LipPass: Lip Reading-based User Authentication on Smartphones Leveraging Acoustic Signals

Li Lu*, Jiadi Yu*§, Yingying Chen†, Hongbo Liu‡ Yanmin Zhu*, Yunfei Liu*, Minglu Li*
*Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, P.R.China
Email: {luli_jtu, jiadiyu, yzhu, liuyunfei, mlli}@sjtu.edu.cn
†Department of Electrical and Computer Engineering, Rutgers University, NJ, USA
Email: yingche@scarletmail.rutgers.edu
‡Department of Computer, Information and Technology, Indiana University-Purdue University Indianapolis, IN, USA
Email: hl45@iupui.edu
§Corresponding Author

***Abstract*—To prevent users' privacy from leakage, more and more mobile devices employ biometric-based authentication approaches, such as fingerprint, face recognition, voiceprint authentications, etc., to enhance the privacy protection. However, these approaches are vulnerable to replay attacks. Although 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. Towards this end, we explore liveness verification of user authentication leveraging users' lip 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 lips leveraging build-in audio devices on smartphones for user authentication. We first investigate Doppler profiles of acoustic signals caused by users' speaking lips, and find that there are unique lip movement patterns for different individuals. To characterize the lip movements, we propose a *deep learning-based method* to extract efficient features from Doppler profiles, and employ *Support Vector Machine* and *Support Vector Domain Description* to construct binary classifiers and spoofer detectors for user identification and spoofer detection, respectively. Afterwards, we develop a *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 90.21% accuracy in user identification and 93.1% accuracy in spoofer detection.**

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, Face recognition, Voiceprint authentications, etc., and relative products are already developed, i.e., Apple Touch ID [2], Alipay Face Recognition Login [3], Wechat Voiceprint Lock [4], etc. However, such authentications are only based on physiological characteristics, suffering from replay attacks [5]. To combat the replay attacks, liveness verification [6] becomes an attractive approach to improve the reliability of user authentication. Luettin et al. [7] propose a visual features-based method to distinguish a face of a live user from a photo. Zhang et al. [5] 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' lip movements, which cannot be forgotten and are robust to noisy environments.

When speaking, people's lips involve in motions. Studies show that such motions present unique lip movement patterns for different individuals [8]. This triggers our research in this work to extract behavioral patterns of lip movements for user authentication on mobile devices, such as smartphones and smartpads. We study whether it is possible to distinguish different user's lip movements leveraging acoustic signals, as acoustic signals have been proved feasible in sensing moving objects [9], [10] 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 the lip reading-based

978-1-5386-4128-6/18/$31.00 ©2018 IEEE

user authentication leveraging acoustic signals, we face several challenges in practice. Firstly, the subtle lip movements need to be captured leveraging acoustic signals. Secondly, the unique behavioral patterns of users' speaking lips 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 computational efficient for smartphones.

In this paper, we first investigate the behavioral patterns of users' speaking lips leveraging acoustic signals. To capture Doppler shift of acoustic signals caused by subtle lip movements, we utilize *signal gradient* in frequency-domain to extract the reflected signals caused by lip movements from a mixed received signal. Through analyzing Doppler profiles of acoustic signals with respect to users' speaking lips, we find that there are unique lip movement patterns for different individuals. Inspired by the observations, we propose a lip reading-based user authentication system, $LipPass$, which reads users' speaking lips leveraging acoustic signals and extracts unique behavioral patterns of users' speaking lips 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 lips under a single word. Given the extracted features, $LipPass$ employs *Support Vector Domain Description* (SVDD) to construct a spoofer detector 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 *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 lip 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 lip movements, and find that there are unique lip movement patterns for different individuals.
- We propose a lip reading-based user authentication system, $LipPass$, which leverages acoustic signals to read users' speaking lips and extract unique behavioral patterns of speaking lips for user authentication.
- We design a deep learning-based method to abstract high-level behavioral characteristics of lip movements, and employ SVM and SVDD to train binary classifiers and spoofer detectors for user identification and spoofer detection, respectively.

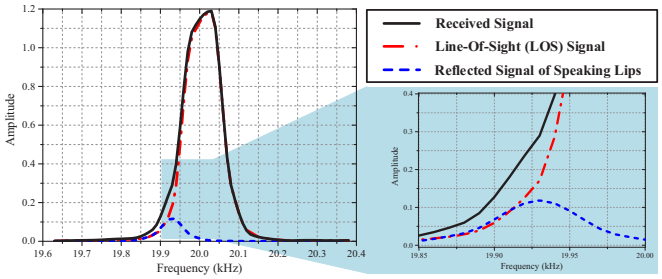


Fig. 1. An example of a mixed received signal including a LOS signal and a reflected signal from speaking lips.

- We develop a binary tree-based authentication approach for 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 90.21% accuracy on average in user identification and 93.1% accuracy in spoofer detection across different environments.

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 implementation details are described in Section IV. The evaluation of the system is presented in Section V. Finally, we review several related work in Section VI and make a conclusion in Section VII.

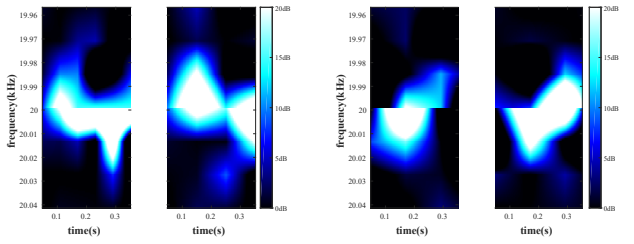
II. Preliminary

Audio devices on smartphones can be exploited to build an acoustic signal field by continually emitting acoustic signals with the speaker and receiving the signals by microphones on a smartphone. A user's lip movements can induce Doppler effect of acoustic signals while the user speaks words. Different users exhibit subtle differences on 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 lips 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:

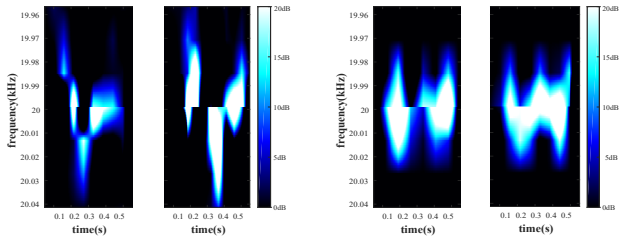
$$\Delta f = \frac{v}{c} \times f_0, \tag{1}$$

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



(a) User 1 (b) User 2

Fig. 2. Doppler profiles of acoustic signals caused by speaking the word 'Hello' under two different users.



(a) User 1 (b) User 2

Fig. 3. Doppler profiles of acoustic signals caused by speaking the word 'World' under two different users.

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 lips is much slower, the corresponding Doppler shift will lie in the frequency band of the LOS signals. Fig. 1 shows an example of a mixed received signal including a LOS signal and a reflected signal from speaking lips. We can see that, in the received signal, the reflected signal caused by speaking lips is buried within the LOS signal.

In order to capture Doppler shift of acoustic signals caused by subtle lip 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 lips 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 lips, the reflected signals from surrounding static objects (e.g., furnitures), and the environmental noises, i.e.,

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

where $s^{e}_{(f)}(t)$ is the LOS signal in time slot t, $s^{rl}_{(f)}(t)$ is the reflected signal from speaking lips 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

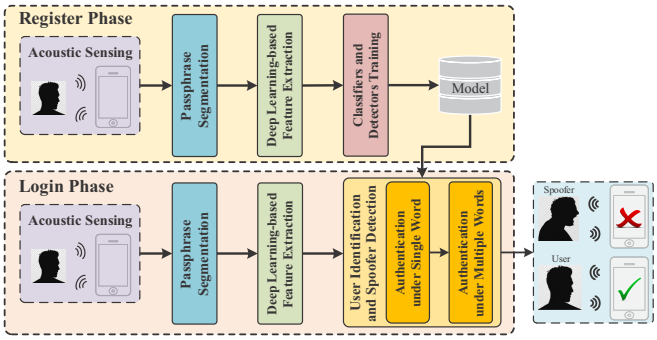


Fig. 4. System architecture of $LipPass$.

the time. Thus, the signal gradient of received signals in frequency-domain from time slot $t-1$ to t, $g(t)$, is:

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

The gradient feature matrix $G = [g(1), g(2), \cdots, g(T)]$ can represent Doppler profiles of users' speaking lips within a duration time T.

Fig. 2 and 3 show two Doppler profiles of acoustic signals caused by speaking two words (i.e., 'Hello' and 'World') from two different users respectively. Compare Fig. 2(a) with 2(b), we observe that Doppler profiles of speaking the word 'Hello' exhibit different variation trends between the two users. Fig. 3(a) and 3(b) show the similar results. Additionally, speaking the same word by the same user produces similar Doppler profiles. These encouraging results demonstrate the great potential that Doppler effect of acoustic signals caused by users' speaking lips can be used in user authentication.

III. System Design

In this section, we present the design of the lip reading-based user authentication system, $LipPass$, which leverages acoustic signals to read users' speaking lips and capture the unique behavioral patterns of lip movements for user authentication.

A. Overview

Fig. 4 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 lips. First, $LipPass$ segments the received signals of the passphrase into several episodes, each representing a single word. Then, $LipPass$ extracts efficient and reliable features from the signal episodes leveraging a deep learning-based method. Finally, based on these features, $LipPass$ employs Support Vector Machine and Support Vector Domain Description to construct binary classifiers and spoofer detectors for user identification and spoofer detection respectively.

In the login phase, $LipPass$ first captures reflected signals when user speaks the same passphrase as that in the register

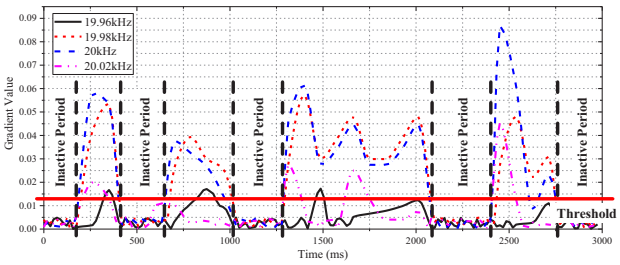


Fig. 5. Doppler profiles of lip movements when a user speaks four words under four frequencies.

phase, then performs passphrase segmentation and feature extraction. In user authentication, $LipPass$ applies a 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, $LipPass$ further employs a weighted voting scheme for user authentication by examining lip movement patterns with multiple words.

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 lips. $LipPass$ first segments the received signals of the given passphrase into episodes, each representing a single word. According to [11], there is usually a short interval (e.g., 300 ms) between speaking two successive words. Fig. 5 shows Doppler profiles of lip 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 to detect all inactive periods in a passphrase and segments the passphrase. The threshold can be set as the mean value of the noises in the surrounding. $LipPass$ would extract features from the signal episode of each single word for classifier training and user authentication.

C. Deep Learning-based Feature Extraction

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 [12], Doppler profiles of users' speaking lips are usually non-linear separated. Therefore, we develop a deep learning-based method, a *three-layer autoencoder-based Deep Neural Network* (DNN) [13], to extract efficient and reliable features from Doppler profiles of users' speaking lips.

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 through an unsupervised manner. Such compressed representations are able

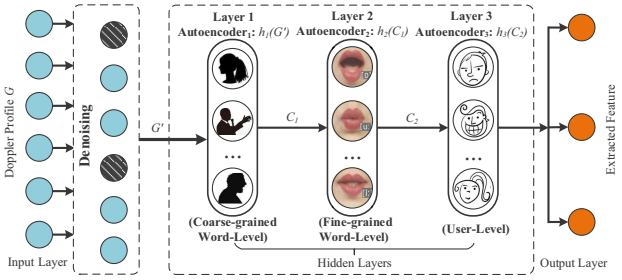


Fig. 6. Architecture of feature extraction through a three-layer autoencoder-based Deep Neural Network.

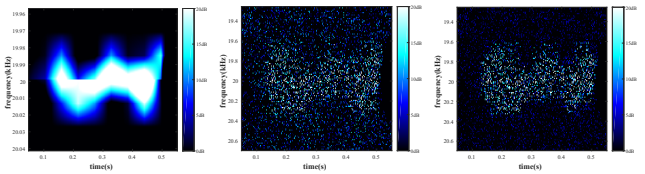
to characterize unique behavioral patterns of users' speaking lips. The autoencoder can map the input X into a set of compressed representation C as $C = \sigma(wX + b)$, where $\sigma()$ is a logistic function defined as $\sigma(x) = \frac{1}{1+e^{-x}}$, w and b are the weight and bias of the autoencoder 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 regulariser for the parameters and sparsity, and λ as well as β are the coefficients of the two L_2 regularisers. 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. 6 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 lips 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 [13] to denoise Doppler profiles G of users' speaking lips 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, the coarse-grained word-level features C_1 can be extracted as output by the autoencoder $h_1(G')$ in the first layer. 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 word-level features C_2 (e.g., phoneme-level features). 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. 7 shows two reconstructed profiles of a user speaking the word 'World' based on extracted features of lip movements. Compare with the original Doppler profile as shown in Fig. 7(a), we can observe that both reconstructed profiles in



(a) Original. (b) Without denoising. (c) With denoising.

Fig. 7. Reconstructed profiles based on extracted features of lip movements.

Fig. 7(b) and 7(c) can recover basic features from the original Doppler profile, and the reconstructed result with denoising shows more significant features than that without denoising.

D. Classifiers and Detectors Training for User Authentication

Given extracted features from Doppler profiles of users' speaking lips through DNN model, we employ *Support Vector Machine* (SVM) [14] to train classifiers and detectors for user identification and spoofer detection.

For a single-user system, when a user registers to $LipPass$, the user is required to speak a predefined passphrase several times, so $LipPass$ can extract the user's unique features from Doppler profiles of the user's speaking lips 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) [15], 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 newly 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 applies the one-versus-rest method to divide the n users' training data into two-classes data, i.e., the n^{th} user's data and prior $(n-1)$ registered users' data, and then employs SVM to train a binary classifier for the n^{th} user based on the two-classes data, which can distinguish the n^{th} user from prior $(n-1)$ registered users. In a multi-users system, $LipPass$ would train a binary classifier for each 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.

E. User Identification and Spoofer Detection

In the login phase, $LipPass$ usually requires users to speak a passphrase including several words. $LipPass$ first identifies each individual and detects spoofers under each word. Then, based on authentication results under single words, $LipPass$ achieves the final authentication result under multiple words.

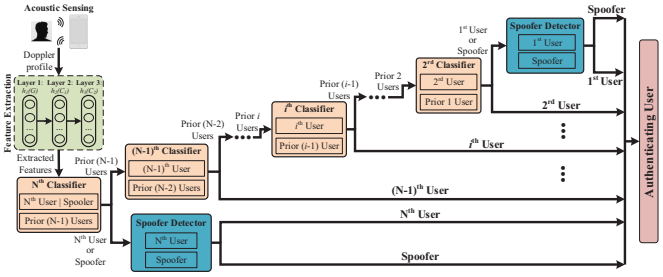


Fig. 8. Architecture of the binary tree-based authentication under single word.

1) Authentication under Single Word: In the register phase, for a user U_i in a n users system, $LipPass$ trains a binary classifier based on U_i's features and prior $(i-1)$ registered users' features to verify 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., $U_{i+1}, U_{i+2}, ..., U_n$) and spoofers, the user could be U_i, one of the subsequent registered users (i.e., $U_{i+1}, U_{i+2}, ..., U_n$) or a spoofer if the i^{th} classifier verifies a login user as U_i. Thus, in the login phase, we propose a *binary tree-based authentication approach* to verify users' identities and detect spoofers. Fig. 8 shows the architecture of the 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 lips, and then segments received acoustic signals into episodes, as well as extracts features of the user's speaking lips from the episodes through DNN model. Based on the n^{th} classifier, $LipPass$ verifies whether the user is the n^{th} user or one of prior $(n-1)$ registered users. If the classifier identifies the user as the n^{th} user, $LipPass$ would feed the user's extracted features to the spoofer detector based on the n^{th} user's features, which will verify whether the user is the n^{th} user or a spoofer. On the contrary, if the n^{th} classifier identifies the user as one of prior $(n-1)$ registered users, the extracted features are further fed to the $(n-1)^{th}$ classifier. By analogy, if the i^{th} classifier identifies the user as the i^{th} user, $LipPass$ can verify that the user is not an arbitrary user of the prior $(i-1)$ users. Additionally, $LipPass$ has verified that the user is not an arbitrary one of the subsequent registered users (i.e., $U_{i+1}, U_{i+2}, \cdots, U_n$) through the $(i+1)^{th}$ to n^{th} classifiers, so $LipPass$ can regard the user as the i^{th} user. For the 1^{st} user, $LipPass$ utilizes the spoofer detector based on the 1^{st} user's features to distinguish the 1^{st} user from spoofers. Finally, $LipPass$ is able to accurately identify a login user as a registered user or spoofer.

The time complexity of the binary tree-based authentication approach is $O(N)$, where N is the number of registered users. Thus, our authentication approach is lightweight and computational efficient for smartphones.

2) Authentication under Multiple Words: To strengthen the robustness of the authentication result, $LipPass$ verifies

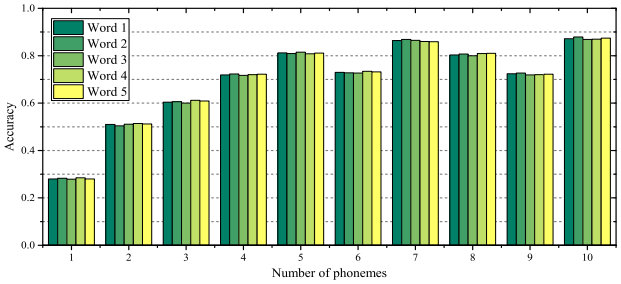


Fig. 9. Relationship between authentication accuracy and the number of phonemes in a single word.

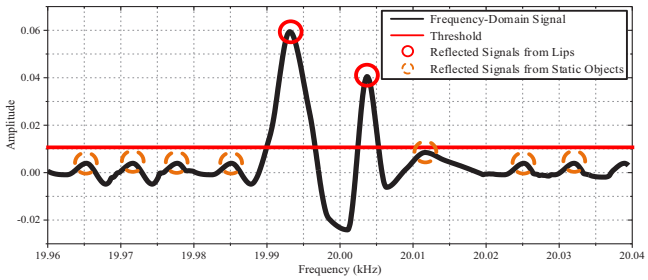


Fig. 10. An example of the reflected signals from speaking lips and static objects.

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 lips. 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. 9 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., $\{w_1, w_2, ..., w_m\}$, we define the confidence of a user U_i as follows:

$$conf_i = \sum_j w_j, j \in \{k | L_k = U_i\}. \quad (5)$$

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

IV. Implementation

$LipPass$ utilizes acoustic signals to read users' speaking lips for user authentication. 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.

A. Eliminating Multi-path Interferences from Body Movements

In practice, users' speaking lips 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 lips. The normal body movements lead to a Doppler shift ranging in $[50, 200]Hz$ [17], and Doppler shift of audible voices ranges in $[500, 2000]Hz$. However, Doppler shift caused by users' speaking lips is $[-40, 40]Hz$. Thus, we apply a *Butterworth Band-Pass Filter* [18] for acoustic signals to obtain the target frequency band, i.e., $[f_0 - 40, f_0 + 40]Hz$, for speaking lips detection, and eliminate other out-band interferences.

B. Removing Multi-path Interferences from Static Objects

When users authenticate through $LipPass$, except for reflected signals from users' lips, there are other reflected signals from static objects, such as desks and chairs. 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 lips. Thus, it is necessary to remove the reflected signals from static objects in received signals.

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

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

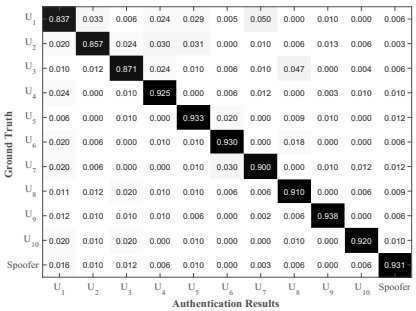


Fig. 11. Confusion matrix of $LipPass$, each entry of which is the average value in four different environments.

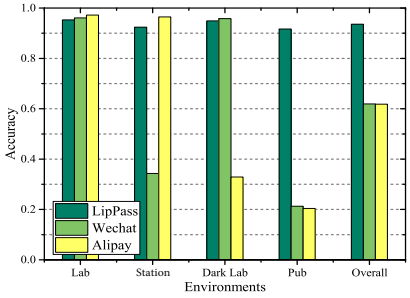


Fig. 12. Authentication accuracy of $LipPass$, voiceprint and face authentications in four different environments.

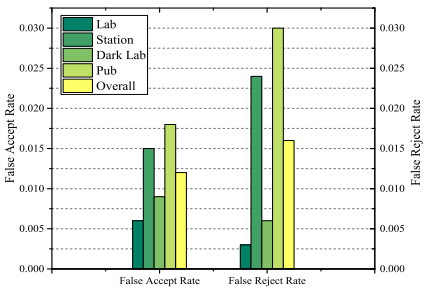


Fig. 13. False accept rate and false reject rate of $LipPass$ in four different environments.

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, to conduct our experiments. Among the 12 volunteers, 10 of them register in the system with $LipPass$ while the rest two volunteers as spoofers. Each volunteer randomly selects a smartphone for the experiment. We predefine 10 passphrases, each of which contains 1-10 words. In each passphrase, we select words with the number of phonemes larger than 4. This is because when the number of phonemes increases to 4, the expected authentication accuracy under single word can be achieved, as Fig. 9 shows. 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. 11 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 90.21% with a standard derivation of 3.52%, and the average accuracy in spoofer detection is 93.1%.

We compare the performance of $LipPass$ with that of Wechat voiceprint lock and Alipay face recognition login. Fig. 12 shows the authentication accuracies of $LipPass$, Wechat voiceprint lock and Alipay face recognition login in four different environments respectively. It can be seen from the figure 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. 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.

We further evaluate the reliability and user experience of $LipPass$ through the false accept and false reject rates. Fig. 13 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. 13 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. 14 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 system through speaking a passphrase less than 4 times, which is acceptable for users in real environments.

C. Performance of LipPass in Response Time

We enable $LipPass$ to trace two time points, i.e., the end time t_{talk} of a user's speaking lips 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. 15 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 $0.62s$, $0.62s$,

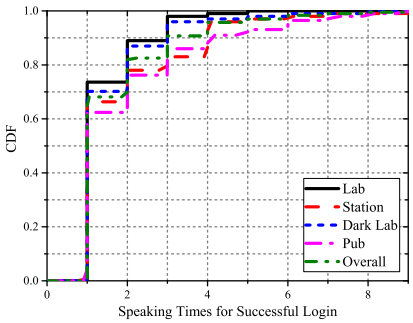


Fig. 14. CDF of the speaking times for successful login in four different environments.

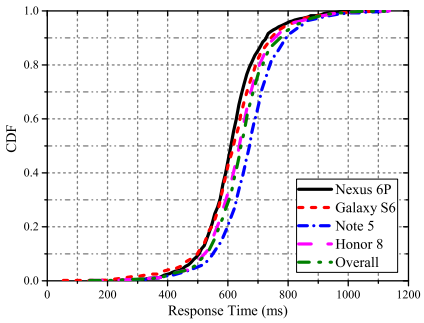


Fig. 15. CDF of the response time under four smartphones.

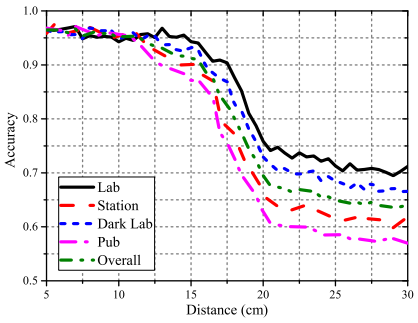


Fig. 16. Relationship between authentication accuracy and distances from microphone to users' lips in four different environments.

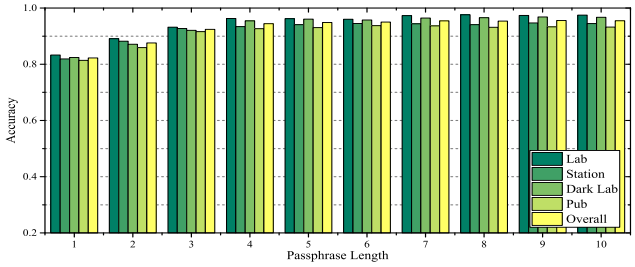


Fig. 17. Authentication accuracy of $LipPass$ under different passphrase lengths in four different environments.

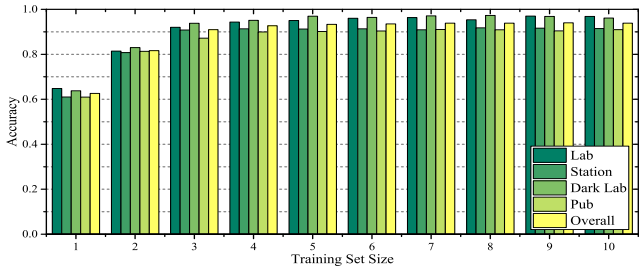


Fig. 18. Authentication accuracy of $LipPass$ under different training set sizes in four different environments.

$0.67s$, and $0.64s$ under the four smartphones respectively. Users are not clearly aware of such a response time, which demonstrates $LipPass$ would authenticate users efficiently.

D. Impact of Distance between Microphone and Users' Lips

Since we utilize acoustic signals to capture users' speaking lips, the signal attenuation cannot be avoided. A longer distance between the microphone and users' lips 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' lips and the microphone through Time of Arrival (ToA). Fig. 16 shows the relationship between the authentication accuracy of $LipPass$ and distance from the microphone to users' lips 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 lips becomes larger as the distance between the microphone and users' lips increases. However, the authentication accuracies in all four environments can achieve 95% authentication accuracy as the distance less than $12cm$.

E. Impact of Passphrase Length

Usually, a longer passphrase brings more behavioral patterns of users' speaking lips, which can provide stronger security guarantee. However, speaking a too long passphrase will induce a poor user experience. Specifically, we sort all passphrases based on their lengths, and obtain the relative authentication results. Fig. 17 shows the authentication accuracy of $LipPass$ under different passphrase lengths in four different environments. We can see from the figure that the authentication accuracy 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 larger than 4. Thus, it is appropriate to select 4 as the passphrase length for $LipPass$.

F. 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 much times of users speaking would lead to a poor user experience in the register phase. 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. 18 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 accuracy of $LipPass$ is 92.69%, and more speaking times would not bring significant increase in authentication accuracy. Thus, we select speaking 3 times for user registering.

VI. Related Work

Acoustic Signals-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 [17] [19], gesture tracking [9], [10], and even silent talking recognition [20]. However, there is no work on leveraging acoustic signals to identify a specific user based on the unique behavioral patterns of the user.

Password-based Authentication. As the most typical and widely used approach for user authentication, password-based approach [21] requires users to remember some specific secure texts as the sole tool for authentication. Since the password is not associated with a specific user, any spoofer who steals the password can pass the authentication.

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 [2], identifies different users through recognizing the fingers' unique patterns. Face recognition-based authentication, such as Alipay Face Recognition Login [3], utilizes image pattern recognition techniques to capture the uniqueness of users' faces. Voiceprint-based authentication, such as Wechat Voiceprint Lock [4], verifies a user through identifying the user's unique speaking voices. 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.

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. [7] propose a visual features-based method to distinguish a face of a live user from that in a photo. Zhang et al. [5] 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, such as ambient lights and audible noises. Unlike existing approaches, our work leverages acoustic signals to read users' speaking lips 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 lips leveraging build-in audio devices on smartphones. Our system takes step forward to support user authentication in not only defending various attacks but also adapting to different environments. We find that Doppler profiles of acoustic signals are affected by lip movements and exhibit unique pattern for different individual. To characterize the lip movements, we design a deep learning-based method to extract efficient and reliable features from Doppler profiles of users' speaking lips. Given the extracted features, binary classifiers and spoofer detectors are trained for user identification and spoofer detection through Support Vector Machine and Support Vector Domain Description, respectively. Finally, we develop a 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.

Acknowledgment

This research is sponsored by National Natural Science Foundation of China (No. 61772338, 61772341, 61472254). This work is also supported by the Program for Changjiang Young Scholars in University of China, the Program for Shanghai Top Young Talents, and the National Top Young Talents Support Program.

References

[1] Symantec, "New norton anti-theft to protect lost or stolen smartphones," 2011. [Online]. Available: https://www.symantec.com/about/newsroom/press-releases/2011/symantec_1004_05

[2] Apple, "Use Touch ID on iPhone and iPad," 2017. [Online]. Available: https://support.apple.com/en-us/HT201371

[3] Alibaba, "Alipay," 2017. [Online]. Available: https://www.alipay.com/

[4] Tencent, "Voiceprint: The new wechat password," 2015. [Online]. Available: http://blog.wechat.com/2015/05/21/voiceprint-the-new-wechat-password/

[5] 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'16*, Vienna, Austria, 2016.

[6] G. Chetty and M. Wagner, "Multi-Level Liveness Verification for Face-Voice Biometric Authentication," in *Biometrics Symposium'06: Special Session on Research at the Biometric Consortium Conference*, 2006.

[7] J. Luettin, N. A. Thacker, and S. W. Beet, "Speaker identification by lipreading," in *Proc. IEEE ICSLP'96*, Philadelphia, PA, USA, 1996.

[8] L. Benedikt, D. Cosker, P. L. Rosin, and D. Marshall, "Assessing the uniqueness and permanence of facial actions for use in biometric applications," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 40, no. 3, pp. 449–460, 2010.

[9] S. Yun, Y.-C. Chen, H. Zheng, L. Qiu, and W. Mao, "Strata: Fine-Grained Acoustic-based Device-Free Tracking," in *Proc. ACM Mobisys'17*, Niagara Falls, NY, USA, 2017.

[10] W. Wang, A. X. Liu, and K. Sun, "Device-free gesture tracking using acoustic signals," in *Proc. ACM Mobicom'16*, New York, USA, 2016.

[11] G. Wang, Y. Zou, Z. Zhou, K. Wu, and L. M. Ni, "We Can Hear You with Wi-Fi!" *IEEE Transactions on Mobile Computing*, vol. 15, no. 11, pp. 2907–2920, Nov. 2016.

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

[13] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and Composing Robust Features with Denoising Autoencoders," in *Proc. ACM ICML '08*, Helsinki, Finland, 2008.

[14] C. Cortes and V. Vapnik, "Support vector machine," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[15] D. M. J. Tax and R. P. W. Duin, "Support vector domain description," *Pattern Recognition Letters*, vol. 20, no. 1113, pp. 1191–1199, 1999.

[16] M. Davies, "Word frequency: based on 450 million word COCA corpus," 2017. [Online]. Available: http://www.wordfrequency.info/intro.asp

[17] S. Gupta, D. Morris, S. Patel, and D. Tan, "Soundwave: Using the doppler effect to sense gestures," in *Proc. ACM CHI '12*, Austin, Texas, USA, 2012.

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

[19] S. Yun, Y.-C. Chen, and L. Qiu, "Turning a mobile device into a mouse in the air," in *Proc. ACM MobiSys '15*, Florence, Italy, 2015.

[20] J. Tan, C.-T. Nguyen, and X. Wang, "SilentTalk: Lip Reading through Ultrasonic Sensing on Mobile Phones," in *Proc. IEEE INFOCOM'17*, Atlanta, USA, 2017.

[21] J. Yan, A. Blackwell, R. Anderson, and A. Grant, "Password memorability and security: empirical results," *IEEE Security Privacy*, vol. 2, no. 5, pp. 25–31, Sep. 2004.