Lip Reading-Based User Authentication Through Acoustic Sensing on Smartphones

Li Lu, Jiadi Yu, *Member, IEEE*, Yingying Chen, *Senior Member, IEEE*, Hongbo Liu, Yanmin Zhu, *Senior Member, IEEE*, Linghe Kong, *Member, IEEE*, and Minglu Li

***Abstract*— To prevent users' privacy from leakage, more and more mobile devices employ biometric-based authentication approaches, such as fingerprint, face recognition, voiceprint authentications, and so on, to enhance the privacy protection. However, these approaches are vulnerable to replay attacks. Although the state-of-art solutions utilize liveness verification to combat the attacks, existing approaches are sensitive to ambient environments, such as ambient lights and surrounding audible noises. Toward this end, we explore liveness verification of user authentication leveraging users' mouth movements, which are robust to noisy environments. In this paper, we propose a lip reading-based user authentication system, *LipPass*, which extracts unique behavioral characteristics of users' speaking mouths through acoustic sensing on smartphones for user authentication. We first investigate Doppler profiles of acoustic signals caused by users' speaking mouths and find that there are unique mouth movement patterns for different individuals. To characterize the mouth movements, we propose a *deep learning-based method* to extract efficient features from Doppler profiles and employ *softmax function*, *support vector machine*, and *support vector domain description* to construct multi-class identifier, binary classifiers, and spoofer detectors for mouth state identification, user identification, and spoofer detection, respectively. Afterward, we develop a *balanced binary tree-based authentication approach* to accurately identify each individual leveraging these binary classifiers and spoofer detectors with respect to registered users. Through extensive experiments involving 48 volunteers in four real environments, *LipPass* can achieve 90.2% accuracy in user identification and 93.1% accuracy in spoofer detection.**

***Index Terms*— Lip reading, user authentication, acoustic signals.**

I. Introduction

MOBILE devices are increasingly pervasive and common in our daily life. Due to the fast and convenient data connections of mobile devices, an increasing number of people use mobile devices as frequent storage medium for sensitive information including personal (e.g., identity ID) and financial (e.g., CVS code of credit cards) information, etc. Thus, more and more users are concerned with the privacy-preserving problem in mobile devices. According to a report from Symantec [1], 78% of users are concerned about losing information on their personal devices and 41.2% of users have lost their mobile devices with sensitive information leakage. Because of the potential risks, it is essential to develop a powerful user authentication to prevent users' sensitive information from leakage on mobile devices.

The most widely deployed user authentication approach is the password. But passwords are usually hard to remember and vulnerable to stealing attacks. To deal with the problem, many biometric-based techniques are developed to perform user authentication on mobile devices, such as Fingerprint [2], Face recognition [3], Voiceprint [4] authentications, etc., and relative products are already developed, i.e., Apple Touch ID [5], Android Face Unlock [6], and Apple 'Hey, Siri' voiceprint identification [7], etc. However, such authentications are only based on physiological characteristics, suffering from replay attacks [8]. To combat the replay attacks, liveness verification [9] becomes an attractive approach to improve the reliability of user authentication. Luettin *et al.* [10] propose a visual features-based method to distinguish a face of a live user from a photo. Zhang *et al.* [8] propose a phoneme localization approach to verify a passphrase whether spoken by a live user or pre-recorded by attackers. However, these recent works are sensitive to ambient environments. For example, face recognition and voiceprint authentications are susceptible to ambient lights and surrounding audible noises respectively, which could lead to significant performance degradations. Towards this end, we explore the liveness verification of user authentication leveraging unique patterns extracted from users' mouth movements, which cannot be forgotten and are robust to noisy environments.

When speaking, people's mouths involve in motions. Studies show that such motions present unique mouth movement patterns for different individuals [11]. This triggers our research in this work to extract behavioral patterns of mouth movements for user authentication on mobile devices. We study whether it is possible to distinguish different user's mouth movements leveraging acoustic signals, as acoustic signals have been proved feasible in sensing moving objects [12]–[14] without deploying customized hardware on mobile devices. In addition, the acoustic signals are robust to ambient light variations and surrounding audible noises. Thus, the lip reading-based user authentication can easily adapt to various environments. Meanwhile, the lip reading-based user authentication can achieve liveness verification naturally and cope with various attacks. To realize

Manuscript received May 22, 2018; revised September 15, 2018; accepted January 2, 2019; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor W. Lou. Date of publication January 23, 2019; date of current version February 14, 2019. This work was supported in part by the NSFC under Grant 61772338, Grant 61772341, Grant 61472254, and Grant 61672349, in part by STSCM under Grant 18511103002, in part by the Program for Changjiang Young Scholars in the University of China, in part by the Program for China Top Young Talents, in part by the Program for Shanghai Top Young Talents, and in part by the China Scholarship Council. *(Corresponding author: Jiadi Yu.)*

L. Lu, J. Yu, Y. Zhu, L. Kong, and M. Li are with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: luli_jtu@sjtu.edu.cn; jiadiyu@sjtu.edu.cn; yzhu@sjtu.edu.cn; linghe.kong@sjtu.edu.cn; mlli@sjtu.edu.cn).

Y. Chen is with the Department of Electrical and Computer Engineering, Rutgers University, New Brunswick, NJ 08854 USA (e-mail: yingche@scarletmail.rutgers.edu).

H. Liu is with the Department of Computer, Information and Technology, Indiana University-Purdue University Indianapolis, Indianapolis, IN 46202 USA (e-mail: hl45@iupui.edu).

Digital Object Identifier 10.1109/TNET.2019.2891733

the lip reading-based user authentication leveraging acoustic signals, we face several challenges in practice. Firstly, the subtle mouth movements need to be captured leveraging acoustic signals. Secondly, the unique behavioral patterns of users' speaking mouths should be extracted for different individuals. Thirdly, the designed authentication system needs to have the capability to accurately identify each individual. Finally, the solution should be lightweight and computationally efficient for smartphones.

In this paper, we first investigate the behavioral patterns of users' speaking mouths leveraging acoustic signals. To capture Doppler shift of acoustic signals caused by subtle mouth movements, we utilize *signal gradient* in frequency-domain to extract the reflected signals caused by mouth movements from a mixed received signal. Through analyzing Doppler profiles of acoustic signals with respect to users' speaking mouths, we find that there are unique mouth movement patterns for different individuals. Inspired by the observations, we propose a lip reading-based user authentication system, *LipPass*, which reads users' speaking mouths through acoustic sensing and extracts unique behavioral patterns of users' speaking mouths for user authentication. First, we propose a deep learning-based method, a *three-layer autoencoder-based Deep Neural Network* (DNN), to extract efficient and reliable features from Doppler profiles of users' speaking mouths under a single word. Given extracted features, *LipPass* then utilizes *softmax function* to build a mouth state identifier for identifying the user's mouth state during speaking the passphrase before user authentication. Next, *LipPass* employs *Support Vector Domain Description* (SVDD) to construct a spoofer detector based on extracted high-level features for a single-user system, which can distinguish a registered user from spoofers. Meanwhile, we also consider a multi-users authentication system to differentiate a group of users, in which users sequentially register to the system one by one. To reduce the computational complexity and improve user experience, *LipPass* constructs a binary classifier for each newly registered user through *Support Vector Machine* (SVM) to differentiate from prior registered users, and thereby develop a *balanced binary tree-based authentication approach* built upon the binary classifiers with respect to each registered user for continuous user authentication. Finally, to strengthen the reliability of the authentication results, we design a *weighted voting scheme* for user authentication by examining the speaking mouth patterns with multiple words. Our extensive experiments demonstrate that *LipPass* is reliable and efficient for user authentication in real environments.

We highlight our contributions as follows.

- We utilize signal gradient in frequency-domain to capture Doppler shift of acoustic signals caused by subtle mouth movements, and find that there are unique mouth movement patterns for different individuals.
- We propose a lip reading-based user authentication system, *LipPass*, which read users' speaking mouths through acoustic sensing and extract unique behavioral patterns of speaking mouths for user authentication.
- We design a deep learning-based method to abstract high-level behavioral characteristics of mouth movements, and employ SVM and SVDD to train binary classifiers and spoofer detectors for user identification and spoofer detection, respectively.
- We develop a balanced binary tree-based authentication approach for a multi-users system to accurately identify each individual leveraging the binary classifiers with respect to each registered user.
- We conduct experiments in four real environments. The results show that *LipPass* can achieve 90.2% accuracy on average in user identification and 93.1% accuracy in spoofer detection across different environments.

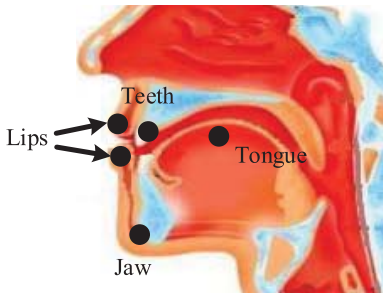

Fig. 1. Illustration of mouth movements during speaking.

The rest of this paper is organized as follows. We first show the preliminary in Section II. Then Section III presents the system design of *LipPass*. The details of multi-path mitigation are described in Section IV. The evaluation of the system is presented in Section V. Finally, we review several related work and make a conclusion in Section VI and VII.

II. Preliminary

In this section, we first describe the mouth movements during users' speaking, then present the method to capture the mouth movements leveraging acoustic signals, and finally show the relative results which validate the feasibility of utilizing Doppler profiles for user authentication.

A. Mouth Movements During Speaking

Human speaking requires precise and highly coordinated movements of multiple components in the mouth [15], including lips, teeth, jaw, tongue, etc. Specifically, mouth movements depict the connection between the lexical units with mouth behavior dynamic during users' speaking. For English speaking, the coordination among multiple components of mouth induces behaviors like lip protrusion and closure, tongue stretch and constriction, as well as jaw angle change, etc.

Fig. 1 illustrates the movements of different components in the mouth during users' speaking. Each word speaking usually involves multidimensional movements of multiple components in the mouth. For instance, the pronunciation of the word 'Hello' consists of two phonemes $[he]$ and $[l'\overline{o}]$. Speaking $[he]$ induces lips horizontal outward movements, tongue tip stretch and jaw angle change, while speaking $[l'\overline{o}]$ requires lips horizontal inward movements and tongue tip constriction. On the other hand, different users' speakings induce different mouth movements, which depict unique behavioral characteristics for different individuals [11]. Also, it is hard for a spoofer to observe the movements of many components in the mouth during speaking, such as tongue and teeth, which depicts the difficulty of imitating users' speaking mouth. Toward this end, we are motivated to capture the mouth movements during speaking and further utilize them for user authentication.

B. Capturing Mouth Movements Through Doppler Effect

Audio devices on smartphones can be exploited to build an acoustic signal field by continually emitting acoustic signals

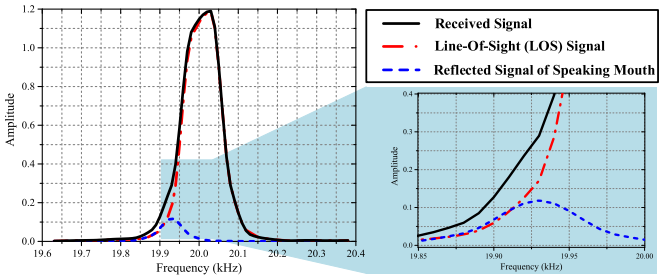

Fig. 2. An example of a mixed received signal including a LOS signal and a reflected signal from speaking mouth.

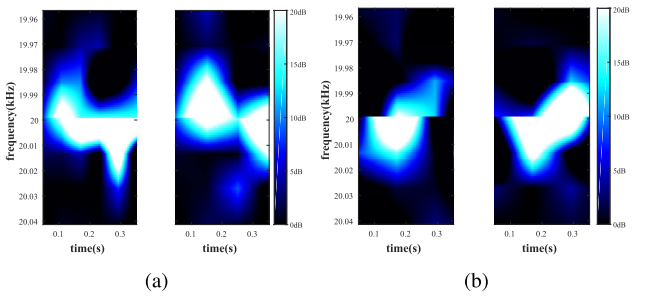

Fig. 3. Doppler profiles of acoustic signals caused by speaking the word 'Hello' under two different users. (a) User 1. (b) User 2.

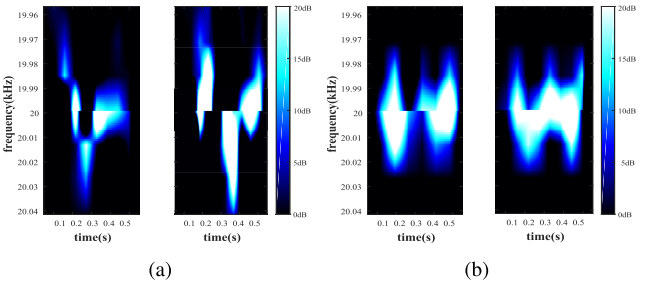

Fig. 4. Doppler profiles of acoustic signals caused by speaking the word 'World' under two different users. (a) User 1. (b) User 2.

with the speaker and receiving the signals by microphones on a smartphone. A user's mouth movements can induce Doppler effect of acoustic signals while the user speaks words. Different users exhibit subtle differences in Doppler shift of acoustic signals while speaking the same words. We are motivated to utilize Doppler effect of acoustic signals to capture the unique behavioral patterns of a user's speaking mouth and perform user authentication on smartphones.

Doppler effect depicts the frequency change caused by the movements of objects relative to the signal source. Specifically, an object moving at speed v relative to the acoustic signal source brings a frequency change:

$$\Delta f = \frac{v}{c} \times f_0, \tag{1}$$

where c and f_0 are the speed and frequency of the acoustic signal respectively. Since a higher frequency results in a more discernible Doppler shift confined by Eq. (1), and most smartphone speaker systems can only produce acoustic signals at up to $20kHz$, we select $f_0 = 20kHz$ as our frequency of pilot tone, which is also out of the humans' auditory perceptual range. We sample the raw data on smartphones at the rate of $44.1kHz$, which is the default sampling rate of acoustic signals under $20kHz$. Then, the original received signals are transformed into frequency-domain signals by performing the 2048-points Fast Fourier Transform (FFT), which achieves a high frequency resolution with an appropriate computational complexity.

Since the speaker and microphone are both integrated in a smartphone, in the received signals, the attenuation of the Line-Of-Sight (LOS) signal (i.e., the signal directly propagated from the speaker to microphone) is far less than that of the reflected signals by objects. Moreover, since the speed of users' speaking mouths is much slower, the corresponding Doppler shift will lie in the frequency band of the LOS signals. Fig. 2 shows an example of a mixed received signal including a LOS signal and a reflected signal from speaking mouth. We can see that, in the received signal, the reflected signal caused by speaking mouth is buried within the LOS signal.

In order to capture Doppler shift of acoustic signals caused by subtle mouth movements, we employ *signal gradient* of received signals in frequency-domain, which denotes the difference of the frequency-domain signals between two successive time slots. Assume a user is stationary and the speaking mouth are the sole moving objects in the authentication scenario. The received signal $s_{(f)}(t)$ consists of the LOS signal, the reflected signal from speaking mouth, the reflected signals from surrounding static objects (e.g., furnitures), and the environmental noises, i.e.,

$$s_{(f)}(t) = s^{e}_{(f)}(t) + s^{rl}_{(f)} + \sum_{i} s^{rs}_{(f)i}(t) + n(t), \tag{2}$$

where $s^{e}_{(f)}(t)$ is the LOS signal in time slot t, $s^{rl}_{(f)}(t)$ is the reflected signal from speaking mouth in time slot t, $s^{rs}_{(f)i}(t)$ is the i^{th} reflected signal from static objects in time slot t, and $n(t)$ is the white noise in the surrounding. Since the smartphone steadily emits a predefined signal from the speaker, and the distance between the speaker and microphone is fixed in a smartphone, the LOS signal is invariant along the time. Also, users are stationary in the authentication scenario, so the reflected signals from static objects are invariant along the time. Thus, the signal gradient of received signals in frequency-domain from time slot $t-1$ to t, $g(t)$, is:

$$\begin{aligned} g(t) &= s_{(f)}(t) - s_{(f)}(t-1) \\ &= s^{rl}_{(f)}(t) - s^{rl}_{(f)}(t-1) + n(t) - n(t-1). \end{aligned} \tag{3}$$

The gradient matrix $G = [g(1), g(2), \cdots, g(T)]$ can represent Doppler profiles of speaking mouth within a duration T.

C. Unique Behavioral Characteristics of Mouth Movements

To demonstrate whether the signal gradient of received acoustic signals can capture the subtle mouth movements during users' speaking, and further extract unique behavioral characteristics of these mouth movements, we conduct an experiment, in which 12 volunteers are required to speak 10 most frequent words [16], and a smartphone is used to emit $20kHz$ acoustic signals and receive acoustic signals reflected by speaking mouths under a sampling rate of $44.1kHz$. Based on received acoustic signals, we analyze Doppler profiles during these speakings to validate the feasibility of capturing mouth movements through signal gradient.

Fig. 3 and 4 show two Doppler profile examples caused by speaking two words (i.e., 'Hello' and 'World') from two different users respectively. Compare Fig. 3(a) with 3(b),

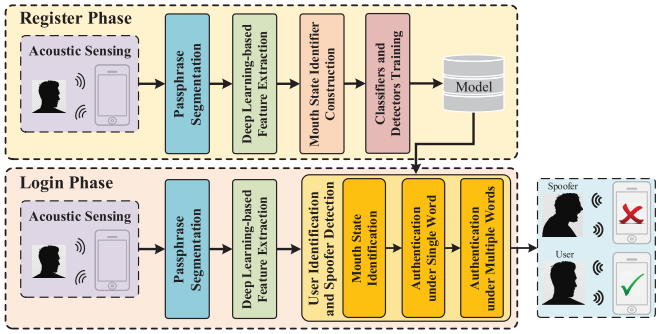

Fig. 5. System architecture of *LipPass*.

we observe that Doppler profiles of speaking the word 'Hello' exhibit different variation trends between the two users. Fig. 4(a) and 4(b) show the similar results. Additionally, speaking the same word by the same user produces similar Doppler profiles. Other experimental results are similar to these Doppler profiles. These encouraging results demonstrate the great potential that Doppler effect of acoustic signals caused by users' speaking mouths can be used in user authentication.

III. System Design

In this section, we present the design of lip reading-based user authentication, *LipPass*, which reads users' speaking mouths through acoustic sensing and capture the unique behavioral patterns of mouth movements for user authentication.

A. Overview

Fig. 5 shows the system architecture of *LipPass*, which includes two phases - the register phase and login phase.

In the register phase, a user speaks a passphrase including several words several times. Meanwhile, a smartphone continually emits predefined ultrasonic acoustic signals and receives the acoustic signals reflected from users' speaking mouths. First, *LipPass* segments the received signals of the passphrase into several episodes, each of which represents a single word. Then, *LipPass* extracts efficient and reliable features from the signal episodes leveraging a deep learning-based method. Next, based on extracted features, *LipPass* constructs a mouth state identifier through softmax function to identify the mouth state during speaking the passphrase. Finally, for each mouth state, *LipPass* employs Support Vector Machine and Support Vector Domain Description to construct binary classifiers and spoofer detectors based on extracted high-level features for user identification and spoofer detection respectively.

In the login phase, *LipPass* first captures reflected signals when a user speaks the same passphrase as that in the register phase, then performs passphrase segmentation and feature extraction. Based on extracted features, in user authentication, *LipPass* first identifies the user's mouth state during speaking the passphrase through the trained mouth state identifier. Then, *LipPass* applies a balanced binary tree-based authentication approach to verify the user whether a registered user or spoofer leveraging the trained binary classifiers and spoofer detectors with respect to registered users. Finally, *LipPass* employs a weighted voting scheme for user authentication by examining mouth movement patterns with multiple words.

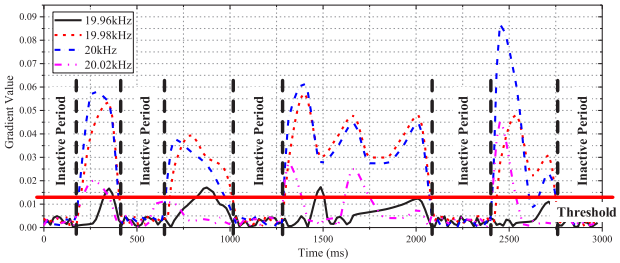

Fig. 6. Doppler profiles of mouth movements when a user speaks four words under four frequencies.

B. Passphrase Segmentation

In both register and login phases, a user speaks a passphrase including several words, and the smartphone receives the acoustic signals reflected by the user's speaking mouth. *LipPass* first segments the received signals of the given passphrase into episodes, each representing a single word. According to [17], there is usually a short interval (e.g., 300 ms) between speaking two successive words. Fig. 6 shows Doppler profiles of mouth movements when a user speaks four words under four frequencies, which are the largest four ones among all Doppler profiles. It can be observed from the figure that the intervals between arbitrary two words are significant. *LipPass* regards each interval between two words as an inactive period. Through empirical studies, Doppler profiles in an arbitrary inactive period are all less than a threshold. Thus, *LipPass* uses a sliding window with a fixed step to detect all inactive periods in a passphrase so as to segment the passphrase. The threshold can be set as the mean value of the noises in the surrounding.

C. Deep Learning-Based Feature Extraction

As mentioned in Section II, we find that the contour of Doppler profiles can depict unique behavioral features during speaking, such as the continuous variations in opening degrees of mouth, etc. Thus, it is straightforward to use the complete Doppler profiles as features for user authentication based on speaking behaviors. However, due to requirement of high resolution in both frequency and time of Doppler profiles to depict the behaviors, the profiles are usually high-dimensional, which induces curse of dimension problem. Hence, it is necessary to extract features from the high-dimensional profiles.

Traditional feature extracting methods abstract features by observing the unique patterns manually. Features extracted by these methods usually have redundant information and are poor in robustness. Although some linear feature extraction approaches (e.g., PCA or LDA) can achieve preferable features by generating the linear decision boundaries [18], Doppler profiles of users' speaking mouths are usually non-linear separated. Therefore, we develop a deep learning-based method, a *three-layer autoencoder-based Deep Neural Network* (DNN) [19], to extract efficient and reliable features from Doppler profiles of users' speaking mouths.

In the proposed three-layer DNN model, each hidden layer consists of an autoencoder network which abstracts the input features as a set of compressed representations in an unsupervised manner. Such compressed representations are able to characterize unique behavioral patterns of users' speaking mouths. The autoencoder can map the input X into a set of compressed representation C as $C = \sigma(wX + b)$, where $\sigma()$

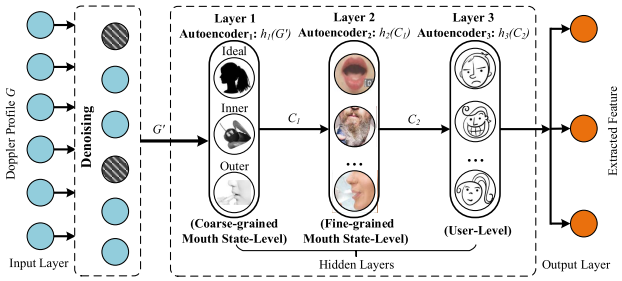

Fig. 7. Architecture of feature extraction through a three-layer autoencoder-based Deep Neural Network.

is a logistic function defined as $\sigma(x) = \frac{1}{1+e^{-x}}$, w and b are the weight and bias of the autoencoder network respectively. The autoencoder is trained with the objective as follows:

$$\min DIF(X, X') = \min \frac{1}{N} \sum_{i=1}^{N} (X^{(i)} - X'^{(i)})^2 + \lambda \Omega_{weights} + \beta \Omega_{sparsity}, \quad (4)$$

where N is the number of training samples, $X^{(i)}$ and $X'^{(i)}$ are the i^{th} element in the original input X and reconstructed input X', $\Omega_{weights}$ and $\Omega_{sparsity}$ are the L_2 regularizer for the parameters and sparsity, and λ as well as β are the coefficients of the two L_2 regularizers. The objective minimizes the differences between the original input X and a relative reconstructed input X', where $X' = \sigma(w^T C + b')$. Such an objective ensures the compressed representation C can abstract most of the original input X's information.

Fig. 7 shows the architecture of feature extraction through a three-layer autoencoder-based DNN model. Given Doppler profiles, $G = [g(1), g(2), \cdots, g(T)]$, of a user's speaking mouth within a duration time T, where $g(t)$ is the signal gradient of received signals in time slot t ($t \in [1, T]$), each layer of DNN model contains an autoencoder $h_i (i = 1, 2, 3)$, which encodes the input into a set of compressed representations as output. To ensure the extracted features robust enough for classification, *LipPass* first applies the denoising autoencoder [19] to denoise Doppler profiles G of users' speaking mouths as the input of DNN model. The input of the first layer is the denoised Doppler profiles G' of users' speaking under a single word, and the autoencoder $h_1(G')$ in the first layer can extract the coarse-grained mouth state-level features C_1 as the outputs, which represent whether the mouth is in ideal state or not. Except for ideal state, C_1 divides unideal mouth state into inner and outer mouth states, which depict unideal states occurred inner (e.g., eating) and outer (e.g., shaving beard) the mouth respectively. Then, the output C_1 of the first level is fed to the second layer. The autoencoder $h_2(C_1)$ in the second layer further extracts the fine-grained mouth state-level features C_2, which depicts the details of mouth state (e.g., eating, shaving beard, etc.). Finally, the autoencoder $h_3(C_2)$ in the last layer takes the output C_2 of the second layer as input, and extracts the user-level features, which represent the unique patterns of a user and can be used for user authentication.

Fig. 8 shows two reconstructed profiles of a user speaking the word 'World' based on extracted features of mouth movements. Compare with the original Doppler profile as shown in Fig. 8(a), we can observe that both reconstructed profiles in Fig. 8(b) and 8(c) can recover basic features from the original Doppler profile, and the reconstructed result with denoising shows more significant features than that without denoising.

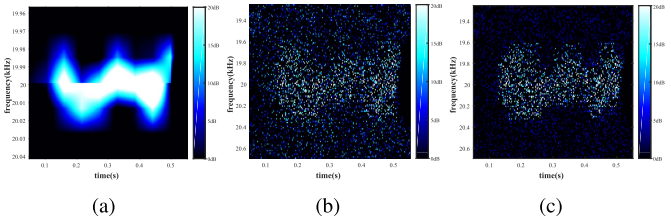

Fig. 8. Reconstructed profiles based on extracted features of mouth movements. (a) Original. (b) Without denoising. (c) With denoising.

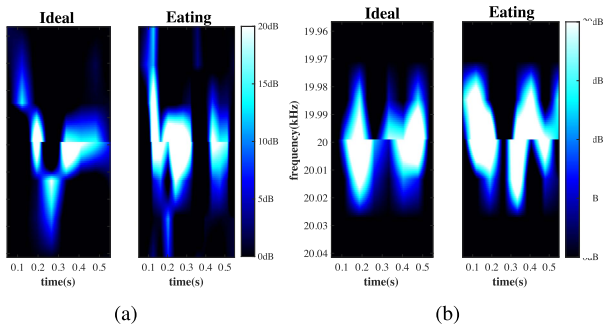

Fig. 9. Doppler profiles of acoustic signals caused by speaking the word 'World' under ideal and eating mouth states of different users respectively. (a) User 1. (b) User 2.

D. Mouth State Identifier Construction

In Section II, we reveal the feasibility of utilizing Doppler profiles caused by speaking mouth for user authentication. In the experiment, the volunteers are required to keep their mouths in an ideal state, i.e., volunteers do not eat or drink something, and do not conduct behaviors about their mouths (e.g., shave the beard) during speaking the passphrase. However, in real environments, a user may not keep his/her mouth in the ideal state during the login phase. The unideal mouth state would induce Doppler profiles which are different from that under ideal mouth state. Fig. 9 shows Doppler profiles of acoustic signals caused by speaking the word 'World' under ideal and eating mouth states of different users respectively. We can observe that Doppler profiles under different mouth states of the same user present significant difference, which is quite different from that in Fig. 4. This result indicates that the mouth state has significant impact on Doppler profiles, which further affects user authentication based on Doppler profiles. Thus, it is necessary to identify the mouth state before authenticating users.

Since higher level compressed features are more stable and robust to small-scale input variations [20], we construct a mouth state identifier based on extracted fine-grained mouth state-level features through *softmax function* [21] to identify the mouth state. Softmax function is a generalization of the logistic function, which is used to handle multi-label classification. Specifically, given a fine-grained mouth state-level feature C, the posterior probability $P(M_i|C)$ of mouth state M_i can be derived through the softmax function as follows:

$$P(M_i|C) = \frac{P(C|M_i)P(M_i)}{\sum_{j=1}^{K} P(C|M_j)P(M_j)}, \quad (5)$$

where K is the number of mouth states, $P(M_i)$ is the prior probability of the same mouth state M_i, and $P(C|M_i)$ is the likelihood of fine-grained mouth state-level feature C under the mouth state M_i. The outputs of softmax function characterize the probability distribution over K profiled mouth state (e.g., eating, drinking, shaving beard, etc.). During the register phase, a user is required to speak the same number of training samples for each mouth state, and thus the prior probability $P(M_i)$ obeys the uniform distribution for all mouth states M_i. The posterior probability $P(M_i|C)$ can also obtained through statistics of training samples. Based on the two probabilities, our system can derive the likelihood $P(C|M_i)$ based on Eq. (5) to train the mouth state identifier in the register phase. In the login phase, based on the extracted fine-grained mouth state-level feature C, the mouth state identifier can calculate the posterior probabilities $P(M_i|C)$ for all mouth states $i \in \{1, \cdots, K\}$, and further regards the mouth state as M_k, where $k = \arg\max_i P(M_i|C)$.

E. User Authentication Classifiers and Detectors Training

After the mouth state identifier is constructed, *LipPass* can identify the mouth state during users' speaking before authenticating a user's identity. To further authenticate users and detect spoofers under each mouth state, we employ *Support Vector Machine* (SVM) [22] to train classifiers and detectors based on extracted user-level features from Doppler profiles of users' speaking mouths through DNN model.

For a single-user system, when a user registers to *LipPass*, the user is required to speak a predefined passphrase several times and provides the relative mouth state during the speakings, so *LipPass* can extract the user's unique features from Doppler profiles of the user's speaking mouth under a specific mouth state as training data. Since we only have the user's training data while lack of spoofers' training data, we apply a special version of SVM, i.e., *Support Vector Domain Description* (SVDD) [23], to train a spoofer detector only using one-class data, i.e., the user's training data, which can distinguish the user from spoofers.

Moreover, it is possible for multiple users to access their private information on a system. Thus, it is necessary to verify a user's identity in a multi-users system. In the register phase, users sequentially register to the authentication system one by one. Since multi-classes classifier construction induces significant computational complexity, it is inappropriate for an authentication system to reconstruct a multi-classes classifier whenever a new user registers to the system. Thus, in order to reduce the computational complexity and improve user experience in the register phase, we employ SVM to train a binary classifier for each user. Assume $(n-1)$ users (i.e., $U_1, \cdots, U_{n-1}$) have registered in the authentication system, and the n^{th} user, U_n, is registering to the authentication system. *LipPass* first divide the n users' data into two-class data, i.e., $n^{th} \sim \lfloor n/2 \rfloor^{th}$ users' data and $\lfloor n/2 \rfloor$–$1 \sim 1^{st}$ users' data, and then employs SVM to train a binary classifier based on the two-class data, which can identify whether a user is one of the posterior $\lfloor n/2 \rfloor$ users. By analogy, *LipPass* further divide the posterior $n/2$ users' data into two-class data and employs SVM to train a classifier, which can identify whether a user is one of the posterior $\lfloor n/4 \rfloor$ users. In a multi-users system, *LipPass* would train $\lceil \lg n \rceil$ binary classifiers for the n^{th} registered user to verify the user's identity. Furthermore, *LipPass* trains a spoofer detector based on the n^{th} user's data through SVDD to distinguish spoofers from the n^{th} user. All binary classifiers and spoofer detectors will be used to authenticate users.

F. User Identification and Spoofer Detection

In the login phase, *LipPass* usually requires users to speak a passphrase including several words, and users may speak the passphrase under different mouth states (e.g., eating, etc.). *LipPass* first identifies the user's mouth state during speaking the passphrase. Then, *LipPass* utilizes relative classifiers and detectors based on identified mouth state to authenticate each individual and detects spoofers under each word. Finally, based on authentication results under single words, *LipPass* achieves the final authentication result under multiple words.

1) Mouth State Identification: LipPass constructs the mouth state identifier through the softmax function in the register phase. Thus, in the login phase, *LipPass* identifies the user's mouth state during speaking passphrase through the constructed identifier.

As mentioned in Section III-D, *LipPass* trains the mouth state identifier through calculating all the likelihood $P(C|M_i)$ under K profiled mouth states. When a user speaks the passphrase to login into the system, the fine-grained mouth state-level features C are extracted through DNN model and fed to the identifier. The mouth state during the speaking can be identified through solving the optimization problem:

$$\begin{aligned} k = \arg \max_{i \in [1,K]} \; & P(M_i|C), \\ s.t. \; & 0 < P(M_i|C) \leq 1, \\ & \sum_{i=1}^{K} P(M_i|C) = 1. \end{aligned} \tag{6}$$

The objective is to find a mouth state k with the maximum posterior probability. The first constraint means each posterior probability is in the range of $(0, 1]$. The second one indicates that given a fine-grained mouth state feature C, the sum of posterior probabilities under all profiled mouth states should equal 1. To solve the optimization problem, the mouth state identifier first calculates all posterior probability $P(M_i|C)$ ($i \in [1, K]$) under the extracted feature C based on Eq. (5), which satisfy the two constraints in Eq. (6). Then, the mouth state with the maximum posterior probability would be identified as the mouth state during the speaking.

To construct such a mouth state identifier, users need to speak the passphrase under each mouth state, which depicts a large number of training samples, and leads to significant user experience degradation in register phase. To deal with the problem, *LipPass* would ask users to provide training samples for their own daily mouth states, i.e., users can define the value of K by themselves, and further provide relative training samples for register. Usually, there are only a few different mouth states when a user passes the user authentication of a smartphone, which depicts a small value of K. Therefore, *LipPass* would not require users to speak much times in register phase.

2) Authentication Under Single Word: After the mouth state during speaking the passphrase is identified, *LipPass* utilizes relative classifiers and detectors to authenticate users and detect spoofers. In the register phase, for a user U_i in a n users system, *LipPass* trains $\lceil \lg i \rceil$ binary classifier based on U_i's features and prior $(i-1)$ registered users' features to verify

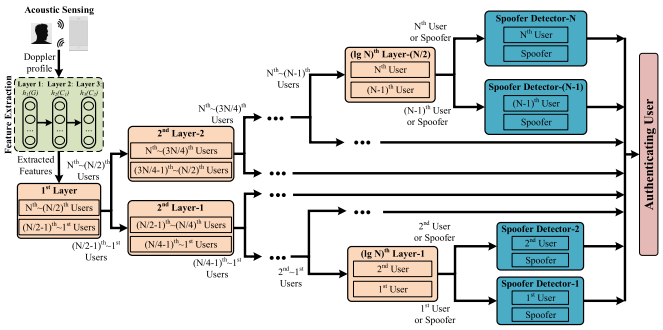

Fig. 10. Architecture of the balanced binary tree-based authentication under single word.

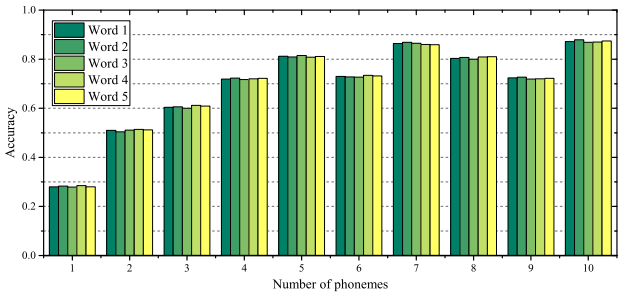

Fig. 11. Relationship between authentication accuracy and the number of phonemes in a single word.

whether the user is the i^{th} user or one of prior $(i-1)$ registered users. Since the i^{th} classifier is trained without any data about the subsequent registered users (i.e., $U_{i+1}, U_{i+2}, \ldots, U_n$) and spoofers, the user could be U_i, one of the subsequent registered users (i.e., $U_{i+1}, U_{i+2}, \ldots, U_n$) or a spoofer if the i^{th} classifier verifies a login user as U_i. Thus, in the login phase, we propose a *balanced binary tree-based authentication approach* to verify users' identities and detect spoofers. Fig. 10 shows the architecture of the balanced binary tree-based authentication under single word.

Assume there are n users registered in a system. When a user logins to the system, *LipPass* first collects Doppler profiles of acoustic signals caused by the user's speaking mouth, and then segments received acoustic signals into episodes, as well as extracts features of the user's speaking mouth from the episodes through DNN model. *LipPass* organizes the trained binary classifiers as a $\lceil \lg n \rceil$-layer balanced binary tree, and adds all spoofer detectors to the tree as leaves, as shown in Fig. 10. Based on the 1^{st} layer, *LipPass* verifies whether the user is one of the posterior $\lfloor n/2 \rfloor$ registered users. If the classifier identifies the user as one of the posterior $\lfloor n/2 \rfloor$ users, *LipPass* would feed the user's extracted features to the second classifier in the 2^{nd} layer, which will further verify whether the user is one of the posterior $\lfloor n/4 \rfloor$ users. On the contrary, if the n^{th} classifier identifies the user as one of prior $\lfloor n/2 \rfloor$ registered users, the extracted features are further fed to the first classifier in 2^{nd} layer. By analogy, *LipPass* would feed the extracted features through one path in the balanced binary tree, and further regard the login user as the i^{th} registered user. To avoid the user is spoofer, the extracted features would be finally fed to a spoofer, which is trained by the i^{th} registered user's data, to validate whether the login user is the registered user or not. Therefore, *LipPass* is able to accurately identify a login user as a registered user or spoofer.

3) Authentication Under Multiple Words: To strengthen the robustness of the authentication result, *LipPass* verifies users' identities and detects spoofers under several words. We propose a *weighted voting scheme* to achieve the final user authentication result under multiple words.

For different words, the number of phonemes are different, which brings different amount of behavioral patterns from speaking mouths. Thus, the authentication accuracies under different number of a word's phonemes may exhibit considerable differences. To exploit the relationship between the authentication accuracy under single word and the number of a word's phonemes, we conduct an extensive experiment under 20 volunteers, which includes 10 males and 10 females. Each volunteer in the experiment is asked to speak several words, whose the number of phonemes varies from 1 to 10. For each number of phonemes, we select 5 most frequent words from Word Frequency Data [16]. For each word, we ask each volunteer to speak it 3 times for the register phase and perform 12 legitimate authentications in the login phase. Fig. 11 shows that the relationship between authentication accuracy and the number of phonemes. We can observe that authentication accuracies under different number of phonemes exhibit significant differences, while the authentication accuracies under the same number of a word's phonemes are almost the same. Therefore, we can utilize the authentication accuracies under different number of a word's phonemes as weights to measure the reliability of authentication results.

Assume the given passphrase includes m words. Through the authentication under single word, *LipPass* can verify a user's identity and obtain m relative authentication results (i.e., $L_1, \cdots, L_m$). Then, based on the m authentication results and relative m weights, i.e., $\{w_1, w_2, \ldots, w_m\}$, we define the confidence of a user U_i as follows:

$$conf_i = \sum_j w_j, \quad j \in \{k | L_k = U_i\}. \tag{7}$$

Based on the confidences of the registered users and the spoofer, *LipPass* can identify a user as the registered user with maximum confidence.

G. Computational Complexity Analysis

Assume there are N registered users, and K profiled mouth states in the system. When a user logins to the system, the mouth movements are sensed by acoustic signals with an amount of M. The computational overhead of *LipPass* can be analyzed as follows.

Model Building: The model building process consists of mouth state identifier construction, user authentication classifiers and detectors training. Since the models are only built once for a user during register phase, the model building would not contribute to the computational complexity.

Passphrase Segmentation: LipPass utilizes a sliding window with a fixed step to detect all inactive periods of received acoustic signals during speaking a passphrase. Hence, the computational complexity of passphrase segmentation is $O(M/w)$, where w is the length of the fixed step.

Feature Extraction: LipPass constructs a three-layer denoising autoencoder-based DNN to extract fine-grained mouth state-level and user-level features. Similar to model building,

DNN is built only once during register phase. To improve the user experience in the register phase, we enable our system to run in front end only when collecting users' speaking passphrase. As for DNN model construction, the system would run in background so that users are unaware with the high computational complexity during the register phase. Then, we further consider the computational complexity of feature extraction through trained DNN during login phase. The computational complexity of feature extraction is $O((1 + c_1c_2d)M)$, where c_1, c_2 and d are the lengths of coarse-grained mouth state, fine-grained mouth state and user-level features respectively.

User Authentication: The user authentication consists of mouth state identification, authentication under single word and multiple words. For mouth state identification, the computational complexity of the softmax-based mouth state identification is $O(K)$, where K is the number of profiled mouth states. Then, for authentication under single word, the computational overhead of the balanced binary tree-based authentication approach is $O(\lg N)$, where N is the number of registered users. Finally, since authentication under multiple words is based on that under single word, the overhead is $O(m \lg N)$, where m is the number of words in a passphrase. Thus, the computational complexity of user authentication is $O(K + m \lg N)$.

Based on the analysis above, the computational complexity of *LipPass* in login phase is $O((1 + \frac{1}{w} + c_1c_2d)M + K + m \lg N)$. Since w, c_1, c_2, d and m are fixed in the design of *LipPass*, the computational complexity can be considered as $O(M + K + \lg N)$. Usually, there are few users sharing the ownership of a system and registering in the system on smartphones, and few mouth states during users' speaking passphrase, which indicates the values of N and K are small. Thus, our authentication approach is lightweight and computationally efficient for smartphones.

IV. Multi-Path Interference Mitigation

LipPass utilizes acoustic signals to read users' speaking mouths for user authentication. In Section II, we assume users are stationary and their speaking mouths are the sole moving objects in the authentication. However, the acoustic signals are vulnerable to multi-path interferences from users' body movements and static objects in the surrounding. Thus, it is necessary to eliminate the multi-path interferences in the authentication.

A. Eliminating Multi-Path Interferences From Normal Body Movements

In practice, users' speaking mouths are not the sole moving objects in authentication scenarios, and there are usually other body movements, such as walking, stretching out hand, and some environmental audible voices, which affect the received signals. However, Doppler shift caused by these motions are quite different from that caused by users' speaking mouths. The normal body movements induce a Doppler shift ranging in $[50, 200]Hz$ [24], and Doppler shift of audible voices ranges in $[500, 2000]Hz$. However, Doppler shift caused by users' speaking mouths is in $[-40, 40]Hz$. Thus, we apply a *Butterworth Band-Pass Filter* [25] to obtain the target frequency band, i.e., $[f_0 - 40, f_0 + 40]Hz$, for speaking mouth detection, and eliminate other out-band interferences.

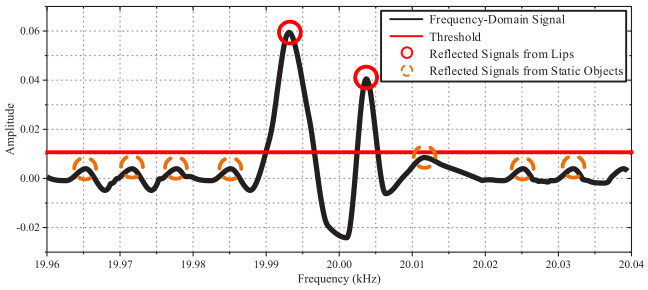

Fig. 12. An example of the reflected signals from speaking mouth and static objects.

B. Filtering Multi-Path Interferences From Minute Body Movements

Except for normal body movements, there are also some minute body movements inducing Doppler shift still ranging in $[-40, 40]Hz$, such as subtle vibrations of users' fingers, etc. Since this kind of interferences is usually unpredictable, we cannot construct a filter with fixed parameters (e.g., barrier frequency) in advance to remove the interferences. Thus, we adopt an adaptive filter named *Fast Transverse Recursive Least Square* (FTRLS) algorithm [26] to adaptively adjust the filter's parameters and remove the interferences induced by minute body movements. Specifically, FTRLS algorithm includes three basic steps: 1) forward estimation, 2) backward estimation and 3) joint process estimation. In the first step, FTRLS obtains a update factor through calculating the prior error of the adaptive filter. Then, based on the update factor, a conversion factor can be further obtained through calculating the posteriori error of the adaptive filter. The update factor and conversion factor are used to estimate objective errors under current parameters for the joint process estimation. Based on the two factors, FTRLS calculates the prior error and updates the weight in the adaptive filter. The computation complexity of FTRLS algorithm is $O(7N)$ [26], where N is the filter order. Hence, the convergence rate of FTRLS algorithm is fast without complex parameter selection.

C. Removing Multi-Path Interferences From Static Objects

When users authenticate through *LipPass*, except for reflected signals from users' mouths, there are other reflected signals from static objects, such as furnitures. Since users are usually not stationary in fact, the reflected signals from static objects are variant with time. Thus, the reflected signals from static objects would also interfere with the reflected signals from users' speaking mouths. Thus, it is necessary to remove the reflected signals from static objects in received signals.

Usually, in the authentication scenario, users' mouths are close to the smartphone (e.g., less than 10 cm), while the distances between static objects and the smartphone are far longer. Thus, the amplitude of reflected signals from static objects is far lower than that from mouths. Fig. 12 shows an example of reflected signals from speaking mouth and static objects. We can observe that the amplitudes of two reflected signals from mouths are far larger than that of other reflected signals from static objects. Thus, we adopt a threshold-based approach to remove reflected signals from static objects, and the threshold can be selected through empirical studies.

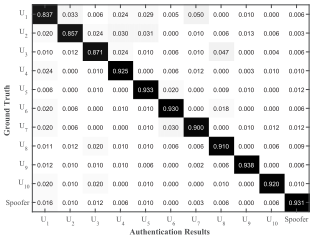

Fig. 13. Confusion matrix of *LipPass*, each entry of which is the average value in four different environments.

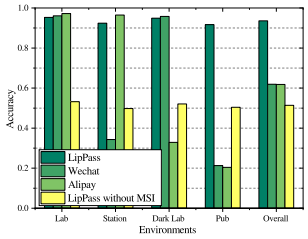

Fig. 14. Authentication accuracy of *LipPass*, Wechat, Alipay and *LipPass* without mouth state identification (MSI).

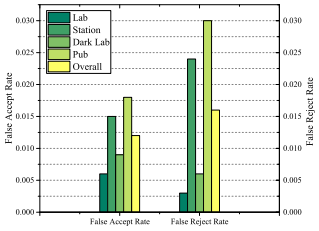

Fig. 15. False accept rate and false reject rate of *LipPass* in four different environments.

V. Evaluation

In this section, we evaluate the performance of *LipPass* under the collected data from 48 volunteers in four different real environments.

A. Experiment Setup and Methodology

We evaluate *LipPass* with four types of smartphones, i.e., a Nexus 6P, a Galaxy S6, a Galaxy Note 5, and a Huawei Honor 8. Our experiments are conducted under 4 different environments, i.e., a laboratory (bright and quiet), a train station (bright but noisy), a dark laboratory (quiet but dark), and a pub (dark and noisy). In each environment, we randomly select 12 volunteers, including 6 males and 6 females whose ages range from 18 to 52, for experiments. Among the 12 volunteers, 10 of them register in the system with *LipPass* while the rest two volunteers as spoofers. During each experiment, each volunteer randomly selects a smartphone and holds the smartphone with the microphone directed towards the mouth. The distance between microphone and volunteer's mouth ranges in $[5, 30]\,cm$. We predefine 10 passphrases, each of which contains 1-8 words. In each passphrase, we select words with the number of phonemes larger than 4. Each volunteer speaks the 10 predefined passphrases 3 times to register in the authentication system, and performs 12 times legitimate authentications for each passphrase.

To evaluate the performance of *LipPass*, we define four metrics as follows,

- *Confusion Matrix*: Each row and each column of the matrix represent the ground truth and the authentication result of *LipPass* 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 for all samples that actually are the i^{th} user.
- *Authentication Accuracy*: The probability that a user who is U is exactly authenticated as U.
- *False Accept Rate*: The probability that a user not a registered user is authenticated as a registered user.
- *False Reject Rate*: The probability that a user not a spoofer is authenticated as a spoofer.

B. Overall Performance

We first evaluate the overall performance of *LipPass* through confusion matrix. Fig. 13 shows the confusion matrix of *LipPass*, each entry of which is the average value in four different environments. We can see that *LipPass* can achieve over 83.7% accuracy in identifying the registered users. The average accuracy of *LipPass* in user identification is 90.2% with a standard derivation of 3.5%, and the average accuracy in spoofer detection is 93.1%.

We also compare the performance of *LipPass* with that of Wechat voiceprint lock and Alipay face recognition login. Fig. 14 shows the authentication accuracies of *LipPass*, Wechat voiceprint lock and Alipay face recognition login in four different environments respectively. It can be observed that the authentication accuracy of *LipPass* is 95.3%, which is similar to that of 96.1% and 97.2% under voiceprint lock and face recognition login in the laboratory respectively. Moreover, the accuracies of *LipPass* are 95.3%, 92.4%, 94.9% and 91.7% in the four environments respectively, which means the differences of *LipPass*' accuracies are insignificant in different environments. On the contrary, Wechat voiceprint lock and Alipay face recognition login suffer significant performance degradation in some environments. For voiceprint lock, the accuracies decrease to 34.3% and 21.3% in noisy environments respectively, i.e., the train station and pub. For face recognition login, the accuracies decrease to 32.9% and 20.4% in dark environments respectively, i.e., the dark laboratory and pub. Furthermore, the overall authentication accuracy of *LipPass* without mouth state identification is 51.4%, which is far less than that of *LipPass*, i.e., 95.3%. This demonstrates that the mouth state has certain impact on the lip reading-based user authentication, and the proposed mouth state identification contributes to the accurate user authentication.

We further evaluate the reliability and user experience of *LipPass* through the false accept and false reject rates. Fig. 15 shows the false accept rates and false reject rates of *LipPass* in four different environments. We can see that the false accept rates are all less than 2%, and the overall false accept rate is 1.2%, which demonstrates that *LipPass* can defend spoofing attacks and is reliable enough. Additionally, it can be seen from Fig. 15 that the false reject rates are all less than 3%, and the overall false reject rate is 1.6%, which demonstrates that *LipPass* can accurately identify a registered user.

We also evaluate the user experience through the speaking times for successful login. Fig. 16 shows CDF of the speaking times for successful login in four different environments. We can see that 95% of users can successfully login to the

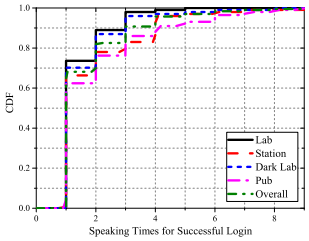

Fig. 16. CDF of the speaking times for successful login in four different environments.

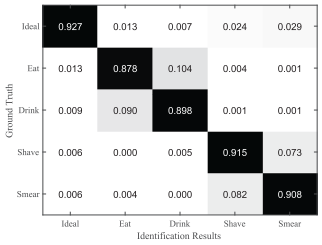

Fig. 17. Confusion matrix of mouth state identification, each entry of which is the average value in four different environments.

system through speaking a passphrase less than 4 times, which is acceptable for users in real environments.

C. Performance of LipPass in Mouth State Identification

Except for the performance of user authentication, we also evaluate the performance of *LipPass* in mouth state identification. We select two unideal states inner the mouth (i.e., eat and drink), and two unideal states outer the mouth (i.e., shave beards and smear face) during speaking passphrases to conduct the experiment. Note that during each experiment, there is no constraint on the unideal mouth states (e.g., the type of food for eating is not constrained, etc.). During the register phase, each volunteer provides the mouth state during the speaking as label for identifier construction. Based on the constructed identifier, *LipPass* identifies the volunteer's mouth state during speaking the passphrase. The identification accuracy is defined as the probability that a mouth state during a speaking which is M_i is exactly identified as M_i.

Fig. 17 shows the confusion matrix of mouth state identification, each entry of which is the average value in four different environments. We can see that the identification accuracies of five mouth states are all above 85%, which indicates *LipPass* can accurately identify the mouth state during speaking. Moreover, it can be observed that the identification accuracies of two inner mouth states (i.e., eat and drink) are a little less than that of other three mouth states. This is because the frequent mouth opening and closing under the two inner mouth states obstruct acoustic signals to continuously capture the movements of some mouth components inner the mouth (e.g., teeth and tongue). We also evaluate the accuracy of mouth state identification in different environments, as shown in Fig. 18. It can be observed that the identification accuracies are all above 85% in four environments, which depicts that the mouth state identification of *LipPass* is also not sensitive to environments.

D. Impact of Distance Between Microphone and Users' Mouths

Since we utilize acoustic signals to capture users' speaking mouths, the signal attenuation cannot be avoided. A longer

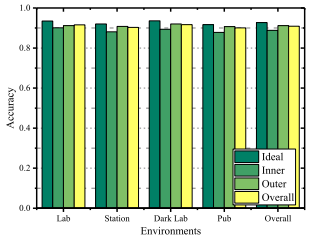

Fig. 18. Accuracy of mouth state identification in four different environments.

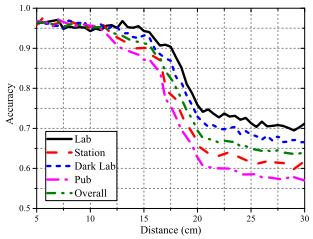

Fig. 19. Relationship between authentication accuracy and distances from microphone to users' mouths in four different environments.

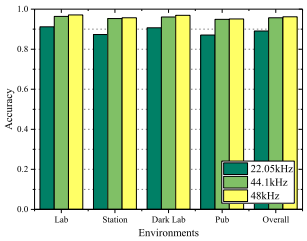

Fig. 20. Authentication accuracy of *LipPass* under different sampling rates.

distance between the microphone and users' mouths may bring a significant signal attenuation of the reflected signals, and further leads to a performance degradation of the authentication system. We enable smartphones to measure the distance between users' mouths and the microphone through Time of Arrival (ToA). Fig. 19 shows the relationship between the authentication accuracy of *LipPass* and distance from the microphone to users' mouths in four different environments. We can observe from the figure that the authentication accuracy of *LipPass* decreases as the distance increases. This is because the signal attenuation of reflected signals from speaking mouth becomes larger as the distance between the microphone and users' mouths increases. However, the authentication accuracies in all four environments can achieve 95% authentication accuracy as the distance less than $12cm$.

E. Impact of Sampling Frequency

Since not all smartphones are in-built with latest audio devices, old smartphones only support to receive acoustic signals under standard sampling frequencies at $22.05kHz$ or $44.1kHz$. We evaluate the impact of different sampling frequencies on *LipPass*. Fig. 20 shows the authentication accuracy of *LipPass* under different sampling frequencies. We can see that *LipPass* does not suffer significant performance degradation under a lower sampling frequency. Under a sampling frequency of $22.05kHz$, *LipPass* can still

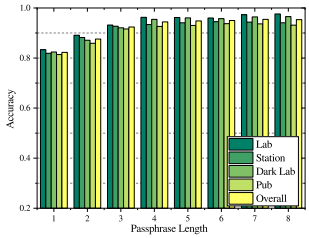

Fig. 21. Authentication accuracy of *LipPass* under different passphrase lengths in four different environments.

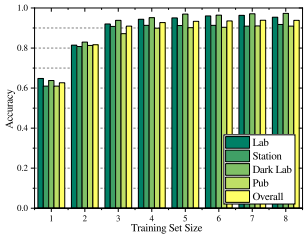

Fig. 22. Authentication accuracy of *LipPass* under different training set sizes in four different environments.

achieve 89.1% overall accuracy. Additionally, although a higher sampling frequency, i.e., $48kHz$, can achieve a better performance, the increases of authentication accuracy is not significant. Specifically, the increase of overall accuracy is 0.5%. These results demonstrate that *LipPass* is robust to the sampling frequency and can work well in old smartphones.

F. Impact of Passphrase Length

Usually, a longer passphrase brings more behavioral characteristics of users' speaking mouths, which can provide stronger security guarantee. However, speaking a too long passphrase will induce a poor user experience. Thus, we evaluate the performance of *LipPass* under different passphrase lengths. Specifically, we sort all passphrases based on their lengths, and obtain the relative authentication results. Fig. 21 shows the authentication accuracy of *LipPass* under different passphrase lengths in four different environments. We can see from the figure that the authentication accuracy of *LipPass* first increases, and then goes stable as the passphrase length increases. Specifically, when the passphrase length increases to 3, the overall authentication accuracy of *LipPass* is above 90%. And the overall authentication accuracy of *LipPass* is stable at around 95% when the passphrase length is larger than 4. Thus, it is appropriate to select 4 as the passphrase length.

G. Impact of Training Set Size

The size of training set is proportional to users' speaking times for registering. In the register phase, more times of users' speaking provides more data for classifiers training. However, too many times of users speaking would lead to a poor user experience in the register phase. Thus, we evaluate the impact of the training set size on *LipPass*. We randomly select 3 volunteers in each environment to conduct the extensive experiment. Each volunteer is required to speak a passphrase with 1-10 times in the register phase, and perform 12 times legitimate authentications in the login phase. Fig. 22 shows the authentication accuracy of *LipPass* under different sizes of training sets in four environments. We can see that as the size of training set increases, the authentication accuracy of *LipPass* first increases and then goes stable. Specifically, to achieve 90% overall accuracy, the speaking times of users is 3 times. When users' speaking times increases to 4 times, the overall authentication accuracy of *LipPass* is 92.7%, and more speaking times would not bring significant increase in authentication accuracy. Thus, we select 3 times for register.

H. Impact of Passphrase With Similar Pronunciation

To validate the impact of passphrase with similar pronunciation, we recruit 10 volunteers to conduct an extensive experiment, in which the volunteers register with five predefined passphrases, and then login with another five passphrases with similar pronunciation. The result shows that the authentication accuracy can achieve over 90%, which indicates that *LipPass* cannot distinguish passphrases with similar pronunciation. This is because the mouth movements of speaking words with similar pronunciation are almost the same. Hence, it is difficult for *LipPass* to distinguish passphrases with similar pronunciation. However, the mouth movements of different users are still distinguishable, so that users' identities still can be validated. For example, if a user registers with 'bed', a spoofer cannot login to the system with either 'bed' or 'bad'. Moreover, we conduct another experiment to evaluate the False Accept Rate (FAR) of *LipPass*. The result shows that the FAR is 1.1%, which is similar with the overall FAR, i.e., 1.3% (as shown in Fig. 15), which indicates that users' identities can be validated under passphrase with similar pronunciation.

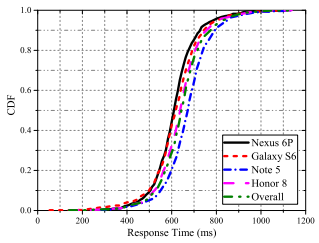

Fig. 23. CDF of the response time under four smartphones.

I. Performance of LipPass in Response Time

In Section III-G, we theoretically analyze the computational complexity of *LipPass*. To further demonstrate *LipPass* can authenticate users responsively, we evaluate the response time of *LipPass* during the login phase. We enable *LipPass* to trace two time points, i.e., the end time t_{talk} of a user's speaking mouth and the time t_{login} when the user logins the system, and obtain *LipPass*' response time $T = t_{login} - t_{talk}$. Usually, the response time of applications is related to the capabilities of smartphones, so we evaluate the response time of *LipPass* under four different smartphones. Fig. 23 shows the response time of *LipPass* under four smartphones. We can see that for 90% of volunteers, the response times are less than $0.73s$, $0.74s$, $0.79s$, and $0.75s$ under Nexus 6P, Galaxy S6, Galaxy Note 5, and Huawei Honor 8, respectively. The average response times are $0.62s$, $0.62s$, $0.67s$, and $0.64s$ under the four smartphones respectively. Users are not clearly aware of such a response time, which demonstrates *LipPass* would authenticate users responsively.

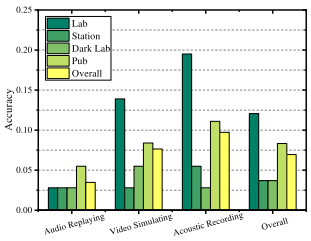

Fig. 24. Authentication accuracy of *LipPass* under different attacks.

J. Attack Resistance of LipPass in Real Environments

To demonstrate *LipPass* can resist various attacks in real authentication scenarios, we conduct an extensive experiment involving 12 volunteers. We select three kinds of attacks to conduct the experiment, i.e., audio replaying (a spoofer records the audio during the user's speaking, and replays it to attack the system), video simulating (a spoofer previously records a video about the mouth movements during the user's speaking, and simulates the mouth movements to attack the system), and acoustic recording (a spoofer emits acoustic signals, receives the reflected acoustic signals from the user's mouth during speaking under an inconspicuous distance, and utilizes it to attack the system). The 12 volunteers are divided into four groups equally, i.e., 3 of them (i.e., 1^{st}-3^{rd} users) register to the system as registered users, and the rest 9 volunteers are divided into 4 groups equally to attack the system through audio replaying (i.e., 4^{th}-6^{th} users), video simulating (i.e., 7^{th}-9^{th} users), and acoustic recording (i.e., 10^{th}-12^{th} users) respectively. For the 3 registered users, they register to the system through speaking a passphrase 3 times. Then, the rest 9 volunteers try to login into the system through the three kinds of attacks, during which each volunteer speaks a passphrase 12 times. We repeat the experiments four times in the four real environments, i.e., laboratory, train station, dark laboratory and pub, respectively.

Fig. 24 shows the authentication accuracy of *LipPass* under three kinds of attacks in four different environments. It can be observed that the overall accuracies under the three kinds of attacks are all less than 10%, which indicates *LipPass* can resist various attacks. Specifically, the authentication accuracies under audio replaying attack in four environments are all less than 6%, which are much less than that under other three attacks. This is because the pre-recorded audio only consists of the voice during users' speaking, instead of mouth movements, and replaying the audio through a speaker cannot reproduce Doppler effects of acoustic signals induced by mouth movements. For the video simulating attack, except for the accuracy in the laboratory approaching 15%, the accuracies in other three environments are all less than 10%. This is because spoofers can only simulate the movements of some outer components in the mouth (e.g., lips and jaw) through a pre-recorded video, instead of movements of inner components (e.g., teeth and tongue). As for acoustic recording attack, the accuracy in the laboratory is approaching 20%, and that in other three environments are all less than 12%. Since this kind of attack directly records the mouth movements during speaking, it is more threating than other three attacks. However, to conduct such an attack in an inconspicuous manner, the distance between spoofer's microphone and user's mouth should be at least $50cm$, which depicts significant accuracy degradation in authentication, and further cannot successfully attack our system.

VI. Related Work

In this section, we review related work about acoustic signal-based applications and existing authentication approaches.

Acoustic Signal-Based Applications: Recently, acoustic sensing attracts considerable attentions since audio devices are widely deployed in mobile devices and the acoustic sensing is non-intrusive. Previous studies propose to use acoustic signals for gesture recognition [24], [27], gesture tracking [12]–[14], and indoor localization [28]. However, there is no work on leveraging acoustic signals to identify a specific user based on the unique behavioral patterns of the user. More recent work [29] utilizes acoustic signals for silent talking recognition, which only recognizes the talking content during speaking. Different from [29], our work can further capture the subtle differences between different users during speaking the same contents.

Password-Based Authentication: As the most typical and widely used approach for user authentication, password-based approach [30] requires users to remember some specific secure texts as the sole tool for authentication. Since the password is not associated with a specific user, weak passwords are vulnerable to stealing attacks [31], while strong passwords are easily forgotten by users [32].

Biometric-Based Authentication: To overcome the vulnerability of password-based authentication, previous works exploit biometric-based authentication approaches, such as fingerprint, face recognition and voiceprint authentications, to identify users. Fingerprint-based authentication, such as Apple Touch ID [5], identifies different users through recognizing the fingers' unique patterns [2]. Face recognition-based authentication, such as Alipay Face Recognition Login [33] and Android Face Unlock [6], utilizes image pattern recognition techniques to capture the uniqueness of users' faces [3]. Voiceprint-based authentication, such as Wechat Voiceprint Lock [34] and Apple 'Hey, Siri' [7], verifies a user through identifying the user's unique speaking voices [4]. However, these existing solutions are vulnerable to replay attacks. For example, attackers can pre-record a video or voice to spoof the face recognition and voiceprint authentication systems. Even the fingerprint-based authentication can be spoofed by the fingerprint film. Recently, Apple launches Face ID [35] leveraging infrared-based technology, which overcomes the disadvantages of typical face recognition-based authentication (e.g., sensitive to environments). However, such an advanced user authentication requires expensive and complex infrastructures, which obstructs the wide deployment in mobile devices.

Authentication With Liveness Verification: To combat the replay attacks, some previous works propose to utilize liveness verification to improve the reliability of user authentication. Luettin *et al.* [10] propose a visual features-based method to distinguish a face of a live user from that in a photo. Zhang *et al.* [8] propose a phoneme localization approach to verify whether a passphrase spoken by a live user or pre-recorded by attackers. However, these works are all sensitive to the ambient environments. For example, face recognition and voiceprint authentications are susceptible to the ambient lights and surrounding audible noises respectively, which leads to significant performance degradations. Recently, there are two simultaneous works [36], [37] which extract users' uniqueness

from mouth movements through Doppler effects of acoustic signals. However, both of these two works do not take the mouth state during speaking into consideration, which depicts potential poor user experience.

Unlike existing approaches, our work leverages acoustic sensing to read users' speaking mouths for user authentication on smartphones, which is robust to different environments and can cope with various attacks.

VII. Conclusion

In this paper, we propose a lip reading-based user authentication system, *LipPass*, by extracting unique behavioral characteristics of users' speaking mouths through acoustic sensing on smartphones. Our system takes step forward to support user authentication in both defending various attacks and adapting to different environments. We find that Doppler profiles of acoustic signals are affected by mouth movements and exhibit unique pattern for different individuals. To characterize the mouth movements, we design a deep learning-based method to extract efficient and reliable features from Doppler profiles. Given extracted features, the mouth state identifier is constructed through softmax function for identifying users' mouth states during speaking. Further, binary classifiers and spoofer detectors are trained based on extracted features for user identification and spoofer detection through Support Vector Machine and Support Vector Domain Description, respectively. Finally, we develop a balanced binary tree-based authentication approach to accurately identify each individual based on the trained classifiers and detectors. Extensive experiments show that *LipPass* is reliable and efficient for user authentication in various environments.

References

[1] Symantec. (2011). *New Norton Anti-Theft to Protect Lost or Stolen Smartphones*. [Online]. Available: https://www.symantec.com/about/newsroom/press-releases/2011/symantec_1004_05

[2] N. Ratha and R. Bolle, *Automatic Fingerprint Recognition Systems*. New York, NY, USA: Springer, 2007.

[3] A. Tefas, C. Kotropoulos, and I. Pitas, "Using support vector machines to enhance the performance of elastic graph matching for frontal face authentication," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 7, pp. 735–746, Jul. 2001.

[4] L. G. Kersta, "Voiceprint identification," *Nature*, vol. 196, no. 4861, pp. 1253–1257, 1962.

[5] Apple. (2018). *Use Touch ID on iPhone and iPad*. [Online]. Available: https://support.apple.com/en-us/HT201371

[6] Google. (2018). *Google Smart Lock*. [Online]. Available: https://get.google.com/smartlock

[7] Apple. (2018). *Siri*. [Online]. Available: http://www.apple.com/ios/siri/

[8] 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*, Vienna, Austria, 2016, pp. 1080–1091.

[9] G. Chetty and M. Wagner, "Multi-level liveness verification for face-voice biometric authentication," in *Proc. Biometrics Symp. Special Session Res. Consortium Conf.*, 2006, pp. 1–6.

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

[11] L. Benedikt, D. Cosker, P. L. Rosin, and D. Marshall, "Assessing the uniqueness and permanence of facial actions for use in biometric applications," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 40, no. 3, pp. 449–460, May 2010.

[12] S. Yun, Y.-C. Chen, H. Zheng, L. Qiu, and W. Mao, "Strata: Fine-grained acoustic-based device-free tracking," in *Proc. ACM MobiSyst.*, Niagara Falls, NY, USA, 2017, pp. 15–28.

[13] R. Nandakumar, V. Iyer, D. Tan, and S. Gollakota, "Fingerio: Using active sonar for fine-grained finger tracking," in *Proc. ACM CHI*, Santa Clara, CA, USA, 2016, pp. 1515–1525.

[14] W. Wang, A. X. Liu, and K. Sun, "Device-free gesture tracking using acoustic signals," in *Proc. ACM Mobicom*, New York, NY, USA, 2016, pp. 82–94.

[15] B. J. Kröger, G. Schröder, and C. Opgen-Rhein, "A gesture-based dynamic model describing articulatory movement data," *J. Acoust. Soc. Amer.*, vol. 98, no. 4, pp. 1878–1889, 1995.

[16] M. Davies. (2018). *Word Frequency: Based on 450 Million Word COCA Corpus*. [Online]. Available: http://www.wordfrequency.info/intro.asp

[17] G. Wang, Y. Zou, Z. Zhou, K. Wu, and L. M. Ni, "We can hear you with Wi-Fi!" *IEEE Trans. Mobile Comput.*, vol. 15, no. 11, pp. 2907–2920, Nov. 2016.

[18] X. Wang and K. K. Paliwal, "Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition," *Pattern Recognit.*, vol. 36, no. 10, pp. 2429–2439, 2003.

[19] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. ACM ICML*, Helsinki, Finland, 2008, pp. 1096–1103.

[20] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, no. 12, pp. 3371–3408, Dec. 2010.

[21] C. M. Bishop, "Pattern recognition," *Mach. Learn.*, 2006.

[22] C. Cortes and V. Vapnik, "Support vector machine," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[23] D. M. J. Tax and R. P. W. Duin, "Support vector domain description," *Pattern Recognit. Lett.*, vol. 20, nos. 11–13, pp. 1191–1199, 1999.

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

[25] I. W. Selesnick and C. S. Burrus, "Generalized digital Butterworth filter design," *IEEE Trans. Signal Process.*, vol. 46, no. 6, pp. 1688–1694, Jun. 1998.

[26] D. T. M. Slock and T. Kailath, "Numerically stable fast transversal filters for recursive least squares adaptive filtering," *IEEE Trans. Signal Process.*, vol. 39, no. 1, pp. 92–114, Jan. 1991.

[27] S. Yun, Y.-C. Chen, and L. Qiu, "Turning a mobile device into a mouse in the air," in *Proc. ACM MobiSyst.*, Florence, Italy, 2015, pp. 15–29.

[28] B. Zhou, M. Elbadry, R. Gao, and F. Ye, "BatMapper: Acoustic sensing based indoor floor plan construction using smartphones," in *Proc. ACM MobiSyst.*, Niagara Falls, NY, USA, 2017, pp. 42–55.

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

[30] J. Yan, A. Blackwell, R. Anderson, and A. Grant, "Password memorability and security: Empirical results," *IEEE Secur. Privacy*, vol. 2, no. 5, pp. 25–31, Sep./Oct. 2004.

[31] E. H. Spafford, "OPUS: Preventing weak password choices," *Comput. Secur.*, vol. 11, no. 3, pp. 273–278, 1992.

[32] P. G. Inglesant and M. A. Sasse, "The true cost of unusable password policies: Password use in the wild," in *Proc. ACM CHI*, Atlanta, GA, USA, 2010, pp. 383–392.

[33] Alibaba. (2018). *Alipay*. [Online]. Available: https://www.alipay.com/

[34] Tencent. (2015). *Voiceprint: The New Wechat Password*. [Online]. Available: http://blog.wechat.com/2015/05/21/voiceprint-the-new-wechat-password/

[35] Apple. (2018). *Face ID—iPhone X—Apple*. [Online]. Available: https://www.apple.com/iphone-x/#face-id

[36] 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*, Dallas, TX, USA, 2017, pp. 57–71.

[37] 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, pp. 36:1–36:18, 2018.

Li Lu received the B.E. degree in computer science and technology from Xi'an Jiaotong University, Xi'an, China, in 2015. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. He is also a Visiting Student with the Department of Electrical and Computer Engineering, Rutgers University. His research interests include mobile and ubiquitous computing, cyber security, and privacy.

Jiadi Yu (M'11) received the Ph.D. degree in computer science from Shanghai Jiao Tong University, Shanghai, China, in 2007. He is currently an Associate Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University. Prior to joining Shanghai Jiao Tong University, he was a Post-Doctoral Fellow with the Data Analysis and Information Security Laboratory, Stevens Institute of Technology, from 2009 to 2011. His research interests include cyber security and privacy, mobile and pervasive computing, cloud computing, and wireless sensor networks. He is a member of the IEEE Communication Society.

Yingying Chen (SM'11) received the Ph.D. degree in computer science from Rutgers University. She is currently a Professor with the Department of Electrical and Computer Engineering, Rutgers University. Prior to joining the Stevens Institute of Technology, she was with Alcatel-Lucent. Her research interests include cyber security and privacy, mobile and pervasive computing, and mobile healthcare. She was a recipient of the NSF CAREER Award, the Google Faculty Research Award, the NJ Inventors Hall of Fame Innovator Award, and the Best Paper Award from the ACM International Conference on Mobile Computing and Networking (MobiCom) 2011. She also received the IEEE Outstanding Contribution Award from the IEEE New Jersey Coast Section each year from 2005 to 2009. Her research has been reported in numerous media outlets, including *MIT Technology Review*, *Wall Street Journal*, and National Public Radio. She is on the editorial boards of the IEEE TRANSACTIONS ON MOBILE COMPUTING, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and the *IEEE Network*.

Hongbo Liu received the Ph.D. degree in electrical engineering from the Stevens Institute of Technology. He has been an Assistant Professor with the Department of Computer Information and Graphics Technology, Indiana University-Purdue University Indianapolis, since 2013. His research interests include mobile and pervasive computing, cyber security and privacy, and smart grid. He was a recipient of the Best Paper Award from ACM MobiCom 2011 and the Best Paper Runner-up Award from IEEE CNS 2013.

Yanmin Zhu (SM'17) received the Ph.D. from the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, in 2007. He was a Research Associate with the Department of Computing, Imperial College London. He is currently a Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University. His research interests include wireless sensor networks and mobile computing. He is a member of the IEEE Communication Society.

Linghe Kong received the B.E. degree in automation from Xidian University, China, in 2005, the Dipl.Ing. degree in telecommunication from TELECOM SudParis (ex. INT), France, in 2007, and the Ph.D. degree in computer science from Shanghai Jiao Tong University, China, in 2012. He was also a joint Ph.D. student at the University of California, San Diego, CA, USA, in 2011, and a Visiting Researcher at Microsoft Research Asia in 2010. He is currently an Associate Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University. Before that, he was a Post-Doctoral Researcher at the Singapore University of Technology and Design in 2013 and at McGill University from 2014 to 2015. His research interests include wireless communication, sensor networks, mobile computing, Internet of Things, and smart energy systems.

Minglu Li is graduated from the School of Electronic Technology, PLA Information Engineering University, in 1985, and the Ph.D. degree in computer software from Shanghai Jiao Tong University in 1996. He is a tenured Full Professor and the Director of the Grid Computing Center, Shanghai Jiao Tong University. His current research interests include cloud computing, big data, and Internet of Things.