|^2 = |\sum_{k=1}^{N_2}\sum_{i=1}^{N_1'} A_{ki}(t)e^{j(2\pi f\frac{d_{ki}(t)}{c})}|^2, \quad (11)$$

where $N_1' < N_1$ represents the paths related to the victim's moving lips. We leave out the expansion of this expression due to space constraints, which will not affect the following analysis. Compared with Eq. (6), the expanding expression is the sum of multiple terms including $|A_{k_1i}(t)A_{k_2i}(t)|\cos[2\pi f\frac{d_{k_1i}(t)-d_{k_2i}(t)}{c}]$. Since $d_{ki}(t) = d_k(t) + d_i(t)$, the term can be rewritten as $|A_{k_1i}(t)A_{k_2i}(t)|\cos[2\pi f\frac{d_{k_1}(t)-d_{k_2}(t)}{c}]$. This term shows that the dual multipath interference makes the lip movement estimation different from the original one. Taking into account the unpredictable phase shift caused by the devices, the difference will be even more significant. In this way, the VLID can successfully detect the replay attacks.

B. Limitations

Our prototype of LVID on Android smartphones has demonstrated its effectiveness, yet the performance of LVID on other mobile systems needs to be further investigated. Moreover, with the popularity of IoT, we are considering to deploy LVID on IoT devices since LVID only requires a commercial speaker and a microphone for capturing samples and a network module for information transmission.

The change of smartphone position during sample capturing will have negative influences on the authentication result since the change of phone position also changes the relative position between the phone and the lips. In the data collection process, the participants are required to hold the smartphones, and may inevitably have small shakes. The experiment results have confirmed that by using MMSE algorithm to eliminate the interference during signal processing, LVID is able to tackle such small disturbances with an ideal performance.

We simply leverage MFCC to extract user-specific features and GMM to make a decision, which results in an accuracy of 89.29% in street, higher than unimodal biometric systems but not enough for practical application. Other feature extraction schemes and classification algorithms can be explored to further improve the performance of LVID.

Our evaluation is conducted with a restricted number of college students. More participants with a more diverse background and wider age ranges will help better understand the performance of our system. Moreover, the evaluation only lasts for several months while a long-term investigation should be conducted to learn how individual characteristics change over time and influence the performance. We believe that a periodically updated user-specific model may be necessary to ensure the robustness of the system.

VIII. Conclusion

In this paper, we propose LVID, a multimodal biometric authentication system on smartphones which combines the advantages of the unimodal biometric authentication system based on either lip movements or voice. Without requiring special hardware or cumbersome operations, LVID simultaneously captures the samples of lip movements and voice with the built-in speaker and microphone on smartphones when the user speaks the passphrase. LVID obtains the fine-grained estimation of lip movements and pure voice sample from the recorded signal and then fuses these two biometrics at the data level for accurate authentication. Extensive experiments show that LVID is effective and robust for user authentication in various environments, and can defend against different types of voice spoofing attacks with a very high accuracy.

References

[1] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1701–1708.

[2] D. Peralta, I. Triguero, R. Sanchez-Reillo, F. Herrera, and J. M. Benitez, "Fast fingerprint identification for large databases," *Pattern Recognit.*, vol. 47, no. 2, pp. 588–602, Feb. 2014.

[3] Q. Zou, L. Ni, Q. Wang, Q. Li, and S. Wang, "Robust gait recognition by integrating inertial and RGBD sensors," *IEEE Trans. Cybern.*, vol. 48, no. 4, pp. 1136–1150, Apr. 2018.

[4] D. D. Zhang, *Biometric solutions: For authentication E-world*. Berlin, Germany: Springer, 2012, vol. 697.

[5] A. K. Jain, P. Flynn, and A. A. Ross, *Handbook Biometrics*. Berlin, Germany: Springer, 2007.

[6] *Google Smart Lock*. Accessed: 2019. [Online]. Available: https://get.google.com/smartlock/

[7] *Voice Biometric on Smartphone*. Accessed: 2018. [Online]. Available: http://shouji.baidu.com/

[8] *The voiceprint of wechat*. Accessed: 2018. [Online]. Available: http://thenextweb.com/apps/2015/03/25/wechat-onios-now-lets-you-log-in-using-just-your-voice/

[9] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Commun.*, vol. 66, pp. 130–153, Feb. 2015.

[10] T. Kinnunen *et al.*, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. INTERSPEECH*, 2017, pp. 1–5.

[11] M. Zhou, Z. Qin, X. Lin, S. Hu, Q. Wang, and K. Ren, "Hidden voice commands: Attacks and defenses on the VCS of autonomous driving cars," *IEEE Wireless Commun.*, to be published. doi: 10.1109/MWC.2019.1800477.

[12] Q. Wang *et al.*, "VoicePop: A pop noise based anti-spoofing system for voice authentication on smartphones," in *Proc. IEEE INFOCOM*, Apr. 2019, pp. 2062–2070.

[13] L. Zhang, S. Tan, and J. Yang, "Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication," in *Proc. ACM CCS*, Oct. 2017, pp. 57–71.

[14] S. Chen *et al.*, "You can hear but you cannot steal: Defending against voice impersonation attacks on smartphones," in *Proc. IEEE ICDCS*, Jun. 2017, pp. 183–195.

[15] Y. Meng *et al.*, "WiVo: Enhancing the security of voice control system via wireless signal in IoT environment," in *Proc. ACM MobiHoc*, Jun. 2018, pp. 81–90.

[16] M. Rusko, T. Marian, S. Darjaa, M. Ritomský, and I. Guoth, "Influence of noise on the speaker verification in the air traffic control voice communication," *J. Acoust. Soc. Amer.*, vol. 141, no. 5, p. 3469, Jun. 2017.

[17] J. Luettin, N. A. Thacker, and S. W. Beet, "Speaker identification by lipreading," in *Proc. IEEE ICSLP*, vol. 1, Oct. 1996, pp. 62–65.

[18] P. Singh, V. Laxmi, and M. S. Gaur, "Speaker identification using optimal lip biometrics," in *Proc. IEEE ICB*, Mar. 2012, pp. 472–477.

[19] H. E. Cetingul, Y. Yemez, E. Erzin, and A. M. Tekalp, "Discriminative analysis of lip motion features for speaker identification and speech-reading," *IEEE Trans. Image Process.*, vol. 15, no. 10, pp. 2879–2891, Oct. 2006.

[20] L. Lu *et al.*, "LipPass: Lip reading-based user authentication on smartphones leveraging acoustic signals," in *Proc. IEEE INFOCOM*, Apr. 2018, pp. 1466–1474.

[21] J. Tan, X. Wang, C.-T. Nguyen, and Y. Shi, "SilentKey: A new authentication framework through ultrasonic-based lip reading," *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol.*, vol. 2, no. 1, p. 36, 2018.

[22] L. Lu *et al.*, "Lip reading-based user authentication through acoustic sensing on smartphones," *IEEE/ACM Trans. Netw.*, vol. 27, no. 1, pp. 447–460, Feb. 2019.

[23] M. Zhou, Q. Wang, T. Lei, Z. Wang, and K. Ren, "Enabling online robust barcode-based visible light communication with realtime feedback," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 8063–8076, Dec. 2018.

[24] A. Ross and A. K. Jain, "Multimodal biometrics: An overview," in *Proc. 12th Eur. Signal Process. Conf.*, Sep. 2004, pp. 1221–1224.

[25] Y.-C. Tung and K. G. Shin, "EchoTag: Accurate infrastructure-free indoor location tagging with smartphones," in *Proc. 21st Annu. Int. Conf. Mobile Comput. Netw.*, Sep. 2015, pp. 525–536.

[26] Y. Zhang, J. Wang, W. Wang, Z. Wang, and Y. Liu, "Vernier: Accurate and fast acoustic motion tracking using mobile devices," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, Apr. 2018, pp. 1709–1717.

[27] W. Wang, A. X. Liu, and K. Sun, "Device-free gesture tracking using acoustic signals," in *Proc. ACM MobiCom*, Oct. 2016, pp. 82–94.

[28] L. Zhang, S. Tan, J. Yang, and Y. Chen, "VoiceLive: A phoneme localization based liveness detection for voice authentication on smartphones," in *Proc. ACM CCS*, Oct. 2016, pp. 1080–1091.

[29] M. Ichino, H. Sakano, and N. Komatsu, "Multimodal biometrics of lip movements and voice using kernel Fisher discriminant analysis," in *Proc. 9th Int. Conf. Control, Automat., Robot. Vis.*, Dec. 2006, pp. 1–6.

[30] H. E. Çetingül, E. Erzin, Y. Yemez, and A. M. Tekalp, "Multimodal speaker/speech recognition using lip motion, lip texture and audio," *Signal Process.*, vol. 86, no. 12, pp. 3549–3558, Dec. 2006.

[31] R. W. Frischholz and U. Dieckmann, "BiolD: A multimodal biometric identification system," *Computer*, vol. 33, no. 2, pp. 64–68, Feb. 2000.

[32] U. Dieckmann, P. Plankensteiner, R. Schamburger, B. Fröba, and S. Meller, "SESAM: A biometric person identification system using sensor fusion," in *Proc. AVBPA*, Berlin, Germany: Springer, 1997, pp. 301–310.

[33] G. C. Martin. *Preston Blair Phoneme Series*. Accessed: 2019. [Online]. Available: http://www.garycmartin.com/mouth_shapes.html

[34] M. Zhou *et al.*, "Patternlistener: Cracking Android pattern lock using acoustic signals," in *Proc. ACM CCS*, Oct. 2018, pp. 1775–1787.

[35] J. Tan, C.-T. Nguyen, and X. Wang, "Silenttalk: Lip reading through ultrasonic sensing on mobile phones," in *Proc. IEEE INFOCOM*, May 2017, pp. 1–9.

[36] Q. Wang, K. Ren, M. Zhou, T. Lei, D. Koutsonikolas, and L. Su, "Messages behind the sound: Real-time hidden acoustic signal capture with smartphones," in *Proc. ACM MobiCom*, Oct. 2016, pp. 29–41.

[37] *Human Voice Frequency Range*. Accessed: 2019. [Online]. Available: http://www.seaindia.in/blog/human-voice-frequency-range/

[38] S. Gupta, D. Morris, S. Patel, and D. Tan, "SoundWave: Using the Doppler effect to sense gestures," in *Proc. ACM CHI*, May 2012, pp. 1911–1914.

[39] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[40] *Stokes's Law of Sound Attenuation*. Accessed: 2019. [Online]. Available: https://en.wikipedia.org/wiki/Stokes%27s_law_of_sound_attenuation

[41] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[42] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

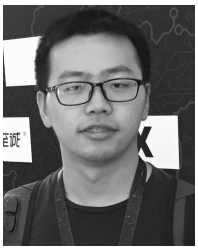
[43] *Android Version Distribution*. Accessed: 2019. [Online]. Available: https://developer.android.com/about/dashboards//

[44] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[45] M. Zhou, Q. Wang, K. Ren, D. Koutsonikolas, L. Su, and Y. Chen, "Dolphin: Real-time hidden acoustic signal capture with smartphones," *IEEE Trans. Mobile Comput.*, vol. 18, no. 3, pp. 560–573, Mar. 2019.

Libing Wu received the Ph.D. degree in computer science from Wuhan University, China, in 2006. He is currently a Professor with the School of Cyber Science and Engineering and the School of Computer Science, Wuhan University. His research interests include network management, trusted software, and wireless sensor networks.

Jingxiao Yang is currently pursuing the M.S. degree with the School of Computer Science, Wuhan University, China. His research interests include mobile system and wireless security. He was a recipient of the First Prize in the National Graduate Contest on Application, Design, and Innovation of Mobile-Terminal, China, in 2017.



Man Zhou received the B.E. degree in information security from Wuhan University, China, in 2016, where he is currently pursuing the Ph.D. degree with the School of Cyber Science and Engineering. His research interests include mobile security, mobile computing, and the IoT security. He was a recipient of the First Prize in the National Graduate Contest on Application, Design, and Innovation of Mobile-Terminal, China, in 2016 and 2017.

Yanjiao Chen received the B.E. degree in electronic engineering from Tsinghua University in 2010 and the Ph.D. degree in computer science and engineering from The Hong Kong University of Science and Technology in 2015. She is currently a Professor with Wuhan University, China. Her research interests include computer networks, wireless system security, and network economy.

Qian Wang (SM'18) received the Ph.D. degree from the Illinois Institute of Technology, USA. He is currently a Professor with the School of Cyber Science and Engineering, Wuhan University. His research interests include AI security, data storage, search and computation outsourcing security and privacy, wireless system security, big data security and privacy, and applied cryptography. He is also a member of ACM. He received the National Science Fund for Excellent Young Scholars of China in 2018. He is also an expert under National 1000 Young Talents Program of China. He was a recipient of the 2018 IEEE TCSC Award for Excellence in Scalable Computing for Early Career Researcher and the 2016 IEEE Asia–Pacific Outstanding Young Researcher Award. He was a co-recipient of several Best Paper and Best Student Paper Awards from the IEEE ICNP 2011, the WAIM 2014, the IEEE TrustCom 2016, and the IEEE ICDCS 2017. He serves as an Associate Editor for the IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING (TDSC) and the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY (TIFS).