VoiceLive: A Phoneme Localization based Liveness Detection for Voice Authentication on Smartphones

Linghan Zhang†, Sheng Tan†, Jie Yang†, Yingying Chen*
†Florida State University, Tallahassee, FL 32306, USA
*Stevens Institute of Technology, Hoboken, NJ 07030, USA
†{lzhang, tan, jie.yang}@cs.fsu.edu, *yingying.chen@stevens.edu

ABSTRACT

Voice authentication is drawing increasing attention and becomes an attractive alternative to passwords for mobile authentication. Recent advances in mobile technology further accelerate the adoption of voice biometrics in an array of diverse mobile applications. However, recent studies show that voice authentication is vulnerable to replay attacks, where an adversary can spoof a voice authentication system using a pre-recorded voice sample collected from the victim. In this paper, we propose VoiceLive, a practical liveness detection system for voice authentication on smartphones. VoiceLive detects a live user by leveraging the user's unique vocal system and the stereo recording of smartphones. In particular, with the phone closely placed to a user's mouth, it captures time-difference-of-arrival (TDoA) changes in a sequence of phoneme sounds to the two microphones of the phone, and uses such unique TDoA dynamic which doesn't exist under replay attacks for liveness detection. VoiceLive is practical as it doesn't require additional hardware but two-channel stereo recording that is supported by virtually all smartphones. Our experimental evaluation with 12 participants and different types of phones shows that VoiceLive achieves over 99% detection accuracy at around 1% Equal Error Rate (EER). Results also show that VoiceLive is robust to different phone placements and is compatible to different sampling rates and phone models.

Keywords

Voice recognition; Liveness detection; Phoneme localization

1. INTRODUCTION

As a primary way of communication, our voice is a particularly attractive biometric for identifying users. It reflects individual differences in both behavioral and physiological characteristics, such as the inflection and the shape of the vocal tract [23]. Such distinctive behavioral and physiological traits could be captured by voice authentication systems for differentiating each individual [17]. Voice authentication leveraging built-in microphones on mobile devices is particularly convenient and low-cost, comparing to the passwords authentication that is difficult to use while on-the-go and requires memorization. Recent advances in mobile technology further accelerate the adoption of voice biometrics in an array of diverse mobile applications.

Indeed, voice authentication has been introduced recently to mobile devices and apps to provide secure access and logins. For example, Google has integrated it into Android operating systems (OSs) to allow users to unlock mobile devices [2], and Tencent has updated its WeChat mobile app to support voice biometric logins [8]. Another appealing use case of voice authentication is to support mobile financial services. For instance, SayPay provides voice biometric solution for online payment, e-commerce, and online banking [5]. And an increasing number of financial institutions, HSBC, Citi, and Barclays for example, are deploying voice authentication for their telephone and online banking systems [3]. This trend is expected to continue growing at a rate of 22.15 percent yearly until 2019, and will result in an estimated $113.2 billion market share by 2017 [4]. Voice authentication thus becomes an attractive alternative to passwords in mobile authentication and is increasingly popular.

Voice authentication however has been shown to be vulnerable to replay attacks in recent studies [16, 33, 14]. An adversary can spoof a voice authentication system by using a pre-recorded voice sample collected from the victim. The voice sample can be any recording captured inconspicuously. Or, an adversary can obtain voice samples from the victim's publicly exposed speeches. The attacker could even concatenate voice samples from a number of segments in order to match the victim's passphrase. Such attacks are most accessible to the adversary due to the proliferation of mobile devices, such as smartphones and digital recorders. They are also highly effective in spoofing authentication systems, as evidenced by recently work [32, 33]. Replay attacks therefore present significant threats to voice authentication and are drawing increasing attention. For example, Google advises users on the vulnerability of their voice logins by displaying a popup message "... a recording of your voice could unlock your device." [1]

Prior work in defending against replay attacks is to utilize liveness detection to distinguish between a passphrase spoken by a live user and a replayed one pre-recorded by the adversary. For example, Shang *et al.* propose to compare an input voice sample with stored instances of past accesses to detect the voice samples have been seen before by the authentication system [31]. This method, however,

CCS'16, October 24-28, 2016, Vienna, Austria

DOI: http://dx.doi.org/10.1145/2976749.2978296

cannot work if the attacker records the voice samples during a non-authentication time point. Villalba *et al.* and Wang *et al.* suggest that the additional channel noises introduced by the recording and loudspeaker can be used for attack detection [32, 33]. These approaches however have limited effectiveness in practice. For example, the false acceptance rates of these approaches are as high as 17%. Chetty and Wagner propose to use video camera to extract lip movements for liveness verification [13], whereas Poss *et al.* combine the techniques of a neural tree network and Hidden Markov Models to improve authentication accuracy [28]. Aley-Raz *et al.* develop a liveness detection system based on "Intra - session voice variation" [10], which is integrated into Nuance VocalPassword [6]. In addition to a user-chosen passphrase, it requires a user to repeat one or more random sentences prompted by the system for liveness detection.

In this paper, we introduce and evaluate a phoneme sound localization based liveness detection system on smartphones. Our system distinguishes a passphrase spoken by a live user from a replayed one by leveraging (i) the human speech production system and (ii) advanced smartphone audio hardware. First, in human speech production, a phoneme is the smallest distinctive unit sound of a language. Each phoneme sound can be viewed as air waves produced by the lungs, and then modulated by the movements of vocal cords and vocal tract including throat, mouth, nose, tongue, teeth, and lips. Each phoneme sound thus experiences unique combination of place and manner of articulation. Consequently, different phoneme sounds could be located at different physical positions in the human vocal tract system with an acoustic localization method. Second, smartphone hardware is now supporting advanced audio capabilities. Virtually all smartphones are equipped with two microphones for stereo recording (one on the top and the other one at the bottom), and are capable of recording at standard 48kHz and 192kHz sampling rates. For example, with the latest Android OSs, Samsung Galaxy S5 and Note3 are capable of stereo recording at 192kHz, which yields 5.21 microseconds' time resolution or millimeter-level ranging resolution[1]. We thus can leverage such stereo recording or dual microphones on smartphones to pinpoint the sound origin of each phoneme within human vocal system for liveness detection.

Ideally, locating a phoneme sound requires at least three microphones with three individual audio channels. Although current two-channel stereo recording cannot uniquely locate the phoneme sound origin, it can capture the time-difference-of-arrival (TDoA) of each phoneme sound to the two microphones of the phone. With the phone closely placed to user's mouth, the differences in TDoA between most phoneme sounds are distinctive and measurable with millimeter-level ranging resolution. Very importantly, each passphrase (usually 5 to 7 words [7, 30]) consists of a sequence of different phoneme sounds that will produce a series of TDoA measurements with various values. We refer to the changes in TDoA values as TDoA dynamic, which is determined by the specific passphrase, the placement of the phone, and a user's unique vocal system. Such TDoA dynamic, which doesn't exist under replay attacks, is then utilized for liveness detection.

In particular, when a user first enrolled in the system, the TDoA dynamic of the user-chosen or system prompted passphrase is first captured by the smartphone stereo recording, and then stored in the system. During online authentication phase, the extracted TDoA dynamic of an input utterance will be compared to the one stored in the system. A live user is detected, if that produce a similarity score higher than a pre-defined threshold. By relaxing the problem from locating each phoneme sound to measuring the TDoA dynamic for a sequence of phonemes, we enable liveness detection on a single phone without any additional hardware. Our system does have the limitation of requiring a user to hold the phone close to her/his mouth with the same pose in both enrollment and authentication processes. The contributions of our work are summarized as follows:

- We show that the origin of each phoneme can be uniquely located within the human vocal tract system by using a microphone array. It lays the foundation of our phoneme localization based liveness detection system.
- We develop VoiceLive, a practical liveness detection system that extracts the TDoA dynamic of the passphrase for live user detection. VoiceLive takes advantages of the user's unique vocal system and high quality stereo recording of smartphones.
- We conduct extensive experiments with 12 participants and three different types of phones under various experimental settings. Experimental results show that VoiceLive achieves over 99% detection accuracy at around 1% EER. Results also show that VoiceLive is robust to different phone placements and is compatible to different sampling rates and phone models.

The remainder of this paper expands on above contributions. We begin with system and attack model, and a brief introduction to phoneme sounds localization.

2. PRELIMINARIES

2.1 System and Attack Model

There exists two types of voice authentication systems: text-dependent and text-independent. We primarily focus on the text-dependent system as it is currently the most commercially viable method and produces better authentication accuracy with shorter utterances [31]. In a text-dependent system, the text to be spoken by a user is the same one for enrollment and verification. Such text could be either a user-chosen or system prompted one. Figure 1 shows the processes of a typical voice authentication system. Our method can also be extended to text-independent systems, which will be discussed in Section 5.

For the attack model, we consider replay attacks, which are the most accessible and effective attacks aiming at spoofing the system by replaying a pre-recorded voice sample of the victim [32]. We consider the replay attacks that take place at two locations, at the microphone point and at the transmission point, as shown in Figure 1. For the sake of simplicity, we refer to the former as a *playback attack* and the latter as a *replace attack*. In a playback attack, an adversary uses a speaker to replay the pre-recorded voice sample in front of the microphones. In a replace attack, an adversary replaces his/her own speech signal as the victim's before or during transmission. This can be done by leveraging the availability of the virtual recorder to bypass the local microphones, or by intercepting and replacing speech signal during transmission.

[1]Assuming the speed of sound is 340m/s, each digital sample represents a distance of 1.77mm.

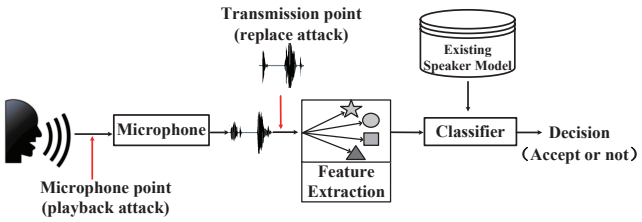
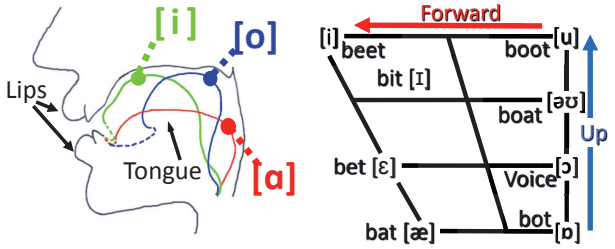


Figure 1: A typical voice authentication system with two possible places of replay attacks.



(a) Example (b) Vowel chart

Figure 2: Tongue positions of English vowels within the oral cavity, and the vowel chart.

2.2 Human Speech Production and Phonemes

The human speech production system involves three vital physiological components: lungs, vocal cords, and vocal tract [27]. When someone exhales, air is expelled from the *lungs*, and then passes over the *vocal cords*, which dilate or constrict to allow or impede the air flow to produce unvoiced or voiced sound. Such sound is then resonated and reshaped by the *vocal tract* that consists of multiple organs such as throat, mouth, nose, tongue, teeth, and lips. The vocal cords modulation, interaction and movement of these organs can alter sound waves and produce unique human sounds.

A phoneme is the smallest distinctive unit sound of a language [27]. The two major phoneme categories are vowels and consonants. In particular, vowels are the phoneme sounds produced when vocal cords constrict air flow (i.e., voiced sound) but with an open vocal tract. The tongue position is the most important physical feature that distinguishes one vowel from another [27]. As different tongue positions lead to different multipath environments inside the oral cavity, we can locate the sound origins of different vowels at different physical locations inside the human oral cavity. As illustrated in Figure 2 (a), when the tongue moves to lower right corner, vowel [ɑ] can be pronounced, whereas when the tongue moves to upper left corner and backward, vowels [i] and [o] can be produced, respectively. More generally, Figure 2 (b) shows the vowel chart which involves two dimensions of tongue movements: up/down movements (i.e., height) and back/forth movements (i.e., backness). Extending or retracting the tongue forward or backward towards the teeth produces a more front or back vowel sound, whereas lowering or raising the tongue towards lower jaw or towards the roof of mouth produces a more open or close vowel.

Unlike vowels, consonants are produced when vocal cords either constrict or dilate air flow and with significant constriction of the air flow in the oral cavity. The articulation place and manner are two major factors that distinguish one consonant from another [27]. The combined effect of place and manner of articulation and voiced/unvoiced sound lead to different consonant sounds emitted from different locations within the human vocal tract system. In particular, place of articulation is the location where the constrictions or obstructions of air stream occur, and can be categorized into 6 groups: bilabial, labiodental, dental, alveolar, palatal, and velar. Figure 3 shows each group and the corresponding consonants. For example, the consonants [p][b][m][w] can be pronounced when the obstruction of air stream occurs at upper and lower lips. The consonants within each group can be further distinguished by the manner of articulation, which describes the configuration and interaction of the speech organs (e.g., the tongue, lips, and palate). There are 6 types of articulation manners including nasal, stop, fricative, affricate, approximate and lateral. For instance, nasal consonant [m] is produced when the air stream is completely blocked by mouth and only passes through the nose. Figure 4 summarizes the categorization of different consonants based on place of articulation and manner of articulation. The bolded font in the figure shows the voiced sounds (e.g., **[b]** and **[v]**), whereas the rest are unvoiced sounds (e.g., *[p]* and *[f]*).

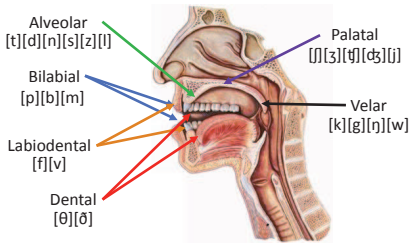


Figure 3: Place of articulation and corresponding consonants.

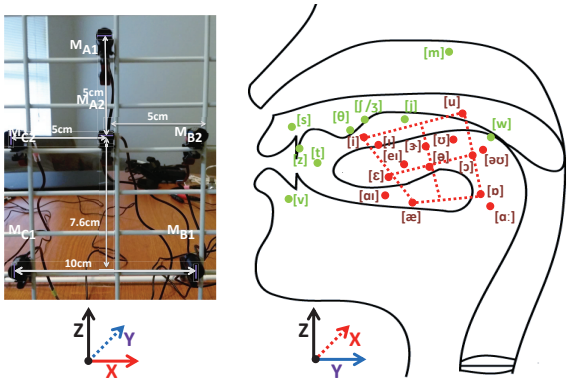
Manner \ Place	Bilabial	Labiodental	Dental	Alveolar	Palatal	Velar
Nasal	[m]			[n]		[ŋ]
Stop	*[p]* **[b]**			*[t]* **[d]**	*[ʃ]* **[ʒ]**	*[k]* **[g]**
Fricative		*[f]* **[v]**	[θ] **[ð]**	*[s]* **[z]**		
Affricate					*[ʧ]* **[ʤ]**	
Approximate						[w]
Lateral				[l]		

Figure 4: Consonants chart based on place and manner of articulation.

2.3 Phoneme Localization using Microphone Array

We next conduct experiments to study how the origin of phoneme sound is located within the human vocal tract system by leveraging a microphone array. We utilize six external microphones organized in three pairs A, B, and C. As shown in Figure 5 (a), the microphones are distributed in the X-Z plane[2] with 5cm and 10cm horizontal distances, and 5cm and 7.6cm vertical distances. Such a distribution could cover the size of a human vocal tract. Each pair is synchronized to measure the TDoA of the sound origin to the two microphones. These pairs produce three independent TDoA values, which could uniquely locate the sound origin in a 3D space. We measure the TDoA in terms of the

[2]Note that the sectional view of the human vocal tract in Figure 5 (b) is on the Y-Z plane.



(a) Layout of Microphones (b) Localized phoneme sounds

Figure 5: Phonemes localization using microphone array.

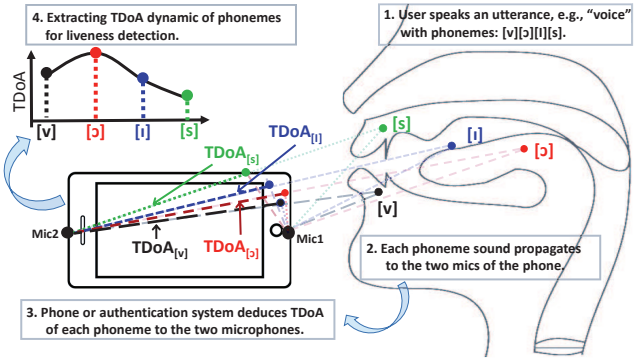


Figure 6: Illustration of phoneme localization using a single phone.

number of delayed samples to the two microphones. As we use 192kHz for recording, the TDoA ranging resolution is 1.77mm. Before the phoneme localization, we test the localization accuracy by emitting chirp sounds at different fixed locations in front of the microphones. We observe that it produces an averaged localization error within 2mm.

We recruit two participants to pronounce each phoneme sound in front of the microphone array multiple trials. Figure 5 (b) illustrates the localized phoneme sound origins for one participant. It shows the sectional view of human vocal tract on Y-Z plane. The red dots show the localized vowel sound origins, whereas the green dots show these of consonant ones. We obtain several important observations from Figure 5. First, the located sound origins of vowels match the tongue positions very well. For example, the vowels connected by the dotted lines in Figure 5 (b) have similar relative positions and overall shape as that of the vowel chart in Figure 2 (b). This is because the tongue position is the deterministic factor of vowel production. Second, some of the consonants have the origins close to the place of articulation, while others are significantly affected by the manner of articulation. For instance, [s],[z] and [t] have the localized sound origins close to alveolar, which is the place of articulation of these sounds, whereas [m] is located in the nasal cavity where the airflows out (i.e., manner of articulation). Moreover, we observe the located phoneme origins are mainly distributed within the mouth and nasal cavities with the size of about 4cm by 4cm, and they show little changes in X axis (i.e., lateral direction of mouth). We also find that different participants produce different localized sound origins for the same phoneme due to the individual diversity in the human vocal tract (e.g., shape and size) and the habitual way of pronouncing phonemes.

3. SYSTEM DESIGN

In this section, we introduce our system design and its core components and algorithms.

3.1 Approach Overview

The key idea underlying our liveness detection system is to perform TDoA ranging for a sequence of phoneme sounds at the two microphones on the phone. As illustrated in Figure 6, a user first speaks an utterance, say "voice" to the phone that closely placed to the user's mouth. Each phoneme sound (i.e., [v] [ɑ] [I] [s] in the example) is then emitted from the user's vocal system and picked up by the two microphones of the phone with stereo recording. The phone processes the recorded sound to deduce the TDoA of each phoneme sound to the two microphones. As most phoneme sounds have measurable TDoA differences to the two microphones, a sequence of phonemes will produce series of TDoA with various values, as shown in Figure 6. We refer to the changes in TDoA measurements as "TDoA dynamic", which is then used for liveness detection.

In particular, the measured TDoA dynamic will be compared with the one extracted when the user enrolled in the system. A live user is detected if the similarity score exceeds the pre-defined threshold. Under playback attacks, the measured TDoA dynamic will be very different from that of a live user due to different sound production systems (i.e., loudspeaker v.s. human vocal system). Under replace attacks, it is extremely unlikely, if not impossible, for an adversary to place a stereo recorder (e.g., smartphone) very close, say 5cm, to the victim's mouth to collect voice samples. Due to the origins of the phoneme sounds are crowded in the mouth and nasal cavities as shown in Figure 5 (b), the TDoA dynamic diminishes rapidly with the increased distance between the recorder and the user's mouth. For example, if the phone is placed 30cm away from the user's mouth, the maximum achievable TDoA range among all phonemes is less than 1cm. With such a small range, most phonemes have the same TDoA measurement to the two microphones of the phone. The measured TDoAs under replace attack thus cannot match the one extracted when the user enrolled in the system.

Virtually all smartphones are equipped with two microphones and are capable of stereo recording. By leveraging a sequence of phoneme sounds in an utterance/passphrase, our approach relaxes the problem of locating each phoneme sound to tracking TDoA dynamic for live user detection. We thus enable the phoneme localization based liveness detection on a single phone without requiring any additional hardware.

Our system does require the user to place the smartphone close to the mouth with the same pose in both enrollment and authentication processes. The effects of different phones and phone displacement are studied in experiment evaluation. Moreover, data protection mechanisms or secure communication protocols should be in place to prevent an attacker from obtaining the plain-text of TDoA dynamic and the dual-channel audio samples [18]. For example, TDoA dynamic could be extracted locally without storing the dual-channel audio sample, and only the encrypted one-channel

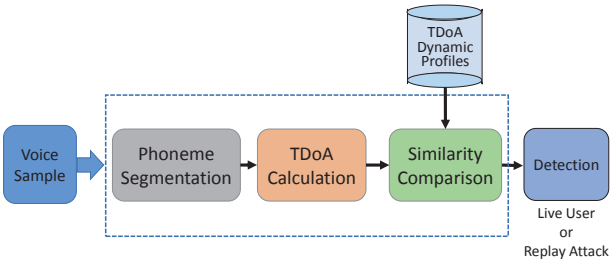


Figure 7: The flow of our liveness detection system.

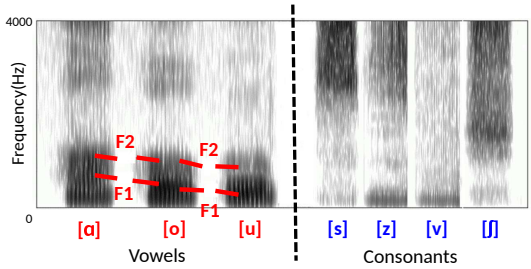


Figure 8: Example: spectrogram of phonemes.

audio sample together with the encrypted TDoA dynamic are transmitted or used for verification and liveness detection.

3.2 System Flow

Realizing our system requires four major components: *Phoneme Segmentation*, *TDOA Calculation*, *Similarity Comparison*, and *Detection*. As shown in Figure 7, the voice sample acquired by two microphones first passes through phoneme segmentation, which extracts phonemes existing in the voice sample. In particular, we combine Hidden Markov Modeling techniques to perform forced alignment on the words recognized from the voice sample to identify each phoneme sound. The words in the voice sample are recognized by acoustic modeling and language modeling algorithms.

Next, the TDOA calculation component is used to calculate the number of delayed samples of each phoneme sound to the two microphones. As acoustic signals can be easily distorted due to multipath propagation, simply correlating phonemes between two channels will result in large error. To address this challenge, we adopt generalized cross-correlation and heuristic-based phase transform weighting approaches for accurate TDoA estimation.

After that, the similarity comparison component measures the similarity of the calculated TDoA dynamic to the one stored in the system. It results in a similarity score, which is then compared with a pre-defined threshold. If the score is larger than the threshold, a live user is detected, otherwise a replay attack is declared. The detection result can be then combined with the traditional voice authentication system to verify the claimed identity of a user.

3.3 Phoneme Segmentation

The underlying principle for phoneme segmentation is that the sound of a phoneme contains a number of different overtone pitches simultaneously, known as formants [23]. By analyzing the sound spectrogram, we are able to discover these overtone pitches or formats to identify each individual phoneme sound. Although the most informative formants are the first three formants, the two first formants, F1 and

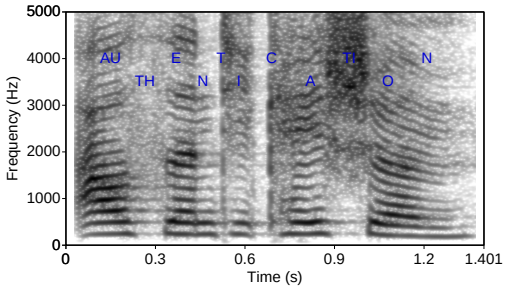


Figure 9: Example of segmented phonemes.

F2, are enough to disambiguate the vowel. As illustrated in Figure 8, it is easy to observe the first two formants, F1 and F2, which contribute to the overtone of each vowel most. It is thus feasible to segment different vowels by looking at the F1 and F2 in the spectrogram. Unlike vowels, consonants' spectrograms display as random mixture of different frequencies, as showed in Figure 8. This static noise-like sound makes it difficult to accurately identify each consonant by simply utilizing formants. We thus adopt forced alignment by using HMM (Hidden Markov Models), which aligns the input voice spectrogram with existing voice samples to distinguish different consonants [20].

In particular, we first recognize the words existing in the voice sample, which could be done by using automatic speech recognition (ASR). We use advanced CMUSphinx [29] to automatically recognize each word in the user's voice sample. More specifically, the voice sample is first parsed into features, which are a set of mel-frequency cepstrum coefficients (MFCC) that model the human auditory system. Then, the MFCCs are combined together with the dictionary, acoustic model, and language model to recognize the words in the voice sample [29].

Given the recognized words, we utilize MAUS as primary method for phoneme segmentation and labeling [21]. In particular, the recognized words are first transferred into expected pronunciation based on standard pronunciation model (i.e., SAMPA phonetic alphabet). Then, the generated canonical pronunciation together with the millions of possible accents of users yield a probabilistic graph including all possible hypotheses and the corresponding probabilities. At last, the system searches the graph space for the path of phonetic units that have been spoken with highest probability using a Hidden Markov Model. Outcomes of the search are segmented and labeled phonetic units. Figure 9 illustrates one example of the resulted phoneme segmentation when one user pronounces the word "authentication". We observe that the segmentation accurately captures both the vowels and the consonants.

3.4 TDOA Calculation

The basic idea of TDoA calculation is to count the number of delayed samples to the two microphones by correlating each segmented phoneme sound between smartphone's two channels. Let's denote mic_1 and mic_2 as the two microphones/channels of the phone, and Δt as the TDoA of one phoneme sound to the two microphones. Given the phoneme sound $mic_1(t)$ recorded at mic_1, we correlate such phoneme sound to the sound signal $mic_2(t+d)$ recorded at the mic_2, with d varying from 0 to $N-1$. Once the best match is found, the corresponding d value is the number of delayed samples between mic_1 and mic_2. In particular, such correla-

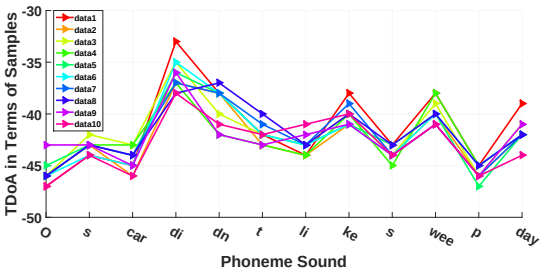


Figure 10: TDoAs of one passphrase for 10 trials.

tion can be done by using a cross-correlation technique [25], as shown below:

$$CC(d) = \frac{\sum_i [(mic_1(i) - \overline{mic_1(i)}) * (mic_2(i+d) - \overline{mic_2(i+d)})]}{\sqrt{\sum_i (mic_1(i) - \overline{mic_1(i)})^2} \sqrt{\sum_i (mic_2(i+d) - \overline{mic_2(i+d)})^2}}, \quad (1)$$

The TDoA Δt can be obtained as:

$$\Delta t = \underset{d}{\operatorname{argmax}}\, CC(d), \quad (2)$$

However, simply applying the cross-correlation method results in an inaccurate estimation of Δt due to the multipath propagation and reverberation effect of acoustic signals. To improve the accuracy, we further utilize generalized cross correlation with phase transformation techniques (PHAT) [22]. By adding a weighting function into cross correlation calculation process, it suppresses the frequency components whose power spectra carry intense additive noises. Meanwhile, PHAT utilizes the cross-power spectral density of two different acoustic signals to improve the system's robustness to reverberation effect. Existing work has shown PHAT can further mitigate the spreading effect that caused by uncorrelated noises at two microphones [22].

Figure 10 shows one example of the TDoA values when one participant performs 10 trials of authentication with the passphrase "Oscar didn't like sweep day". The X axis shows each phoneme sound, whereas Y axis shows the TDoA in terms of number of delayed samples. We observe that TDoA dynamics of these trials are highly similar and stable, with only 1 to 2 samples variation under 192kHz sampling rate. The results show that TDoA calculation is able to catch the user's unique speech production system accurately.

3.5 Similarity Comparison

Once the TDoA dynamic is extracted, we first normalize these TDoA values to the same scale as those stored in the user profile. Such normalization is used to deal with the issues of device diversity and phone displacement. The phone a user used to enroll in the system could be different from the one he/she used for authentication. As different phones differ in size or distance between the two microphones, the absolute TDoA values of the same phoneme could be different. Similarly, if the user places the phone at a location slightly different from that when he/she enrolled in the system, the absolute TDoA values vary slightly. Normalizing the TDoAs to the same scale could effectively mitigate these issues.

To compare the similarity of the TDoA dynamic with the user profile, we utilize both the correlation coefficient and the probability. In particular, the correlation coefficient

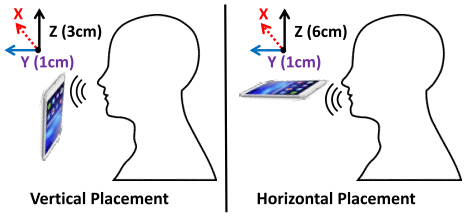


Figure 11: Two different phone placements diagram.

measures the degree of linear relationship between two sequences [35]. Other than calculating the absolute difference, it quantifies the similarities in the changes of two sequences. The correlation coefficient ranges from -1 to +1. A value of near +1 indicates a high degree of similarity, whereas a value near 0 indicates a lack of similarity.

For the probability based method, we assume the TDoA ranging error of each phoneme follows an independent standard Gaussian distribution. Given the TDoA value $TDoA_i$ in the extracted TDoA dynamic, the probability that it matches the one in the user profile is represented as:

$$P(TDoA_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(TDoA_i - \overline{TDoA_i}\right)^2} \quad (3)$$

whereas σ is the standard deviation of the error and $\overline{TDoA_i}$ is the corresponding TDoA value in the user profile. During the user enrollment phase, we ask each user to speak a passphrase three times to extract the averaged TDoA and the standard deviation of each phoneme for similarity comparison. Given the probability value of each phoneme, we simply average the probability values of all phonemes as the indicator of the similarity score.

Correlation coefficient and probability are two metrics targeting on different characteristics of the TDoA dynamic. We refer to the former as *Correlation*, and latter as *Probability*. Moreover, we develop a combined scheme that simply combines the similarity scores of the correlation and probability based methods. We refer to such a method as *Combined method*, which takes advantages of both the correlation coefficient and the probability.

4. PERFORMANCE EVALUATION

In this section, we evaluate our liveness detection system under replay attacks including both *playback* and *replace* attacks[3]. We also evaluate the robustness of our system to different types of phones, sampling frequencies, phone displacements, and lengths of passphrases.

4.1 Experiment Methodology

Phones and Placements. We evaluate our system with three types of phones with different sizes and audio chipsets. In particular, we experiment with Samsung Galaxy Note3, Galaxy Note5 and Galaxy S5. The distance between the two microphones (i.e., one on the top and one at the bottom) for stereo recording is about 15.1cm for Note3, 15.3cm for Note5, and 14.1cm for S5. The audio chipset of Note3 is Qualcomm Snapdragon 800 MSM8974, whereas it is Wolfson WM1840 for Note5, and Audience's ADNC ES704 for S5. The operating system of these phones is Android 6.0 Marshmallow, which enables the phones to perform stereo recording at 48kHz, 96kHz and 192kHz sampling frequencies. These frequencies represent ranging resolutions of 7.08mm,

[3]This project has obtained IRB approval.

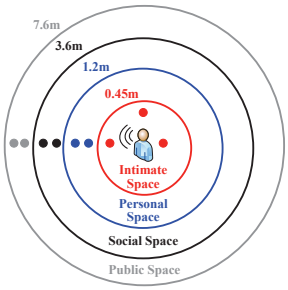


Figure 12: Illustration of locations of replace attacks and different types of social distances.

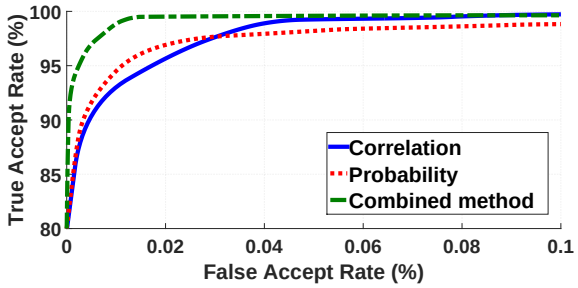
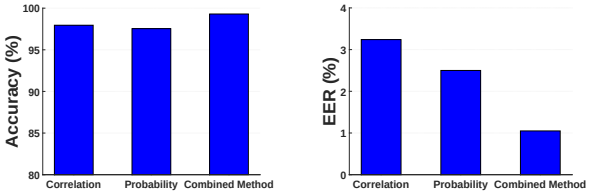


Figure 13: Playback Attacks: ROC curves under different methods.



(a) Accuracy (b) EER

Figure 14: Playback Attacks: Accuracy and EER.

3.54mm, and 1.77mm, respectively. We use 192kHz as our primary sampling frequency and present the corresponding results unless otherwise stated. We also experiment with two types of phone placements, as shown in Figure 11. One is vertical placement with the phone placed close to user's mouth vertically. We call such placement our primary placement and present the performance of such placement unless otherwise specified. For the vertical placement, the phone is about 3cm and 1cm away from user's mouth on Z and Y axis, respectively. The other one is horizontal placement with the phone placed close to the user's mouth horizontally. The phone is about 6cm and 1cm away from user's mouth on Z and Y axis, respectively. We choose these placements because they have relatively large achievable TDoA ranges, which is discussed in Section 5.

Data Collection. Our experiments involve 12 participants including 6 males and 6 females whose ages range between 25 to 38. These participants are either graduate students or university researchers, who are recruited by emails. The participants are informed of the purpose of our experiments and are required to act as if they were conducting voice authentication. Each participant chooses 10 different passphrases of their own and performs 10 times legitimate authentications for each passphrase after enrollment. To enroll in the system, each participant speaks a passphrase three times to extract the averaged TDoA and the standard deviation of each phoneme for similarity comparison. For online verification, users only speak the passphrase once. Each participant speaks the passphrase with her/his habitual way of speaking. The lengths of the passphrases are ranging from 2 words to 10 words with proximately half of them are 2-4 words, one quarter of them are 5-7 or 8-10 words. The experiments are conducted in both the office and home environments with background and ambient noises, such as people chatting and HVAC noise.

Attacks. We experiment with two types of replay attacks: *playback attacks* and *replace attacks*. For playback attacks, we replay participants' voice samples in front of the smartphone that performs stereo recording for authentication. We utilize three different types of loudspeaker including DELL AC411 wireless speaker system, Samsung Galaxy note5 and S5 speakers, to replay each pre-recorded voice sample. In addition, half of the playback attacks are conducted with stationary loudspeakers that are within 10cm away from the smartphone (i.e., *Static Playback Attacks*); while the other half are conducted with mobile loudspeakers targeting on mimicking TDoA changes of users by moving the loudspeakers around the smartphone (i.e., *Mobile Playback Attacks*).

For replace attacks, we place a smartphone with stereo recording close to the target user when the user is performing legitimate voice authentication. In such cases, the adversary obtains a two-channel voice sample of the target and then uploads that directly to the voice authentication system. The only difference between the two-channel voice sample obtained by the adversary and the one in legitimate authentication is the recording distance. We adopt the Edward T. Hall's proxemics theory [15] to emulate how close an adversary could place the phone next to the user's mouth. As shown in Figure 12, the minimum distances between people are categorized by the relationship and types of interactions between them. It includes intimate distance, personal distance, social distance, and public distance. With such a guideline, we chose the recording distances between the attacker's phone to the user's mouth as 30cm, 50cm, 100cm, 150cm, 200cm, 300cm, and 450cm, which simulates different types of relationships. We also consider the circumstances where the attacker could hide behind or at the side of the user. The recording distances for such cases are limited by the size of user's head, and are around 40 cm and 25 cm away to user's mouth, as shown in Figure 12.

Metrics. We use the following metrics to evaluate the performance of our liveness detection system. *False Accept Rate (FAR)*: the probability that the liveness detection system incorrectly declares a replay attack as a live user. *False Reject Rate (FRR)*: the probability that our system mistakenly classifies a live user as a replay attack. *Receiver Operating Characteristic (ROC)*: it describes the relationship between the True Accept Rate (i.e., the probability to identify a live user as a live user) and the FAR when varying the detection threshold. *Equal Error Rate (EER)*: it shows a balanced view of the FAR and FRR and is defined as the rate at which the FAR equals to the FRR. *Accuracy*: it measures the overall probability that the system could detect a live user and reject a replay attack.

4.2 Overall Performance

We first evaluate the overall performance of our liveness detection system under two types of replay attacks: *playback attacks* and *replace attacks*.

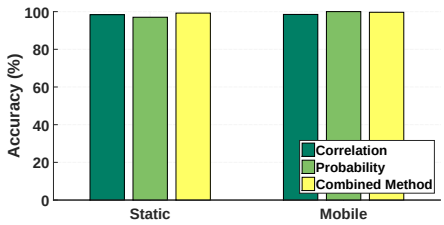


Figure 15: Static and Mobile Playback Attacks: Accuracy under different methods.

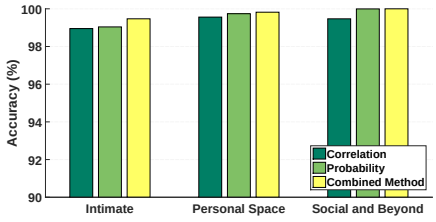


Figure 16: Replace Attacks: Accuracy of different methods with different social distances.

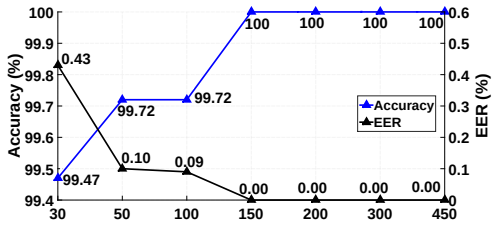


Figure 17: Replace Attacks: EER and Accuracy of Combined method under different distances.

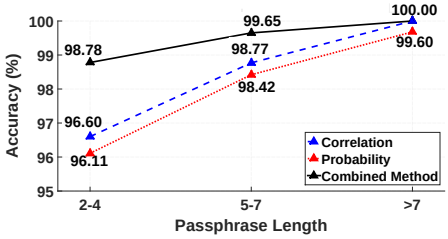


Figure 18: Accuracy under different lengths of passphrase.

Playback Attack. Figure 13 shows the ROC curves of different methods in detecting live users under playback attacks. We observe that our system is highly effective in detecting live users and rejecting playback attacks. These three methods provide more than 94% detection rate with less than 1% FAR. In particular, the correlation and probability based methods have comparable performance. The correlation method provides a detection rate of 95% with 1% FAR. The combined method has the best performance and results in over 99% detection rate with less than 1% false accept rate.

Moreover, Figures 14 depicts the overall accuracy and EER of different methods under playback attacks. We observe that the combined method provides the best accuracy and EER, which are 99.30% and 1.05% respectively. The correlation method produces an accuracy of 97.95%, which is slightly better than that of the probability method (i.e., 97.54%). However, probability method results in a better EER than that of the correlation method. In particular, probability method has an EER of 2.50% and correlation method has an EER of 3.24%. The above results show that VoiceLive is highly accurate in detecting live users under playback attacks, and the combined method provides the best results since it takes advantages of both the correction and the probability based methods.

We next take a closer look at how our system performs under static and mobile playback attacks. In our experiments, we observe the static playback attacks produce similar TDoA values for different types of phoneme sounds. Although playback attacks under mobile scenarios could result in TDoA changes, the resulted changes in TDoA cannot match with the ones in the user profile. It is because the attacker couldn't mimic the sound position transition the same as that of the human vocal system. As shown in Figure 15, our system is highly effective in live user detection under both static and mobile playback attacks. The combined method achieves 99.2% accuracy under static scenarios and 99.65% accuracy under mobile cases.

Replace Attack. We next evaluate the effectiveness of our system in defending against the replace attacks. Figure 16 illustrates the accuracy of different methods with replace attacks conducted under different social distances. In particular, the testing positions of replace attacks fall into three categories: *intimate* (<45cm), *personal space* (45cm to 1.2m), and *social and beyond*(>1.2m). We observe that our system can effectively detect the live users and reject the replace attacks under each category of social distances. For example, the combined method provides 99.47% detection accuracy under intimate relationship, 99.82% under personal relationship, and 100% under social relationship and public space. And all the methods provide over 98.95% detection accuracy across different categories.

Figure 17 shows the details on the accuracy and EER of the combined method under each social distance. We find that both the accuracy and EER are improved with an increased social distance. In particular, the EER decreases from 0.43% to 0% and the accuracy is improved from 99.47% to 100% when the distance is increased from 30cm to 150cm. When the attacker is further away, our system can detect all the live user cases and reject all the replace attacks. This is because when increasing the distance between the phone and user's mouth, the TDoA dynamic diminishes rapidly.

We also investigate the replace attacks launched from 25cm behind the user and 40cm from the side of user. The EER of the combined method under these two cases are 0.33% and 0%, respectively. Such results are comparable to the EER in Figure 17. This shows our system is capable of detecting replace attack conducted from different directions.

4.3 Impact of Passphrase Length

Generally, a passphrase with longer length provides stronger security. It also produces more phoneme sounds that generate more changes in the TDoA measurements. We thus study the performance of our system with different lengths of passphrases. In particular, we sort all the passphrases into three categories including short passphrases with 2 to 4 words, appropriate passphrasses with 5 to 7 words, and long passphrases with 8 to 10 words. Note that researchers and professionals in voice authentication suggest that a passphrase should contain at least 5 words so as to provide sufficient security level [7].

Figure 18 and Figure 19 illustrate the accuracy and EER

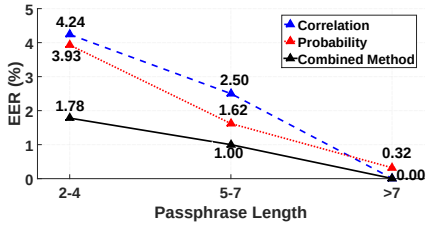


Figure 19: EER under different lengths of passphrase.

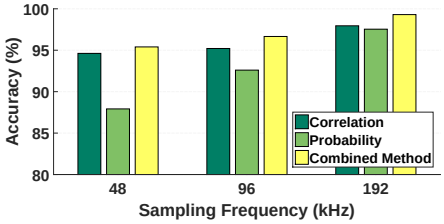


Figure 20: Accuracy under different sampling frequencies.

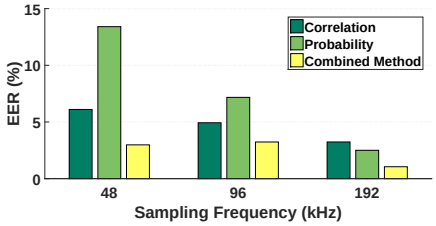


Figure 21: EER under different sampling frequencies.

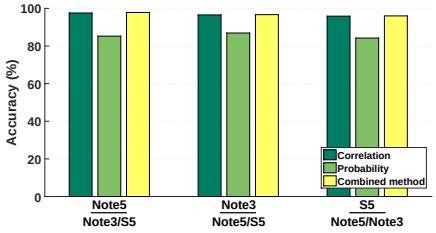


Figure 22: Accuracy of using one phone as enrollment and the other two as online authentication.

for different lengths of passphrases, respectively. We observe that both the accuracy and EER are improved for all the methods when we increase the length of the passphrase. In particular, the accuracy is improved from 98.78% to 100% and the EER is reduced from 1.78% to 0% for combined method, when we increase the length from 2-4 words to more than 7 words. In addition, with an appropriate length of passphrase (i.e., 5 to 7 words), the combined method results in 99.65% accuracy and 1% EER. The results confirm our observation that a longer passphrase leads to more TDoA changes of phoneme sounds, which improves the performance of the live user detection.

4.4 Effect of Sampling Frequency

As not all the smartphones are installed with the latest OS, older version of OSs can only support the standard sampling frequency at 48kHz or 96kHz. We thus study how robust our system is to lower sampling frequencies. Figure 20 and Figure 21 show the accuracy and EER under different sampling frequencies respectively. We observe that our system works effectively under all of these three sampling frequencies. Although with higher sampling rates, it does have better accuracy and EER due to higher ranging resolution. In particular, for combined method the accuracy is over 95% under 48kHz, and over 97% under 96kHz, whereas the EER is at around 3% for both 48kHz and 96kHz. These results show that our liveness detection system could work with both the state-of-the-art smartphones as well as low-end smartphones.

4.5 Impact of Different Phones

As one user may use one phone to enroll in the system but uses another one to perform online authentication, we study how our system behaves under different phones. Specifically, we experiment with users to use either Note5, Note3, or S5 to enroll in the system and then utilize the other two for online authentication. These three types of phones differ in size and audio hardware as described in the experimental setup. Figure 22 shows the accuracy of different methods when using one phone as enrollment and the others as online authentication. We observe that our system still provides accurate liveness detection. In particular, the combined method results in accuracy of 97.82%, 96.67%, 96% when using Note5, Note3, and S5 as the phone for enrollment while the other two for authentication, respectively. Moreover, we observe that these three methods have comparable performance no matter which phone is used for enrollment or authentication. Although the accuracy is slightly higher when using the same phone (i.e., about 99%), our system still produces very accurate detection results with different phones. Such observations show that our system is robust and compatible to different phone models.

4.6 Robustness to Phone Displacement

When performing online authentication, our system requires the user to place the phone at a similar position to that when the user enrolled in the system. We thus study our system's performance if there exists displacement of the phone between the position for enrollment and online authentication. Specifically, we experiment with different degrees of phone displacements when performing authentication, i.e., 1cm, 2cm, and 3cm away from the position that user enrolled in the system. Such displacements occur on each axis: Left (X axis), Down (Z axis), and Forward (Y axis). Figure 23 and Figure 24 show the accuracy and EER under different degrees of phone displacements, respectively. We observe that although a higher degree of displacement results in lower accuracy and higher EER, our system overall still provides accurate detection results: the accuracy is more than 98% for all the displacements and EER is also maintained at a very low rate, ranging from 1% to 3%, when using combined method. Moreover, we find that our system is more sensitive to the displacements on Y axis (i.e., Forward) and less sensitive to X axis (i.e., Left). This is because by moving the phone forward, the maximum achievable TDoA range will be reduced quickly with the increased distance, as shown in Section 5. The results may also indicate VoiceLive is not sensitive to the small movements of the phone (e.g., hand movements). In addition, the time duration of speaking a passphrase is usually about 2-3 seconds. The movements of the phone within such a short duration are usually small and may have limited effect.

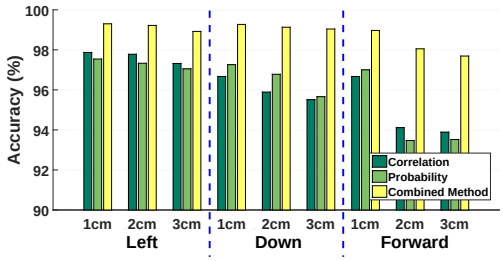


Figure 23: Accuracy under different degree of phone displacement.

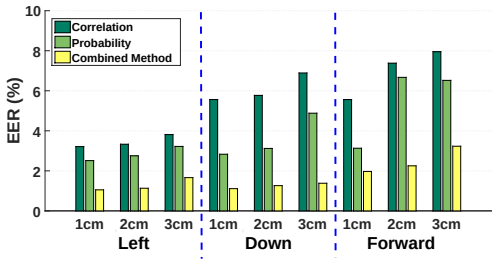


Figure 24: EER under different degrees of phone displacement.

4.7 Effect of Phone's Placement

Different users might have different ways to place the phone close to their mouths during the authentication process. We thus compare the performance of our system under two types of placements, vertical and horizontal, as described in experimental setup. Figure 25 illustrates the accuracy comparison of these two placements. We observe that our system achieves very high accuracy for both placements, with the accuracy slightly higher when the phone is placed vertically. Specifically, the accuracy under horizontal placement is 97.41%, 97.26%, and 99.13% for correlation, probability, and combined method respectively. Figure 26 shows the EER under two displacements. We have similar observation to that of the accuracy. In particular, EER is 3.3%, 2.69%, and 1.33% for correlation, probability, and combined method respectively. Results show that our system works very well for different phone placements including both horizontal and vertical placements.

5. DISCUSSION AND FUTURE WORK

Achievable TDoA Range. The achievable TDoA range is determined by the distance between two microphones and is affected by the relative position between the phone and user's mouth. Figure 27 shows the achievable TDoA range with Samsung Galaxy Note3 by using the sound origin model we built in Figure 5. The distance between two microphones is 15.1cm. Figure 27 plots the sectional view on Y-Z plane and the coordinate (0,0) is the location of the mouth. Each (y, z) point in vertical placement indicates the center of the phone when place the phone vertically, whereas it represents the bottom microphone of the phone when place phone horizontally. The color at each position represents the achievable TDoA range at that position. As we can see from Figure 28, the maximum achievable TDoA range is around 6cm for vertical placement, whereas it is about 4cm for horizontal placement if we place the phone very close to user's mouth. The reason we cannot achieve maximum TDoA range as the distance between two microphones is that the origin of the

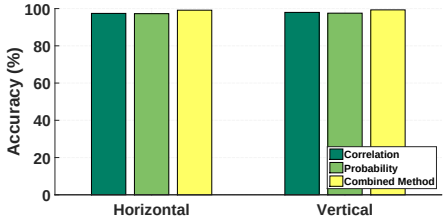


Figure 25: Accuracy of horizontal and vertical placements.

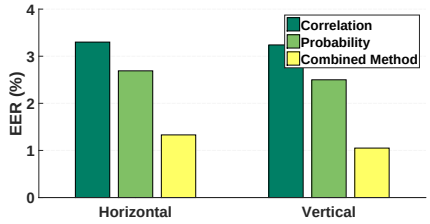
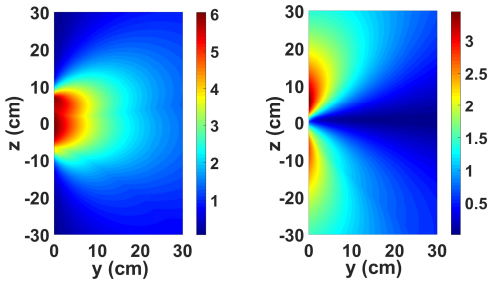


Figure 26: EER of horizontal and vertical placements.

phoneme sound is crowded in user's mouth and nasal cavities (i.e., all are at similar directions to the two microphones). Such maximum achievable TDoA ranges (i.e., 6cm and 4cm) could ideally distinguish 33 and 23 different phoneme sounds under 192kHz sampling rate, respectively. We also find that the achievable TDoA range decreases rapidly when increasing the distance between the phone and the mouth. For example, with the phone placed at 30cm away from user's mouth, the achievable TDoA range decreases to less than 1cm, which makes it hard to capture any TDoA dynamic of a passphrase. This is why our system is robust to the replace attack, where an adversary attempts to record the TDoA dynamic under different social distances.

Potential Active Attacks. In our experiments, we only evaluate our system under the scenarios that an adversary uses similar recording hardware as the one used by the legitimate users. However, it is possible for an attacker to use advanced hardware to record the voice samples and further deduce the TDoA dynamic that can match the victim's profile. In particular, an attacker can leverage a microphone array to locate each phoneme within the victim's vocal system. As the maximum achievable TDoA range decreases rapidly with the increased distance between the recorder and the user's mouth, it requires the microphone array to support an ultra-high sampling rate so as to have sufficient ranging resolution to uniquely locate each phoneme. For example, with the microphone array placed 30cm away from the user, the maximum achievable TDoA range is less than 1cm. To uniquely locate each phoneme, the ranging resolution should be at least 0.2mm, which is ten times of that supported by 192kHz. Current professional digital recorders (e.g., DirectStream Digital (DSD) recorders that worth thousands of dollars and have the sizes similar to a desktop mainframe) that support 2.8224MHz and 5.6448MHz sampling rates can be leveraged to locate each phoneme without placing the recorder very close to victim's mouth.

After locating each phoneme, the attacker can deduce the TDoA dynamic of the victim based on the relative position between the phone and the victim's mouth. This does require the attacker to observe how the phone is placed to the victim's mouth. Given the obtained TDoA dynamic,



(a) Vertical placement (b) Horizontal placement

Figure 27: The achievable TDoA range under both vertical and horizontal placements.

the attacker is further required to reproduce the voice samples that satisfy the TDoA constraints. It could be done by creating a synthetic two-channel audio stream. With such an audio stream, the attacker can either conduct a replace attack or a playback attack to bypass the VoiceLive.

In our future work, we will study the feasibility of conducting such active attacks. Particularly, we will evaluate whether or not current acoustic localization systems could achieve the level of localization accuracy required in the active attacks. The potential countermeasure is to detect the synthetic two-channel audio stream. VoiceLive could integrate with existing speaker verification techniques, such as the higher order Mel-cepstral coefficients [14, 12], which are able to detect speech synthesize attacks. We will evaluate the effectiveness of detecting the synthetic two-channel audio stream with these techniques in our future work.

Extension to Text-Independent System. As a text-independent system operates on arbitrary utterances, we cannot rely on the TDoA dynamic of a passphrase for liveness detection. However, a text-independent system requires collecting a large number of utterances from the user to train its speaker models. We therefore can extract the TDoA value of each phoneme sound to build a model similar to that of the Figure 5 by re-using the training data when the system trains the speaker models. During the online authentication phase, we extract the TDoA value of each phoneme from the incoming utterances and then could build another model. Such a model (could be a sub-model of the trained model) can then be matched with the one trained during the training phase. It is thus still possible to use the location of each phoneme sound for liveness detection in text-independent systems.

Diversity in Human Vocal System. An individual's vocal system differs in the shape and size of the larynx, nasal passages and vocal tract. In addition, different individuals have their own habitual ways of pronouncing the same word, which results in different cadences, accents and pronunciations. We thus investigate how similar are the extracted TDoA dynamics for different users with the same passphrase. Figure 28 depicts the similarity of the extracted TDoA dynamics for the same passphrase between four users: A, B, C, and D. Each user speaks the same passphrase 10 times, and we measure the similarity within each user and between users using Pearson correlation coefficients. We observe that the correlation coefficients for the same user under different trials are very high, at around 0.9, whereas they are generally below 0.6 between different users. This indicates that the diversity in TDoA dynamic does exist, which is similar to that of individual vocal system and user's habitual way of pronouncing. It also shows that it is promising to use the TDoA dynamic or the location of phoneme as a new biometric trait for user authentication. In our future work, we will study the possibility of verifying or identifying the speaker by making a model using the location of the phoneme sound.

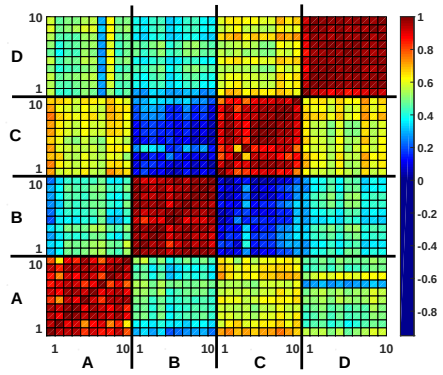


Figure 28: Similarity of TDoA dynamics between different users.

6. RELATED WORK

In recent years, more and more mobile devices and apps are embracing voice biometric for mobile authentication. However, voice authentication is subject to spoofing attacks, as indicated in recent studies [16, 33, 14, 26]. Voice spoofing attacks can be divided into four categories, which are described below together with countermeasures.

Replay Attack. An adversary can spoof a voice authentication system by using a pre-recorded voice sample of the victim [24]. To defend against such attacks, Shang *et al.* propose to compare a new access voice sample with stored instances of past access attempts [31]. If this results in an extremely high similarity score, a replay attack is identified. As an alternative, Villalba *et al.* utilize the increased noise and reverberation of replaying far-field recordings for attack detection [32], whereas Wang *et al.* use the additional channel noise of the recording and loudspeaker for attack detection [33]. However, the effectiveness of these approaches is very limited in practice (e.g., the FAR rate could be as high as 17%.). Chetty and Wagner utilize video camera to detect lip movements for liveness detection [13], whereas Poss *et al.* aim to improve authentication accuracy by combining the techniques of a neural tree network and Hidden Markov Models [28]. Aley-Raz *et al.* develop a liveness detection system, which requires a user to repeat one or more random sentences prompted by the system for attack detection [10].

Impersonation Attack. It refers to attacks where an adversary tries to mimic the victim's voice without utilizing any computer or professional devices. Recent work shows that impersonation attack could be defended very efficiently by using advanced speaker models, such as GMM-UBM [11] and i-vector models [16]. Existing voice authentication systems with such advanced speaker models thus are resistant to impersonation attacks.

Speech Synthesize Attack. This type of attack indicates an attacker has the ability to synthesize the victim's voice by utilizing speech synthesize technologies. Earlier work done by Lindberg and Blomberg [24] shows that the FAR can be increased to as high as 38.9% with less sophisticated speaker models. Recent work done by De Leon *et al.* shows that by adopting both GMM-UBM and SVM tech-

nologies, voice authentication systems are able to lower the FAR of the system to 2.5% [14]. Also, Chen et al. [12] show that by employing higher order Mel-cepstral coefficients, the EER can be lowered to 1.58%.

Voice Conversion Attack. It aims at manipulating or converting existing voice samples from other users so that they would resemble the target's voice. In the early work, researchers demonstrate such attacks can significantly affect the authentication system [19]. Recent studies by Mukhopadhyay *et al.* show that current speaker verification systems based on UBM-GMM and ISV speaker models are vulnerable to voice conversion attacks [26]. To defend against voice conversion attacks, Wu *et al.* [34] developed an authentication system with PLDA component that could achieve 1.71% FAR, whereas Alegre *et al.* utilize PLDA and FA technologies, which result in the FAR rate of 1.6% [9].

7. CONCLUSION

In this work, we developed a liveness detection system for voice authentication that requires only stereo recording on smartphones. Our system VoiceLive is practical as no additional hardware is required during the authentication process. VoiceLive performs liveness detection by measuring TDoA changes of a sequence of phoneme sounds from the two microphones of a smartphone. It distinguishes a live user from a replay attack by comparing the TDoA changes of the input utterance to the one stored in the system. Our experimental evaluation demonstrates the viability of distinguishing between a live user and a replay attack under various experimental settings. Our experimental results also show the generality of our system, as we experiment with different phone types, placements and sampling rates. Overall, VoiceLive can achieve over 99% accuracy, with an EER as low as 1%.

8. ACKNOWLEDGEMENTS

We thank our shepherd, Dr. Nitesh Saxena, and the anonymous reviewers for their insightful feedbacks. This work was partially supported by the National Science Foundation Grants CNS-1514436, SES-1450091, CNS-1505175, CNS-1652447 and CNS-1514238.

9. REFERENCES

[1] Android voice recognition. http://www.popsci.com/new-android-can-recognize-your-voice.
[2] Google smart lock. https://get.google.com/smartlock/.
[3] Hsbc offers voice biometric. http://www.bbc.com/news/business-35609833.
[4] Mobile voice biometric security. http://voicevault.com/hsbc-embraces-mobile-voice-biometric-security-technology/.
[5] Saypay technologies. http://saypaytechnologies.com/.
[6] Vocalpassword. http://www.nuance.com/ucmprod/groups/enterprise/@web-enus/documents/collateral/nc_015226.pdf.
[7] Voicekey mobile applications. http://speechpro-usa.com/product/voice_authentication/voicekey#tab2.
[8] Wechat voiceprint. http://blog.wechat.com/2015/05/21/voiceprint-the-new-wechat-password/.
[9] F. Alegre, A. Amehraye, and N. Evans. A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns. In *IEEE BTAS*, 2013.
[10] A. Aley-Raz, N. M. Krause, M. I. Salmon, and R. Y. Gazit. Device, system, and method of liveness detection utilizing voice biometrics, May 14 2013. US Patent 8,442,824.
[11] T. B. Amin, J. S. German, and P. Marziliano. Detecting voice disguise from speech variability: Analysis of three glottal and vocal tract measures. *The Journal of the Acoustical Society of America*, 2013.
[12] L.-W. Chen, W. Guo, and L.-R. Dai. Speaker verification against synthetic speech. In *2010 IEEE Chinese Spoken Language Processing (ISCSLP)*, 2010.
[13] G. Chetty and M. Wagner. Automated lip feature extraction for liveness verification in audio-video authentication. *Proc. Image and Vision Computing*, 2004.
[14] P. L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga. Evaluation of speaker verification security and detection of hmm-based synthetic speech. *IEEE Processing of Audio, Speech, and Language*, 2012.
[15] E. Hall. Handbook for proxemic research. *Anthropology News*, 1995.
[16] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, T. Leino, and A.-M. Laukkanen. I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry. In *INTERSPEECH*, 2013.
[17] A. Jain, R. Bolle, and S. Pankanti. *Biometrics: personal identification in networked society*. Springer Science & Business Media, 2006.
[18] T. Kevenaar. Protection of biometric information. In *Security with Noisy Data*. 2007.
[19] T. Kinnunen et al. Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech. In *IEEE ICASSP*, 2012.
[20] A. Kipp, M.-B. Wesenick, and F. Schiel. Automatic detection and segmentation of pronunciation variants in german speech corpora. In *IEEE ICSLP*, 1996.
[21] T. Kisler, F. Schiel, and H. Sloetjes. Signal processing via web services: the use case webmaus. In *Digital Humanities Conference*, 2012.
[22] C. H. Knapp and G. C. Carter. The generalized correlation method for estimation of time delay. *IEEE Processing of Acoustics, Speech and Signal*, 1976.
[23] P. Ladefoged. *A course in phonetics*. Hardcourt Brace Jovanovich Inc. NY, 2014.
[24] J. Lindberg, M. Blomberg, et al. Vulnerability in speaker verification-a study of technical impostor techniques. In *Eurospeech*, 1999.
[25] J. Liu, Y. Wang, G. Kar, Y. Chen, J. Yang, and M. Gruteser. Snooping keystrokes with mm-level audio ranging on a single phone. In *ACM MobiCom*, 2015.
[26] D. Mukhopadhyay, M. Shirvanian, and N. Saxena. All your voices are belong to us: Stealing voices to fool humans and machines. In *European Symposium on Research in Computer Security*, 2015.
[27] J. P. Olive, A. Greenwood, and J. Coleman. *Acoustics of American English speech: a dynamic approach*. Springer Science & Business Media, 1993.
[28] J. C. Poss, D. Boye, and M. W. Mobley. Biometric voice authentication, June 10 2008. US Patent 7,386,448.
[29] M. K. Ravishankar. Efficient algorithms for speech recognition. Technical report, DTIC Document, 1996.
[30] M. A. Redford. *The handbook of speech production*. John Wiley & Sons, 2015.
[31] W. Shang and M. Stevenson. Score normalization in playback attack detection. In *IEEE ICASSP*, 2010.
[32] J. Villalba and E. Lleida. Detecting replay attacks from far-field recordings on speaker verification systems. In *Biometrics and ID Management*. 2011.
[33] Z.-F. Wang, G. Wei, and Q.-H. He. Channel pattern noise based playback attack detection algorithm for speaker recognition. In *IEEE ICMLC*, 2011.
[34] Z. Wu, T. Kinnunen, E. Chng, and H. Li. A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case. In *IEEE APSIPA ASC*, 2012.
[35] J. Yang, Y. Chen, and W. Trappe. Detecting spoofing attacks in mobile wireless environments. In *SECON*, 2009.