Hearing Your Voice is Not Enough: An Articulatory Gesture Based Liveness Detection for Voice Authentication

Linghan Zhang, Sheng Tan, Jie Yang
Florida State University
Tallahassee, Florida, USA
{lzhang,tan,jie.yang}@cs.fsu.edu

ABSTRACT

Voice biometrics is drawing increasing attention as it is a promising alternative to legacy passwords for mobile authentication. Recently, a growing body of work shows that voice biometrics is vulnerable to spoofing through replay attacks, where an adversary tries to spoof voice authentication systems by using a pre-recorded voice sample collected from a genuine user. In this work, we propose VoiceGesture, a liveness detection system for replay attack detection on smartphones. It detects a live user by leveraging both the unique articulatory gesture of the user when speaking a passphrase and the mobile audio hardware advances. Specifically, our system re-uses the smartphone as a Doppler radar, which transmits a high frequency acoustic sound from the built-in speaker and listens to the reflections at the microphone when a user speaks a passphrase. The signal reflections due to user's articulatory gesture result in Doppler shifts, which are then analyzed for live user detection. VoiceGesture is practical as it requires neither cumbersome operations nor additional hardware but a speaker and a microphone that are commonly available on smartphones. Our experimental evaluation with 21 participants and different types of phones shows that it achieves over 99% detection accuracy at around 1% Equal Error Rate (EER). Results also show that it is robust to different phone placements and is able to work with different sampling frequencies.

CCS CONCEPTS

• **Security and privacy** → *Biometrics*; *Mobile and wireless security*;

KEYWORDS

Voice authentication; Liveness detection; Articulatory gesture

1 INTRODUCTION

Biometrics has gained increasing attention and significance as it is a promising alternative to legacy passwords for user authentication. Among various biometric modalities (such as fingerprint, iris and facial), voice has wide applicability as it is the primary mode of communication, enabling biometric samples to be acquired remotely through existing landline, cellular and VoIP communication channels without additional hardware. Unlike other biometrics, voice biometrics has the advantage of natural integration with passwords or face authentication in mobile devices for multi-factor authentication. Over recent years, voice authentication has matured to become a low-cost and reliable method for authenticating users in a wide range of applications such as access control, forensics and law enforcement [54].

Particularly, with the advances of mobile technologies, voice authentication is becoming increasingly popular in a growing range of mobile applications. For instance, voice biometrics has been integrated with smartphone operating systems and mobile apps for secure access and login. Examples include Google's "Trusted Voice" for Android devices [10], Lenove's voice unlock feature for its smartphones [1], and Tencent's "Voiceprint" feature in WeChat for voice based app login [7]. Moreover, voice authentication has also been progressively deployed in e-commerce and mobile banking. For example, Saypay, a biometric authentication solutions provider, provides voice authentication services for online transactions in e-commerce [4]. And a considerable number of financial institutes, such as HSBC, USAA, National Australia, Citi and U.S. Bank, have started testing or are deploying voice recognition mobile apps and ATMs to allow customers to bank without requiring passwords or card swipes [3]. Voice authentication thus has increasingly gained interest in mass-market adoption, as also evidenced by the predicted market share of $184.9 billion in 2021 [12].

Recently, a growing body of research has demonstrated the vulnerability of voice authentication systems to spoofing through replay attacks [21, 24, 49, 51], where an adversary tries to spoof the authentication system by using a pre-recorded voice sample collected from a genuine user [48]. The replay attacks are easy to carry out, requiring neither sophisticated equipments nor specific expertise. They are also increasingly practical due to the wide availability of low-cost, high-quality recording and playback devices. The popularity of social media further makes it relatively easy for an adversary to obtain voice samples from the intended target user. Importantly, such low-cost and low-effort attacks have been shown to be highly effective in spoofing the voice authentication systems. For instance, simply replaying a pre-recorded voice command of a user could unlock her/his mobile devices that have voice-unlock feature (e.g., Android devices) [10]. An extensive study in 2017 shows that replay attacks increase the equal error rate (EER) of state-of-art voice authentication systems from 1.76% to surprisingly 30.71% [24]. Replay attacks thus pose serious threats to the voice authentication systems and have drawn much attention recently.

To defend against replay attacks, liveness detection system is required to distinguish between the legitimate voice samples of

CCS '17, October 30-November 3, 2017, Dallas, TX, USA
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-4946-8/17/10…$15.00
https://doi.org/10.1145/3133956.3133962

live users and the replayed ones. Traditional methods mainly rely on the acoustic characteristics of an input utterance. Such methods, however, are only effective when the input utterance contains significant additive or convolution noises, for example, when the voice samples are collected surreptitiously [40]. They fail when the recordings took in benign acoustic environments with high-quality recorders as such recordings are close to indistinguishable from the genuine ones [43]. An adversary could also obtain a copy of genuine voice recording and directly supply to the authentication system, bypassing the local microphone. Such high quality recordings and playbacks make it extremely hard, if not possible, for detecting replay attacks with only the acoustic characteristics. For instance, the replay attack detection 2017 challenge shows that current acoustic characteristics based detection methods only achieve EER of 24.65% on average [24]. The acoustic characterization based approaches therefore have very limited effectiveness in practice.

Current voice authentication service providers, such as VoiceVault [9] and Nuance [5], mainly rely on the challenge-response based approaches for liveness detection. In particular, the user is prompted to repeat a closed set of sentences in addition to the user enrolled passphrase [15]. Such a method however increases the operation overhead of the user and is cumbersome due to an explicit user cooperation is required besides the standard authentication process. More recently, Chen *et al.* [17] develop a smartphone based liveness detection system by measuring the magnetic field emitted from loudspeakers. It however requires the user to speak the passphrase while moving the smartphone with predefined trajectory around the sound source. Moreover, Zhang *et al.* [55] propose a smarthphone based solution, which measures the time-difference-of-arrival (TDoA) changes of a sequence of phoneme sounds to the two microphones of the phone when a user speaks a passphrase for liveness detection. However, it requires a user to hold the phone at a specific position. While effective, the above-mentioned approaches introduce cumbersome operations as they require either additional steps during authentication or holding or moving the phone in some redefined manners.

In this paper, we introduce VoiceGesture, a smartphone based liveness detection system that achieves the best of both worlds - i.e., it is highly effective in detecting live users, but does not require the users to perform any cumbersome operations. In particular, our system achieves around 1% EER and works when the users hold the phones with their habitual ways of speaking on the phones, i.e., have the phone held either to user's ear or in front of the mouth.

Our system leverages a user's articulatory gestures when speaking a passphrase for liveness detection. Human speech production relies on the precise, highly coordinated movements of multiple articulators (e.g., the lips, jaw and tongue) to produce each phoneme sound. It is known as articulatory gesture, which involves multidimensional movements of multiple articulators [29]. Unlike human, loudspeaker produces sound relying on solely the diaphragm that moves in one dimension (i.e., forward and backward). Thus, by sensing the articulatory motions when speaking a passphrase, a human speaker can be distinguished from a loudspeaker. Moreover, there exist minute differences in articulatory gesture among people due to individual diversity in the human vocal tract (e.g., shape and size) and the habitual way of pronouncing phoneme sounds [36]. Such minute differences could be further leveraged to detect an adversary who tries to mimic the articulatory gesture of a genuine user.

Our system exploits the mobile audio hardware advances to sense and extract user-specific features of articulatory gesture when a user speaks a passphrase to a smartphone. Although the increasingly high definition audio capabilities supported by smartphones are targeted at audiophiles, such advanced capabilities can also be leveraged to sense the motions of the articulators during speech production. In particular, current popular smartphones (e.g., Galaxy S5, S6, and iPhone 5 and 6) are capable to record and playback acoustic sounds at a very high frequency of 20kHz. Such a high frequency has significant implication as it is inaudible to human ear and is easily separable from human voice. Moreover, current audio chips are able to playback and record at 192kHz sampling frequency, which is also supported by smartphone OSs (e.g., Android 6.0 released in 2015) [2, 34]. The high sampling frequency enables us to extract fine-grained frequency domain features to capture both the articulator motions as well as the minute differences of articulator gesture among people.

Our system thus re-uses the smartphone as a Doppler radar, which transmits a high frequency acoustic tone at 20kHz from the built-in speaker and listens to the reflections at the microphone during the process of the voice authentication. The movements of a user's articulators when speaking a passphrase/utterance lead to the Doppler frequency shifts at around 20kHz, which are recorded together with the user's voice sample. Our system then separates the voice sample for conventional voice authentication and extracts user-specific features in the frequency shifts for liveness detection. More specifically, in the user enrollment process, the user-specific frequency shift features are extracted based on the spoken passphrase and then stored in the liveness detection system. During online authentication process, the extracted features of a user input utterance are compared against the ones in the system. If it produces a similarity score higher than a predefined threshold, a live user is declared. To evaluate the performance of our system, we conduct experiments with 21 participants and three different types of phones under various experimental settings. Experimental results show that our system is highly effective in detecting live users and works with users' habitual ways of talking on the phone. The contributions of our work are summarized as follows.

- We show that the mobile audio hardware advances can be leveraged to sense the articulatory gesture of a user when she speaks a passphrase. We also show that it is feasible to capture the minute differences in articulatory gesture among different people when speaking the same phoneme sounds.
- We develop VoiceGesture, a liveness detection system that extracts user-specific features in the doppler shifts that resulted from the articulatory gesture when speaking a passphrase for live user detection. VoiceGesture is practical as it requires neither cumbersome operations nor additional hardware but a speaker and microphone that are commonly available on smartphones.
- Our extensive experimental results show that VoiceGesture achieves over 99% detection accuracy at around 1% EER. Results also show that VoiceGesture is able to work with different phone models and sampling frequencies.

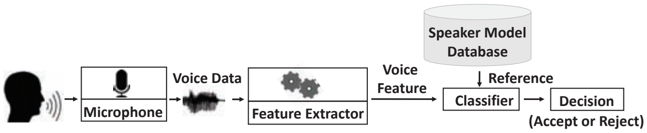


Figure 1: A typical text-dependent authentication system.

The remainder of the paper expands on the above contributions. We begin with system and attack model, and a brief introduction to the articulatory gesture sensing.

2 PRELIMINARIES

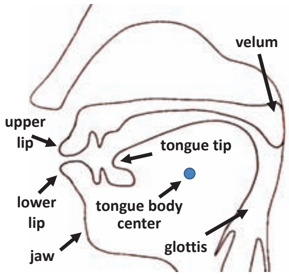
2.1 System and Attack Model

Voice authentication is the process of verifying the claimed identity of a user by extracting the acoustic features that reflect both behavioral and physiological characteristics of a user [48]. In this work, we primarily focus on the text-dependent system, in which a user-chosen or system prompted passphrase is used for user authentication. As a text-dependent system offers high authentication accuracy with shorter utterances, it is generally more suitable for user authentication than text-independent system [49]. A typical text-dependent voice authentication system is shown in Figure 1. Nevertheless, our liveness detection system could be extended to a text-independent system [55].

We consider replay attacks in our work as they are easy to implement by using the wide availability of low-cost and high-quality digital recording and playback devices. To acquire a victim's voice samples, an adversary can either place a recording device surreptitiously in close proximity to the victim or utilize the victim's publicly exposed speeches. An adversary can also extract and concatenate the voice segments to match the victim's passphrase to launch replay attacks. In particular, we consider two types of replay attacks: *playback attack* and *mimicry attack*. In a playback attack, an adversary uses a loudspeaker to replay a pre-recorded passphrase of an intended target user. Given that attackers may know the defending strategy of the liveness detection system, they could conduct more sophisticated mimicry attacks, in which an adversary tries to mimic the articulatory gesture of a genuine user. To perform a mimicry attack, the adversary can use a far-field speaker to replay a pre-recorded passphrase and simultaneously mimic the victim's articulatory gesture corresponding to the replaying passphrase. In mimicry attacks, we also consider that the attacker can observe how a genuine user pronounces the passphrase, for example by taking a video of the genuine user, and then practice before conducting the attack.

2.2 Articulatory Gesture

Human speech production requires precise and highly coordinated movements of multiple articulators [29]. Specifically, articulatory gesture is used to describe the connection between the lexical units with the articulator dynamic when producing speech sounds. For English speech production, the coordination among multiple articulators produces gestures like lip protrusion, lip closure, tongue tip and tongue body constriction, and jaw angle. For example, three articulators including upper lip, lower lip and jaw are involved when a speaker conducts the gesture of lip closure, which could lead to the phoneme sounds of *[p]*, *[b]* and *[m]*.



Phonemes	Articulators	Motion Variables
p, b, m	upper lip lower lip jaw	lip vertical movement lip vertical movement jaw angle
t, d, n, s, z, sh, zh	tongue tip tongue body jaw	radial and angular radial and angular jaw angle
k, g, vowels	tongue body jaw	radial and angular jaw angle
m, n,	velum	velum movement
p, t, k, s, sh	glottal	glottal movement
uw, uh, ow	lip protrusion	horizontal movement

Figure 2: Articulators, phonemes and the corresponding articulatory gestures.

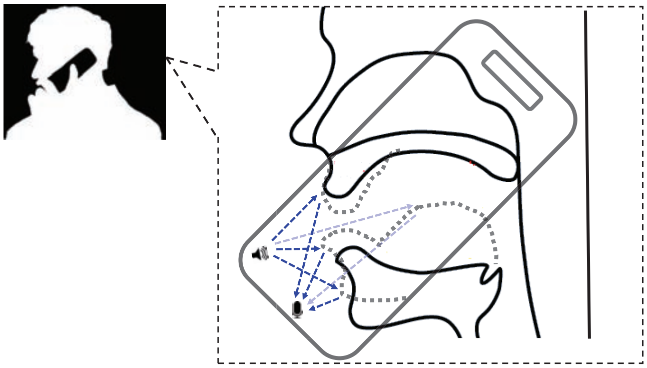


Figure 3: An illustration of sensing the articulatory gesture when a user speaks a passphrase on the phone.

Figure 2 illustrates various articulators and their locations as well as the phonemes and the corresponding articulatory gestures. Each phoneme sound production usually involves multidimensional movements of multiple articulators. For instance, the pronunciation of the phoneme *[p]* requires upper and lower lips horizontal movements and jaw angle change. Moreover, although some phonemes share the same type of articulator gesture, the movement speed and intensity could be different. For example, both *[d]* and *[z]* require the tongue tip constriction, however they differ in terms of the exact tongue tip radial and angular position.

2.3 Sensing the Articulatory Gesture

To sense the articulatory gesture, we leverage the phenomenon of Doppler effect, which is the change in the observed wave frequency as the receiver and the transmitter move relative to each other. A common example of Doppler effect is the change in the pitch of an ambulance's siren as it approaches and departs from a listener. Figure 3 shows one example of sensing the particularity gesture when a user speaks a passphrase by holding the phone to his left ear. The build-in speaker of the phone emits a high frequency tone, which is reflected by multiple articulators of the user. The reflections are then recorded by the built-in microphone of the same phone. In our context, the articulators reflecting the signals from the speaker can be thought of as virtual transmitters that generate the reflected sound waves. As the articulators move towards the microphone, the crests and troughs of the reflected sound waves arrive at the microphone at a faster rate. Conversely, if the articulators move away from the microphone, the crests and troughs arrive at a slower rate. In particular, an articulator moving at a speed of v with an angle of α from the microphone results in a Doppler shift [35] (i.e., frequency change Δf) of:

$$\Delta f \propto \frac{v\cos(\alpha)}{c} f_0, \tag{1}$$

where f_0 is the frequency of the transmitted sound wave and c is the speed of sound in the medium.

We observe from Equation (1) that a higher frequency of the emitted sound (i.e., f_0) results in a larger Doppler shift for the same articulator movements. We thus choose to emit a high frequency sound at 20kHz, which is close to the limit of the built-in speaker/microphone of current popular smartphones. Such a high frequency signal maximizes the Doppler shifts caused by the articulatory gesture and is also inaudible to human ear.

Moreover, the observed Doppler shift depends on the the moving direction of the articulator (i.e., α). An articulator moving away from the microphone results in negative Doppler shift, while an articulator moving towards the microphone leads to a positive Doppler shift. As each phoneme pronunciation involves multidimensional movements of multiple articulators, the resulted Doppler shifts at the microphone are a superposition of sinusoids at different shifts. For instance, the phoneme sound *[o]* requires the lip closure gesture, which involves upper lip and jaw moving towards the microphone and lower lip and tongue moving away from the microphone. We thus could observe a set of Doppler shifts including both positive and negative shifts that can be used to distinguish different articulatory gestures.

In addition, a faster speed (i.e., v) results in a larger Doppler shift. The magnitude of the Doppler shift thus can be further utilized to distinguish different gestures or people that produce the same phoneme sound with various speeds. Furthermore, the reflections from the articulators that closer to the microphone result in stronger energy due to the signal attenuation in the medium. For example, the lip movement usually results in a higher energy in its Doppler shift than that of the tongue tip. The energy distribution of the Doppler shifts thus provide another dimension of information for differentiating articulatory gestures.

2.4 Loudspeaker

Unlike the human, loudspeaker relies on solely the diaphragm that moves in one dimension to produce sound wave [11]. As shown in Figure 4, the diaphragm is moving forward and backward to increase and decrease the air pressure in front of it, thus creating sound waves. The diaphragm is usually driven by the voice coil[1], which converts electrical signals to magnetic energy. By increasing and decreasing the amount of electrical current, the voice coil produces a magnetic field of varying strength, which interacts with the internal permanent magnet. The permanent magnet thus attracts or repels both the voice coil and the attached diaphragm to move froward or backward. The specific movements of the diaphragm are controlled by the frequency and intensity of the input audio signal. For instance, the input sound that possesses a high pitch results in fast movement of the diaphragm, while when a user turning up the volume, the diaphragm pushes harder to produce a higher pressure in the air.

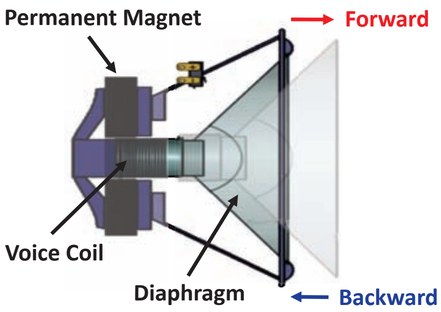


Figure 4: An illustration of a loudspeaker.

A loudspeaker could be distinguished from a live speaker based on the movement of articulators. First, they differ in terms of the movement complexity and the number of the articulators. In addition, the movement of human articulator does not always produce sound, whereas the movement of diaphragm certainly results in sound wave. Figure 5 shows the Doppler shifts sensed by the probe sound at 20kHz for a loudspeaker replay and a live user, respectively. The frequency distribution inside each pair of vertical bars in the figure corresponds to the Dopplor shifts resulted from one phoneme sound. We could observe that the Doppler shifts of the loudspeaker look relatively clean due to much simpler diaphragm movements. The Doppler shifts caused by the complex movements of multiple articulators of a live user spread out over a much larger volume of space. For instance, to pronounce the phoneme *[ai]*, a human speaker first opens his mouth on vertical direction and then gradually changes to horizontal direction. This procedure involves massive movements that result in a diverse of Doppler shifts than that of a loudspeaker.

2.5 Individual Diversity of Articulator Gesture

There exist minute differences in articulatory gesture among people when producing the same phoneme due to the individual diversity in the human vocal tract and the habitual way of pronunciation. For example, research shows that different people adopt different movement trajectories of articulators to produce the same utterance [32]. Also, the physiological features of vocal tract vary among people, such as the size and shape of lips and tongue [41]. Moreover, there is a diverse articulatory strategies for sound production. For instance, some speakers' jaw movement is closely connected with tongue body gesture, while others are not [22].

To assess whether we are able to capture the minute differences in users' articulatory gestures with current smartphone, we use the articulator movement speed among people as one example [31]. Figure 6 shows the statistics of the movement speeds of both upper lip and jaw for five people when producing the same phoneme sound.

[1] Although the diaphragm is driven by stators for electrostatic loudspeaker, it still relies on the movements of diaphragm for sound production.

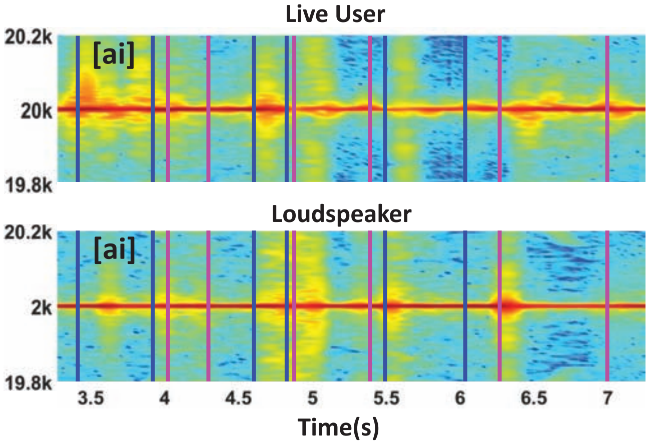


Figure 5: Doppler shifts of a live user and a speaker replay.

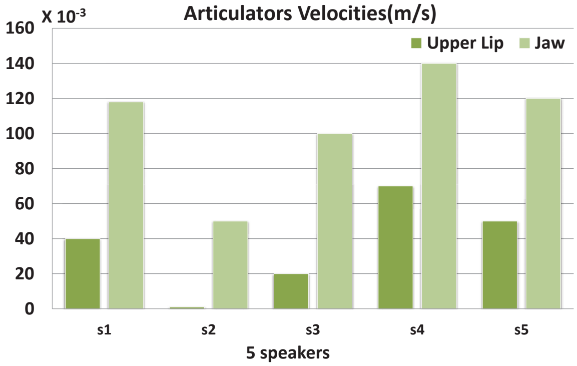
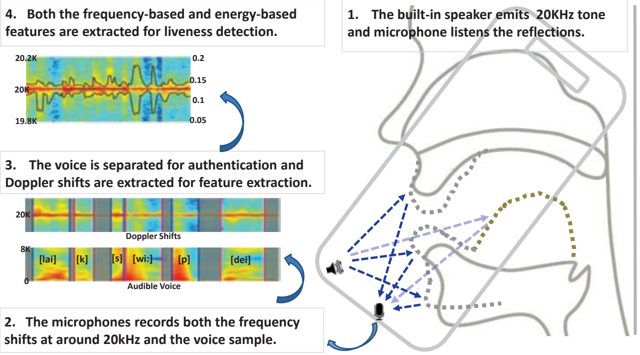


Figure 6: Velocity diversity in upper lip and jaw gestures [27].

We observe a diverse range of movement speed. The averaged difference of the speed is 0.04m/s for upper lip and 0.06m/s for jaw, respectively. Given the duration of producing a phoneme sound is around 250ms [47], we could achieve 1Hz resolution under 192kHz sampling frequency when analyzing the Doppler shifts of each individual phoneme. With the probe sound at 20kHz, 1Hz Doppler shift corresponds to an articulator speed of 0.017m/s, which provides much higher sensitivity than that of the speed difference in both upper lip and jaw movements (i.e., 0.04m/s and 0.06m/s). We thus could be able to differentiate different people even if they are pronouncing the same phoneme sound with 20kHz prob sound wave at 192kHz sampling frequency. Of course, the differences in articulatory gesture are expected to be much smaller under the mimicry attacks, where an adversary mimics the articulatory gesture of a genuine user. Nevertheless, each articulatory gesture involves movements of multiple articulators, which provide more information to detect the attacks. In addition, each passphrase consists of a sequence of phoneme sounds, which dramatically increase the possibility to distinguish between a genuine user and an attacker.


1. The built-in speaker emits 20KHz tone and microphone listens the reflections.

2. The microphones records both the frequency shifts at around 20kHz and the voice sample.

3. The voice is separated for authentication and Doppler shifts are extracted for feature extraction.

4. Both the frequency-based and energy-based features are extracted for liveness detection.

Figure 7: Illustration of the articulatory gesture based liveness detection on smartphone.

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 leverage the mobile audio hardware advances to sense the articulatory gesture of a sequence of phoneme sounds when a user speaks passphrases to a smartphone. As illustrated in Figure 7, the built-in speaker at the bottom of the phone starts to emit an inaudible acoustic tone at 20kHz once the authentication system is triggered. When a user speaks a passphrase, the built-in microphone records user's voice as well as the inaudible acoustic tone and its reflections. Speaking a passphrase involves multidimensional movements of multiple articulators, which result in Doppler frequency shifts in the reflected signals. In particular, the articulators moving toward (away from) the microphone lead to positive (negative) Doppler shifts. While the articulators that closer to the microphone result in stronger energy in the Doppler shifts, the articulators move at faster speeds lead to large Doppler shifts. Once finish recording, the voice sample of the user (which is usually located below 10kHz) is separated for conventional voice authentication, leaving the high frequency band at around 20kHz for extracting features in the Doppler shifts. The system extracts features based on both frequency shift distribution and energy distribution in the observed Doppler shifts. The extracted features are then compared against the ones obtained when user enrolled in the system for live user detection.

A live user is declared if the similarity score exceeds a predefined threshold. Under playback attacks, the extracted features of Doppler shifts are different from the ones obtained from a live user due to the fundamental difference between the human speech production system and the loudspeaker sound production system. Under mimicry attacks, the extracted features can capture the minute differences through a sequence of phoneme sounds due to individual diversity of human vocal tract and the habitual way of pronunciation. Also, it is possible for an attacker to place a recording device (e.g., a smartphone emitting and recording at 20kHz) surreptitiously in close proximity to a legitimate user to record the Doppler shifts when the user speaks a passphrase. As the Doppler shift pattern is tied to the phone placement, the recorded Doppler shifts by the attacker are different from the ones sensed by the legitimate device (e.g., user's smartphone) as long as the attacker has a phone placement different from the one that the legitimate user used for enrollment.

Our system works when the users hold the phones with their nature habits as opposed to the prior smartphone based solutions that require users to hold or move the phone in some predefined manners. Comparing to the commercially used challenge-response based solutions, our system does not require any cumbersome operations besides the conversional authentication process. Once it integrated with voice authentication system, the liveness detection is totally transparent to the users.

Our system however does require the built-in speaker and microphone to playback and record sound wave at a high frequency. The audio hardware on current popular smartphones (e.g., Galaxy S5, S6, and iPhone 5 and 6) has frequency response well above 20kHz. As mobile devices are increasingly supporting high definition audio capabilities, we envision the low-end phones could also reliably record and playback sound wave at high frequencies in the very near future. Moreover, certain data protection methods should be deployed to prevent an attacker from obtaining the plain-text of the extracted features. For example, the feature extraction could be done locally at smartphone and only the encrypted features are transmitted for liveness detection.

3.2 System Flow

Realizing our system requires five major components: *Doppler Shifts Extraction*, *Feature Extraction*, *Wavelet-based Denoising*, *Similarity Comparison*, and *Detection*. As shown in Figure 8, the acoustic signal captured by the phone's microphone first passes through Doppler shifts Extraction process, which extracts the Doppler shifts for each phoneme sound in the spoken utterance. We rely on the audible voice sample of the user for separating each phoneme and the corresponding Doppler shifts. In particular, we apply Hidden Markov Modeling (HMM) based forced alignment to recognize and separate each phoneme in the voice sample. Then, we map the segmentation to the inaudible frequency range at around 20KHz frequency to extract the Doppler shifts of each individual phoneme.

Next, the Feature Extraction component is used to extract both energy-band and frequency-band features from the Doppler shifts. Specifically, our system first partitions the Doppler shifts into several sub-bands based on both energy and frequency levels. It then extracts both the frequency-based and energy-based contours within each sub-band. These extracted frequency-based and energy-based contours capture the movements of multiple articulators in terms of relative positions and relative velocities.

Then we utilize wavelet-based denoising technique to further remove the mixed noises by decomposing each contour into approximation and detailed coefficients. A dynamic threshold is applied to the detailed coefficients to remove the noisy components while retaining sufficient details. After that, we reconstruct the features by combining approximation coefficients and denoised detailed coefficients.

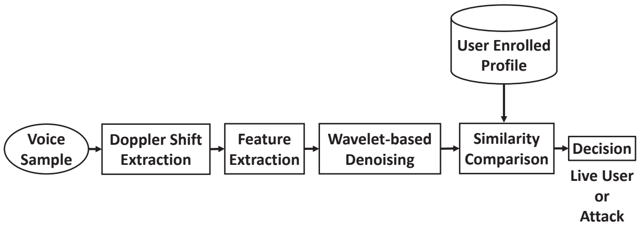


Figure 8: The flow of our liveness detection system.

At last, our system matches the frequency-based and energy-based features with the ones stored in the liveness detection system by using cross correlation coefficient. It yields a similarity score, which is compared against a predefined threshold. If the score is higher than the threshold, a live user is detected, otherwise an attack is declared.

3.3 Doppler Shifts Extraction

Once finish recording, our system first separates the voice sample of the user (i.e., below 10kHz) for conventional voice authentication. Then, we rely on the audible voice sample to separate each individual phoneme and the corresponding Doppler shifts at around 20kHz. Specifically, we convert the recorded signal from the time domain to frequency domain by performing Short-Time Fourier Transform (STFT) with a window size as 250ms. Figure 9 shows one example of the spectrogram of the recorded signal when a user speaks "*like sweep day*". We can find that the audio voice sample is less than 10kHz and the Doppler shifts are usually within 200Hz at around 20kHz. Such a large gap ensures the voice sample will not be affected by the high frequency of 20kHz and its Doppler shifts. Given the spectrogram of the recorded signal, we aim to extract the Doppler shifts for each individual phoneme while removing the pauses due to transaction between phoneme sounds and also the transaction between words (i.e., the shaded bars in the figure).

To perform phoneme segmentation, we utilize the fact that each phoneme consists of numerous distinctive overtone pitches, also known as formants [30]. By inspecting the sound spectrogram, we are able to identify different phonemes by recognizing those formants. In particular, the first two formants with the lowest frequencies are referred to as F1 and F2, which contains the most information can be used to distinguish the vowels. Thus, by analyzing the F1 and F2 in the sound spectrogram, we are able to segment different vowels within given voice sample. Unlike vowels, each consonant is displayed as a mixture of various frequencies randomly. Consequently, only using formants to perform precise segmentation of consonants is very challenging. We thus utilize HMM (Hidden Markrov Mddels) based forced alignment to solve this problem. This method [25] distinguishes different consonants by comparing the input voice sample spectrogram with existing spectrograms and finding the best alignment.

Specifically, we first utilize automatic speech recognition (ASR) to identify each word in the voice sample. We adopt state-of-art

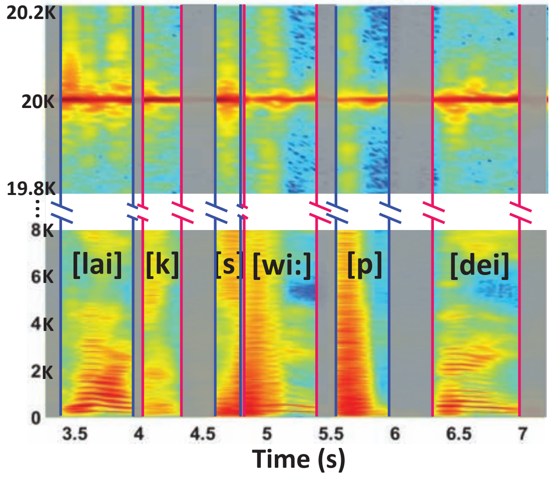


Figure 9: An illustration of Doppler shifts extraction based on phoneme sounds.

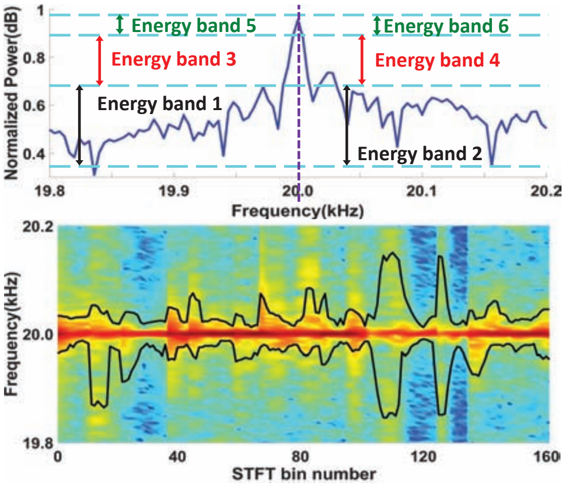


Figure 10: An example of energy sub-band and energy-based frequency contours.

CMUSphinx [38] to perform such task automatically. After identifying existing words in the voice sample, we then perform consonant segmentation and labeling utilizing MAUS [26]. First, based on standard pronunciation model (i.e., SAMPA phonetic alphabet), the identified words will be transformed into expected pronunciation. Next, by combining the canonical pronunciation with millions of possible accents of users, a probabilistic graph will be generated. It contains all possible phoneme combinations and the corresponding probabilities. The system then adopts Hidden Markrov Mddel to perform path search in the graph space and find the combination of phonetic units with the highest probability. The results of the search are the segmented and labeled phonemes for each word. Finally, our system matches the time stamp of each phoneme segmentation to 20KHz frequency range to extract corresponding Doppler shifts.

One example is shown in Figure 9, which illustrates six segmented phoneme sounds (i.e., $[lai]$, $[k]$, $[s]$, $[wi\,:]$, $[p]$, $[dei]$) and the corresponding Doppler shifts at around 20kHz. We observe that the phonemes like $[lai]$ and $[dei]$ display more intensive Doppler shifts than these of the phonemes like $[k]$ and $[p]$. This is because when pronouncing $[lai]$, larger movements from multiple articulators including lips, jaw and tongues are required. In contrast, when pronouncing $[p]$, only small movements from lips and jaw are involved.

3.4 Feature Extraction

After we obtain the Doppler shifts of all the phonemes, we first normalize them as the same length as those stored in the user profile. Such a normalization is used to mitigate the effect of different speech speed of the user when performing voice authentication. Then, we resplice the normalized Doppler shifts of each phoneme together. To eliminate the interferences due to other movements such as nearby moving objects or body movements, we further utilize a Butterworth filter with cut off frequencies of 19.8KHz and 20.2KHz to remove these out of band noises.

Next, we extract two types of features from the Doppler shifts: energy-band frequency feature and frequency-band energy feature. The first type of feature quantifies the relative movement speeds among multiple articulations. By dividing energy level of all the frequency shifts into several different bands, we are able to separate different parts of articulators based on their distances to the microphone. A higher energy of the captured Doppler shifts, a closer movement occurred with respect to the microphone. Before energy band partition, we first normalize the energy level of each segmented phoneme into the same scale (i.e., from 0 to 1). Such a normalization is used to mitigate the energy shift caused by inconsistency of a user when speaking an utterance to the smartphone.

We partition the energy into three levels based on the energy distribution, resulting in 6 sub-bands as each energy level includes both positive and negative Doppler shifts, as shown in the top graph of the Figure 10. Specifically, Sub-band 5 and 6 with power level in between 0.95 to 0.99 represent the strongest Doppler shift signals captured by microphone. Those Doppler shift signals are reflected by the articulators that are closest to the microphone, such as the upper and lower lips. Sub-band 3 and 4 include the power level ranging from 0.7 to 0.9. They represent the Doppler shifts caused by the articulator motions that have further distances comparing with that of the first category, for example, the jaw movement. And sub-band 1 and 2 with energy level smaller than 0.7 but larger than 0.4 consist of motions dominated by articulator components with the farthest distance to the microphone, such as the tongue movement. Given each sub-band, we use the centroid frequency as the feature and combine all the centroid frequencies of each phoneme together, resulting in one frequency contour for each band.

The bottom part of Figure 10 demonstrates two energy-band frequency contours (i.e., band 1 and 2) extracted from the sentence "*Oscar didn't like sweep day*" spoken by a live user. Those two bands represent articulators (e.g., the tongue) with longer distance to the microphone. From Figure 10, we observe at a STFT bin number of 80, the frequency shift is lower than surrounding area, indicating a less movement velocity of an articulator which corresponds to the

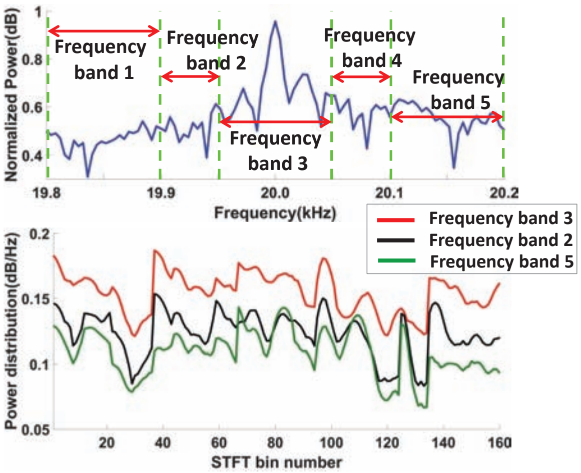


Figure 11: An example of frequency sub-band and frequency-based energy contours.

pronunciation of phoneme *[k]*. Meanwhile, the largest frequency shift in the band can be observed at a STFT bin number of 110, which corresponds to the phonemes *[wi:]*, indicating intensive motion of an articulator captured during the pronunciation.

The second type of feature is the frequency-band energy feature, which quantifies the relative movement positions among multiple articulations across phonemes. As a faster movement velocity results in a larger magnitude of Doppler shift, we thus can compare the energy levels of different articulator movements that have the same movement velocity. In particular, we divide the frequency shifts into 5 major sub-bands in both positive and negative directions, as shown in upper part of the Figure 11. We starts with sub-band 3, which covers frequency shift from -50Hz to 50Hz. The corresponding movements are more likely dominated by articulators with lower movement velocity. Next, sub-band 2 and 4 include frequency shift from 50Hz to 100Hz and -100Hz to -50Hz, respectively. The last two sub-band 1 and 5 include the frequency shift from 100Hz to 200Hz and -200Hz to -100Hz, respectively. They cover the components with the highest movement speed. Similar to the frequency contour, we calculate the average energy level at each frequency sub-band, and then splice the resulted energy level together to form an energy contour.

The lower part of Figure 11 demonstrates three frequency-band energy contours at the band 2, 3 and 5. We observe that the frequency band 3 contour has higher energy level comparing with the other two bands. It is because while speaking an utterance, the lower facial region of a user also move slightly. Although with very slow speed, the large size of the lower facial region leads to much more or stronger signal reflections, resulting in much higher energy than that of each individual articulator. The frequency band 5 contour demonstrates the lowest energy level among three bands and implies the motion is more likely to be caused by the articulator further from the microphone, such as the tongue. And in fact, the tongue is the most flexible part of the articulator and the tongue's motion could be reliably recorded with an open mouth during the pronunciation process.

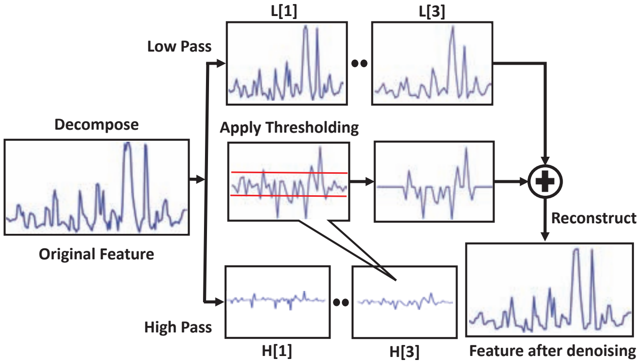


Figure 12: An example of wavelet-based denoising.

3.5 Wavelet-based Denoising

The purpose of wavelet based denoising is to further remove the noisy component mixed in the extracted features. Those components could be caused by hardware imperfection or surrounding environment interferences and noises. Our system thus utilizes wavelet denoising technique that is based on Discrete Wavelet Transform (DWT) to further analyze the signal in both time and frequency domain [39]. It decomposes input signal into two components: approximation coefficients and detailed coefficients. The approximation coefficients depict the trend of input signal, representing large scale features. Meanwhile, the detailed coefficients retain the small scale characteristics, which mixed with both fine details of the signal and noisy components. Our goal is to extract the fine details while removing the mixed noises. To achieve this, we apply a dynamic threshold to the detailed coefficients to remove the noise components.

Figure 12 shows the process of wavelet-based denoising component. Our system first decomposes the each extracted contour into approximation and detailed coefficients by going through low pass and high pass filters. We run this step recursively for 3 levels. After obtaining multiple levels of detailed coefficients, a dynamic threshold is applied to each level of detail coefficients to filter out the mixed noises (i.e., the readings with small values). Then, we combine the original approximation coefficients with the filtered detail coefficients. After that, we use the inverse DWT to reconstruct the denoised contour. The reconstructed features could facilitate accurate liveness detection, especially for those Doppler shifts with similar articulatory gestures.

3.6 Similarity Comparison

To compare the similarity of each extracted contour feature with the corresponding one in the user profile, we use the correlation coefficient technique, which measures the degree of linear relationship between two input sequences [53]. The resulted correlation coefficient ranges from -1 to $+1$, where the value closer to $+1$ indicates a higher level of similarity and a value closer to 0 implies a lack of similarity.

In particular, given a series of n values in each energy-band frequency or frequency-band energy contour A and the corresponding pre-built user profile B, written as A_i and B_i, where $i = 1, 2, \dots, n$.

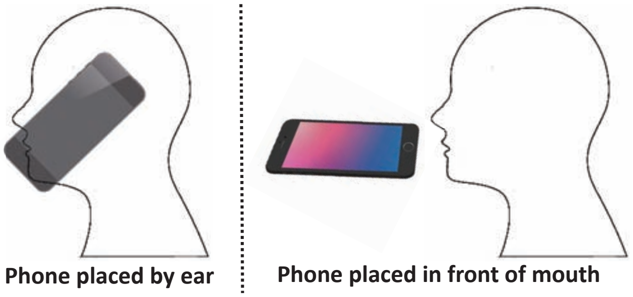


Figure 13: Two different phone placements diagram.

The Pearson correlation coefficient can be calcualted as:

$$r_{AB} = \frac{\sum_{i=1}^{n}(A_i - \bar{A})(B_i - \bar{B})}{(n-1)\delta_A \delta_B}, \quad (2)$$

where $\bar{A}$ and $\bar{B}$ are the sample means of A and B, δ_A and δ_B are the sample standard deviations of A and B.

To detect a live user, we use energy-based frequency contours (i.e., energy-based feature), frequency-band energy contours (i.e., frequency-based feature), and combined feature of these two (combined feature), respectively. Given the correlation coefficients of all contours, we simply compare the averaged coefficient to a predefined threshold for live user detection. Although a more sophisticated classification method, for example a machine learning based classification, could be used, our primary evaluation in this work is the validation of the system methodology.

4 PERFORMANCE EVALUATION

In this section, we present the the experimental performance of our liveness detection system under both *replay* and *mimic* attacks. The project has obtained IRB approval.

4.1 Experiment Methodology

Phones and Placements. We employ three types of phones including Galaxy S5, Galaxy Note3, and Galaxy Note5 for our evaluation. These phones differ in terms of sizes and audio chipsets. Specifically, the lengths of S5, Note3 and Note5 are 14.1cm, 15.1cm and 15.5cm respectively, whereas the chipsets are Wolfson WM1840, 800 MSM8974 and Audience's ADNC ES704, respectively. All the audio chips and the speaker/microphones of these phones can record and playback 20kHz frequency sound. The operating systems of those phones are the Android 6.0 Marshmallow that released in 2015, which supports audio recording and play back at 192kHz sampling frequency. We thus evaluate our system with the sampling frequencies including 48kHz, 96kHz and 192kHz. We present the results for 192kHz sampling frequency in the evaluation unless otherwise stated. Additionally, we consider two types of phone placements as shown in Figure 13 that people usually used to talk on the phone: have the phone held either to user's ear or in front of the mouth.

Data Collection. Our experiments involves 21 participants including 11 males and 10 females. The participants are recruited by emails including both graduate students and undergraduate students. These participants include both native and non-native English speakers with ages from 21 to 35. We explicitly tell the participants that the purpose of the experiments is to perform voice authentication and liveness detection. Each participant chooses his/her own 10 different passphrases. For each passphrase, they repeat three times to enroll in the authentication system and use the averaged features to establish the profile of user. To perform legitimate authentication, each participant tries 10 times for each passphrase, which totals 2100 positive cases. The lengths of those passphrases range from 2 to 10 words with one third are 2 to 4 words, one third are 5 to 7 words, and the rest are 8 to 10 words. In addition, to evaluate the individual diversity among users, we ask 12 out of the 21 participants to pronounce the same passphrase. Our experiments are conducted in classrooms, apartments, and offices with background and ambient noises such as HVAC noises and people chatting.

Attacks. We evaluate our system under two types of replay attack: *playback attacks* and *mimicry attacks*. Both forms of attacks are considered in our evaluation sections unless claimed otherwise. The playback attacks are conducted with loudspeakers including the standalone speakers, the built-in speakers of mobile devices, and the earbuds. In particular, a DELL AC411 loudspeaker, the build-in speaker of Note5 and a pair of Samsung earbud are used to playback the participants' voice samples in front of the smartphone that performing voice authentication. Specifically, each form of these speakers replays voice samples from 10 participants, and the build-in speaker/earbud and the loudspeaker contributes 3 and 4 trials for each of the 10 passphrases respectively, amounting to 1000 replay attacks. All replay attacks are captured by an identical

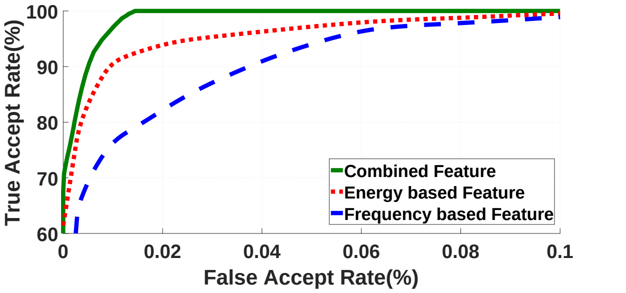


Figure 14: All Attacks: ROC curves under different measurements.

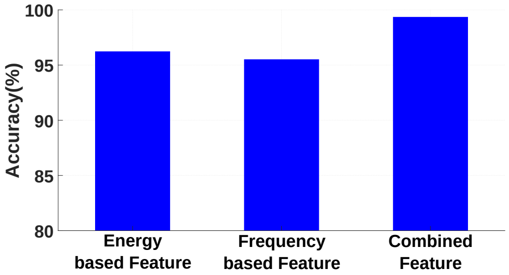
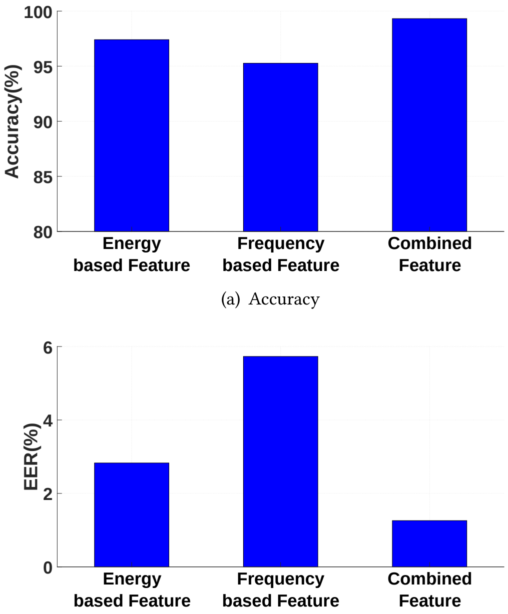


Figure 15: All Attacks: Accuracy under different measurements.



(a) Accuracy

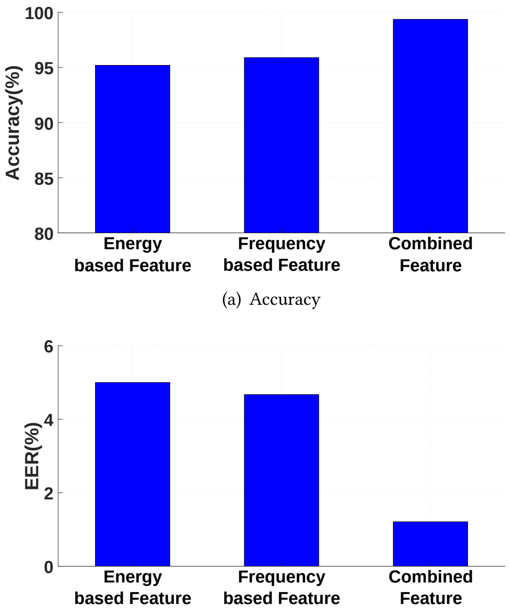
(b) EER

Figure 16: Replay Attacks: Accuracy and EER.

phone with the same holding position that the participants used for authentication.

For mimicry attacks, we first record the articulatory gesture of the participants when they speaking the passphrase by using a digital video recorder. The video recording only covers the lower facial region for privacy concerns. Such a lower facial region including the articulator movement of upper and lower lips, tongue and jaw. Then other participants are invited to watch the video carefully and repeatedly practice the pronunciation by mimicking the articulatory gesture in the video. In particular, they are instructed to mimic the speed of talking, the intensity and range of each articulator movement, the speech tempo and etc. After they claim that they have learned how the person in the video speaks and moves the articulators, they start to conduct the mimicry attacks in front of the smartphone that used for voice authentication. We recruit 4 attackers and each mimics 6 participants. For each victim/participant, 5 trials for each of 5 passphrases are mimicked. There are in total 600 mimicry attack attempts.

Metrics. We evaluate our system with the following metrics. *False Accept Rate (FAR)* is the likelihood that the system incorrectly declares a replay attack as a live user. *True Accept Rate (TAR)* is the probability that the system detects a live user correctly. *Receiver Operating Characteristic (ROC)* curve describes the relationship between the TAR and the FAR when varying the detection threshold. *False Reject Rate (FRR)* is the probability that the system mistakenly classifies a live user as a replay attack. *Equal Error Rate* records the rate when FAR equals to FRR. *Accuracy* presents the possibility that the system accepts live users and rejects attacks. It is the proportion of the true positive and true negative cases in all the evaluated cases.



(a) Accuracy

(b) EER

Figure 17: Mimicry Attacks: Accuracy and EER.

4.2 Overall Performance

We first present the overall performance of our system in detecting live users under both playback and mimicry attacks. Figure 14 depicts the ROC curves of our system under both types of attacks. We observe that with 1% FAR, the detection rate is as high as 98% when using the combined features. Such an observation suggests that our system is highly effective in detecting live users under both replay and mimic attacks. Moreover, we find that the energy-based feature results in better performance than that of the frequency-based feature. For example, with 1% FAR, the frequency-based feature provide the detection rate at around 90%. Furthermore, we observe that the participants who have smaller scale of articulatory movements generate higher false accept rate. Additionally, Figure 15 shows the overall accuracy under both attacks. Similarity, we observe that combined feature has the best performance, with an accuracy at about 99.34%, whereas the energy-based feature alone achieves an accuracy of 96.22%. The time to perform an authentication is about 0.5 seconds on a laptop server. The above results demonstrate the effectiveness of our system in detecting live users. Also, the energy-based feature and frequency-based feature can complement each other to improve the detection performance.

Playback Attack. We next detail the performance under playback attacks. Figure 16 shows the performance in terms of accuracy and EER under replay attacks. We observe that the combined feature results in the best performance. It has an accuracy of 99.3% and an EER of 1.26%. In particular, with only one type of feature, we can achieve an accuracy of 97.41% and an EER of 2.83%. These results show that the two types of feature can complement with each other and the combined feature is very effective in detecting live user under playback attacks.

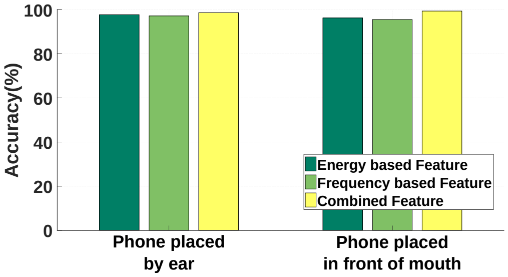


Figure 18: Accuracy of different phone placements.

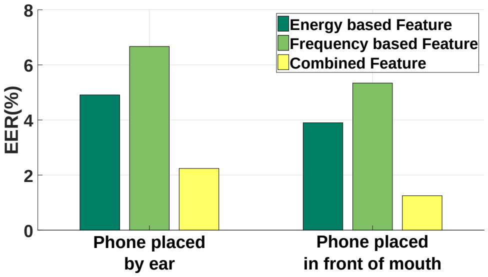


Figure 19: EER of different phone placements.

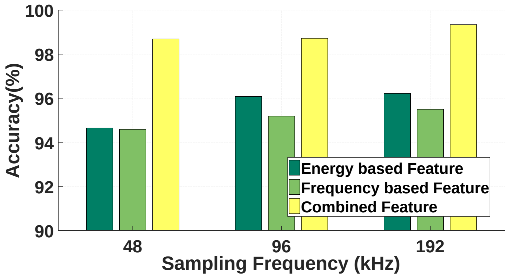


Figure 20: Accuracy under different sampling frequencies.

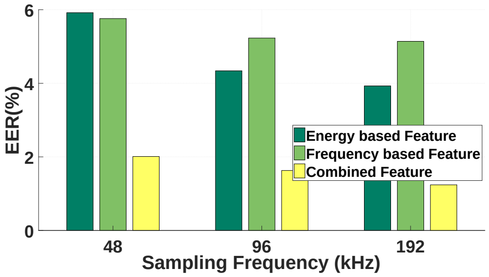


Figure 21: EER under different sampling frequencies.

Mimicry Attack. Next, we study the detailed performance under mimicry attacks. Figure 17 shows both the the accuracy and EER of our system. Again, the combined feature achieves the best accuracy at about 99.3% and an EER of 1.21%. Unlike the playback attack scenario, the frequency-based feature has better performance than that of the energy-based feature. In particular, the frequency-based feature has an accuracy of 95.9% and an EER of 4.67%. The above results suggest that the extracted features from the Doppler shifts of a sequence of phoneme sounds could capture the differences of the articulatory gesture between an attacker and a live user under mimicry attacks. Thus, our system is effective in detecting live users under mimicry attacks.

4.3 Impact of Phone's Placement

Different users may have different habits to talk on the phone in terms of how to hold the phone while speaking. We thus compare the performance under two placements of the phone (i.e., hold the phone to ear and hold the phone in front of the mouth) that people usually feel comfortable to use. Figure 18 presents the performance comparison of the accuracy, whereas Figure 19 shows the comparison of the EER. In high level, the results show that our system is highly effective under both placements. In particular, when placing the phone to the ear, we have the best accuracy as 98.61%, while the best accuracy for placing the phone in front of the mouth is slightly higher. This is due to the fact that placing the phone in front of the mouth can capture the movement of the tongue better as the microphone is directly facing the mouth. Similarly, placing the phone to the ear has slightly worse EER, i.e., at 2.24%, whereas it is about 1.2% for the other placement. Nevertheless, our system works well under both placements and could accommodate different users who have different habits to hold the phone while talking. This property of our system indicates our system doesn't require the user to hold the phone at a specific position or move the phone in a predefined manner as opposed to the prior smartphone based solutions.

4.4 Impact of Sampling Frequency

We next show that how well our system can work with some low-end phones that can only playback and record at 48kHz or 96kHz sampling frequency. Figure 20 depicts the accuracy of our system under 48kHz, 96kHz and 192kHz sampling frequencies. We notice that a higher sampling frequency results in a better performance. This is because a higher sampling frequency could capture more details of the articulatory gestures and has a better frequency resolution. In particular, the combined feature achieves an accuracy of 98.72% for 96kHz sampling frequency, and 98.69% for 48kHz sampling frequency. Moreover, Figure 21 shows the EER under those three sampling frequencies. We find the 96kHz sampling frequency has an EER of 1.63%, whereas it is 2.01% for 48kHz sampling frequency. These results indicate that our system still works very well at a lower sampling frequency. Thus, our system is compatible to these older version smartphones.

4.5 Impact of Different Phones

Our system also supports the users to use different types of phones for enrollment and online authentication. Specifically, we experiment with three different phones including S5, Note3 and Note5. In the experiments, the participants use one of these three phones to enroll in the system but use the other two phones for online voice authentication. The performance of our system is in Figure 22. Results show that our system works well under such scenarios. In particular, the combined feature provides an accuracy of 96.58%,

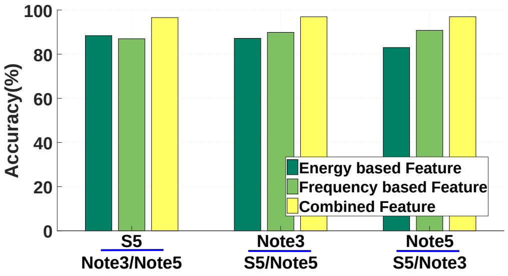


Figure 22: Accuracy of using one phone for enrollment and the other two for online authentication.

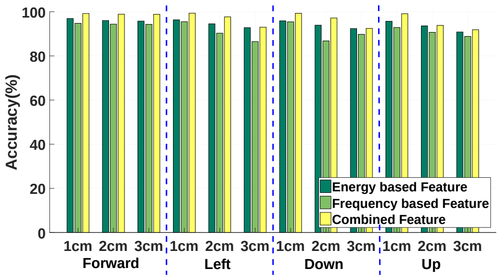


Figure 23: Accuracy under different degree of phone displacement.

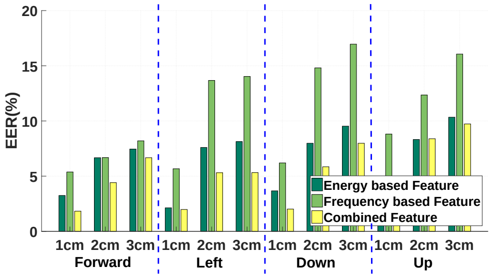


Figure 24: EER under different degree of phone displacement.

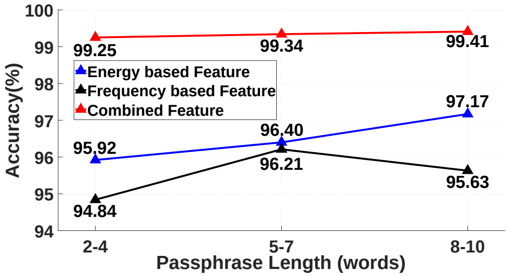


Figure 25: Accuracy under different length of passphrase.

96.93% and 96.98% when using S5, Note3, and Note5 as the enrollment phone, respectively. Results also indicate that the performance is comparably well no mater which phone is used for enrollment. Although the accuracy is slightly worse than that of using the same phone for enrollment and authentication, our system is still able to accommodate different types of phones.

4.6 Robustness to Phone Displacement

In this study, we investigate the performance of our system when experiencing phone displacement between the enrollment phase and online authentication phase. The phone displacement could happen, for example when a user place the phone slightly different from that of the enrolled position or due to hand shakes when a user is talking while walking. Specifically, we exam three degrees of phone displacements, i.e., 1cm, 2cm and 3cm away from the original spot of enrollment in four possible directions, which are Forward from the mouth, Left to the mouth, and Down or Up against the mouth. Figure 23 and Figure 24 depict the accuracy and EER of these scenarios respectively. Generally, a high degree of displacement will decrease the accuracy and increase the EER of our system. Indeed, the average accuracy when displace the phone at 1cm is 99.25% on average, and it is 96.91% and 94.05% on average for 2cm and 3cm displacement, respectively. As for the EER, they are are 1.89%, 5.99% and 7.38% for 1cm, 2cm, and 3cm displacements, respectively.

Furthermore, we notice that the performance is more sensitive to Down and Up displacements. This is due to the fact that the Up and Down displacements is more likely to change the relative positions of multiple articulators to the microphone, thus resulting in the worst performance. Such an observation is consistent with the methodology of our liveness detection system, which relies on the multidimensional movements of multiple articulators for live user detection. However, the displacement in practice is small (e.g., within 1cm) as the size of a user's mouth is small and a user usually intends to put the microphone close to the mouth. Additionally, within the 1-2 seconds time duration of speaking passphrases, we expect small movements of phone to user's mouth, which only have limited effect. Nevertheless, our method provides around 97% accuracy with 2cm phone displacements in all directions. The results in general show that our system is robust to the phone displacement and could tolerate a relative large phone displacement.

4.7 Impact of Passphrase Length

Next, we show how the length of each passphrase affects the performance of our system. Security professionals usually suggest to choose a passphrase with more than 5 words so as to provide a desired security [6]. In the light of this, we classify the passphrases into three categories according to their lengths: 2 to 4 words, 5 to 7 words, and 8 to 10 words. Figure 25 displays the accuracy of our system with different lengths of passphrases. We could observe that when increasing the length of the passphrase, the accuracy slightly improved from 99.25% to 99.41%. This is expected as a longer passphrase results in more articulatory gestures for differentiating a live user from an attacker. Moreover, we observe the improvement is not obvious, since we extract 11-dimensional features from each phoneme, which suggests that 2 to 4 words passphrases containing around 10 to 20 phonemes could provide sufficient information for live user detection.

5 DISCUSSION

Unconventional Loudspeaker. In our work, we have tested conventional loudspeakers including the standalone speakers, the built-in speakers of mobile devices, and the earbuds. Nevertheless, there exists unconventional loudspeakers that do not relies on the diaphragm movement for sound production. For example, Piezoelectric Audio Speakers have a totally different working principle comparing with electro dynamic speakers as there is no voice coils or diaphragms. Each piezoelectric speaker relies on a ceramic disc that interacts when it feels a certain voltage difference. An increase of the signal amplitude Vpp (Voltage peak to peak) results in a larger piezo deformation and leads to a larger sound output. Still, such a mechanism is fundamentally different from that of the human speech production system. It is expected that the proposed liveness detection system works with such unconventional loudspeakers. Another example of unconventional loudspeaker is the Electrostatic Loudspeaker (ESL), which still relies on the diaphragm movements for sound production. It is however, driven by two metal grids or startors instead of voice coil. As our liveness detection system relies on the movements of articulators for live user detection. Playing back with such a loudspeaker can still be detected as a replay attack.

Individual Diversity. In our evaluation, we have tested our system when an attacker mimics the articulatory gesture of a genuine user by observing how the user pronouncing the passphrase. We now show how the performance looks like when an attacker has no prior-knowledge on how the legitimate user speaks. That is, the attacker use his own way of pronouncing the passphrase. This case is equipotent to compare the Doppler shifts of the articulatory gesture between two people who speak the same passphrase with their own habitual ways. Figure 26 shows the accuracy comparison. We observe that we could be able to achieve much higher accuracy at close to 100%. The result demonstrates that it is relative easier to capture the individual diversity than that of a mimicry attack.

Limitations. Our system is evaluated with a limited number of young and educated subjects. It will be useful to evaluate the system with a larger number of participants with a more diverse background to better understand the performance. Moreover, the system is evaluated only for several months. A long-term study could be conducted to consider the case that the individual characteristics is likely to change over lifetime, such as changed mouth cavities or a user grows a beard. Nevertheless, we believe updating user profile periodically could potentially mitigate such a limitation. At last, the system does require the users to hold the phones close to their mouths to reliably capture the articulatory gesture. This limits the applicable scenarios of the system. For instance, the system is less applicable to the cases where the phone is not held in the user's hand but instead is placed somewhere in close proximity.

6 RELATED WORK

Although the number of mobile applications that use voice biometric for authentication is rapidly growing, recent studies show that voice biometrics is vulnerable to spoofing attacks [14, 21, 24, 44, 49]. Such attacks can be further divided into following four categories.

Replay Attack. Numerous work has pointed out existing verification systems can not efficiently defend against replay attacks [18, 20, 46]. A recent study [24] shows the EER of voice authentication systems can increase from 1.76% to 30.71% under replay attacks. Acoustic feature based methods for attack detection have wide applicability, but they all have very limited effectiveness [14, 21, 45, 49, 52]. Current commercial voice authentication system like VoiceVault and Nuance, mostly rely on the challenge-response based methods to detect replay attacks. Such methods however require explicit user cooperation in addition to standard voice authentication process, which could be cumbersome. Recent proposed smartphone based solutions however require the user to hold or move the phone in some predefined manners. In particular, Chen *et al.* [17] propose a smartphone based voice authentication system by measuring the magnetic fields emitted from loudspeakers and thus differentiating them from the live users. It however requires the users to rotate the smartphone around their heads while speaking the given passphrase. VoiceLive [55] measures the time-difference-of-arrival (TDoA) changes to the two microphones of the smartphone to pinpoint the sound origins within a live user's vocal track for liveness detection. While effective, it does require the phones to be held in front of the users' mouths and thus force the majority of the users (who hold the phone by their ears) to change their habitual ways of speaking on phones. In contrast, our system is transparent to users and covers more user cases as it works when holding the phones either to the user's ears or in front of their mouths. Moreover, our system is less susceptible to environmental noises as it senses articulatory gestures by actively emitting high frequency sound waves (which could be easily separated from noises) as oppose to passively listen to the voices that mixed with background noises in VoiceLive.

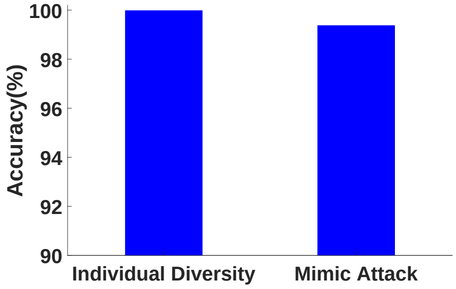


Figure 26: Individual diversity v.s. Mimicry attacks.

Speech Synthesize Attack. This type of attacks assume the attacker is capable of synthesizing the target's voice by using speech synthesize techniques. De Leon *et al.* [18] proposed a relative phase shift feature for GMM-UBM or SVM based speaker verification system. Experiments show this feature can lower the FAR to 2.5%. Also, Wu *et al.* [51] evaluated state-of-the-art systems on defending speech synthesis attacks. Results show an overall average EER of less than 1.5% is achieved. Recent work done by Leow *et al.* [33] utilized concatenation artifacts on the spectrogram to detect US-based synthesized speech attacks, it obtains an EER of 15.2% for 16000Hz utterances. Furthermore, Adobe's recent work VoCo [8] allows the users to edit texts and synthesize corresponding speeches of a given speaker with only 20 minutes of the voice samples of him/her. Though hasn't been evaluated as replay attacks, VoCo might introduce severe potential threats to voice authentication systems as well.

Voice Conversion Attack. The attacker has the ability to imitate victim's voice through voice conversion or manipulation process using existing user voice samples. Conducting voice conversion requires expertise or specialized equipment, however involves no human efforts. Earlier study by Kinnunen *et al.* [23] indicated text-independent speaker verification systems are vulnerable against voice conversion attacks using telephone speech. Besides, Kons *et al.* [28] evaluate several common speaker verification systems, including the I-vector, GMM-NAP, and HMM-NAP based systems, under voice conversion attacks. Results show that they overall gain 4-fold increases in EER, and FAR of HMM-NAP system increase from 1% to 36%. Recently, Wu et al. [50] developed an authentication system with PLDA component that defends against voice conversion attacks with 1.71% FAR, whereas Alegre et al. utilize PLDA and FA technologies and achieve the FAR rate of 1.6% [13]. Further, Sizov *et al.* [42] propose an i-vectors and PLDA based general countermeasure to unknown types of voice conversion attacks, it's claimed that this method could bring EER to as low as 0.54%.

Impersonation Attack. Different from other types of attack, it indicates an adversary would launch an attack without using any professional devices by only relying on impersonating the target's voice. Comparing with other types of attacks, impersonation attacks are less accessible and less risky to authentication systems. Indeed, Wu *et al.* [48] suggest the impersonators may capable of mimicking the F0 pattern and speaking rates of the victims, but it's barely possible for them to fake the spectral characteristics like formants. Therefore impersonation attacks may fool human listens but not authentication systems. Recent work shows by utilizing state-of-the-art speaker model like i-vector models [20] and GMM-UBM [16], the impersonation attack can be effectively mitigated. Furthermore, even with professional mimicry artists or linguists, those common speaker authentication systems maintain considerable effectiveness [19, 37].

7 CONCLUSIONS

In this paper, we developed a voice liveness detection system that requires only a speaker and a microphone that are commonly available on smartphones. Our system, VoiceGesture, is practical as no cumbersome operations are required besides the conversional voice authentication process. Once it is integrated with voice authentication system, the liveness detection is transparent to the users. VoiceGesture performs liveness detection by extracting features in the Doppler shifts that caused by the articulatory gesture when a user speaks a passphrase. Extensive experimental evaluation demonstrates the effectiveness of our system under various conditions, such as with different phone types, placements and sampling rates. Overall, VoiceGesture can achieve over 99% accuracy, with the EER at around 1%.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their insightful feedbacks. This work was partially supported by the National Science Foundation Grands CNS-1464092, CNS-1514238, DGE-1565215, and CNS-1505175.

REFERENCES

[1] 2012. Lenove voice unlock. https://www.techinasia.com/baidu-lenovo-voice-recognition-android-unlock. (2012).

[2] 2015. Android 6.0. https://www.android.com/versions/marshmallow-6-0/. (2015).

[3] 2015. Bank adopt voice mobile application. http://newagebanking.com/finsec/the-new-mobile-banking-password-your-voice/. (2015).

[4] 2015. Saypay Technologies. http://saypaytechnologies.com/. (2015).

[5] 2015. VocalPassword. http://www.nuance.com/ucmprod/groups/enterprise/@web-enus/documents/collateral/nc_015226.pdf. (2015).

[6] 2015. Voicekey Mobile Applications. http://speechpro-usa.com/product/voice_authentication/voicekeytab2. (2015).

[7] 2015. Wechat Voiceprint. http://blog.wechat.com/2015/05/21/voiceprint-the-new-wechat-password/. (2015).

[8] 2016. Adobe VoCo. http://www.bbc.com/news/technology-37899902. (2016).

[9] 2016. VoiceVault. http://www.nuance.com/landing-pages/products/voicebiometrics/vocalpassword.asp. (2016).

[10] 2017. Google Smart Lock. https://get.google.com/smartlock/. (2017).

[11] 2017. Loudspeaker. https://en.wikipedia.org/wiki/Electrostatic_loudspeaker. (2017).

[12] 2017. Voice recognition market share. http://www.marketsandmarkets.com/PressReleases/speech-voice-recognition.asp. (2017).

[13] Federico Alegre, Asmaa Amehraye, and Noah Evans. 2013. A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns. In *IEEE BTAS*.

[14] Federico Alegre, Artur Janicki, and Nicholas Evans. 2014. Re-assessing the threat of replay spoofing attacks against automatic speaker verification. In *Biometrics Special Interest Group (BIOSIG), 2014 International Conference of the.* IEEE, 1–6.

[15] Almog Aley-Raz, Nir Moshe Krause, Michael Itzhak Salmon, and Ran Yehoshua Gazit. 2013. Device, system, and method of liveness detection utilizing voice biometrics. (May 14 2013). US Patent 8,442,824.

[16] Talal B Amin, James S German, and Pina Marziliano. 2013. Detecting voice disguise from speech variability: Analysis of three glottal and vocal tract measures. *The Journal of the Acoustical Society of America* (2013).

[17] Si Chen, Kui Ren, Sixu Piao, Cong Wang, Qian Wang, Jian Weng, Lu Su, and Aziz Mohaisen. 2017. You Can Hear But You Cannot Steal: Defending against Voice Impersonation Attacks on Smartphones. (2017).

[18] Phillip L De Leon, Michael Pucher, Junichi Yamagishi, Inma Hernaez, and Ibon Saratxaga. 2012. Evaluation of speaker verification security and detection of HMM-based synthetic speech. *IEEE Processing of Audio, Speech, and Language* (2012).

[19] Rosa González Hautamäki, Tomi Kinnunen, Ville Hautamäki, and Anne-Maria Laukkanen. 2015. Automatic versus human speaker verification: The case of voice mimicry. *Speech Communication* 72 (2015), 13–31.

[20] Rosa González Hautamäki, Tomi Kinnunen, Ville Hautamäki, Timo Leino, and Anne-Maria Laukkanen. 2013. I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry.. In *INTERSPEECH*.

[21] Artur Janicki, Federico Alegre, and Nicholas Evans. 2016. An assessment of automatic speaker verification vulnerabilities to replay spoofing attacks. *Security and Communication Networks* 9, 15 (2016), 3030–3044.

[22] Keith Johnson, Peter Ladefoged, and Mona Lindau. 1993. Individual differences in vowel production. *The Journal of the Acoustical Society of America* 94, 2 (1993), 701–714.

[23] Tomi Kinnunen et al. 2012. Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech. In *IEEE ICASSP*.

[24] Tomi Kinnunen, Md Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee. 2017. The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection. (2017).

[25] Andreas Kipp, Maria-Barbara Wesenick, and Florian Schiel. 1996. Automatic detection and segmentation of pronunciation variants in German speech corpora. In *IEEE ICSLP*.

[26] Thomas Kisler, Florian Schiel, and Han Sloetjes. 2012. Signal processing via web services: the use case WebMAUS. In *Digital Humanities Conference*.

[27] H Betty Kollia, Vincent L Gracco, and Katherine S Harris. 1995. Articulatory organization of mandibular, labial, and velar movements during speech. *The Journal of the Acoustical Society of America* 98, 3 (1995), 1313–1324.

[28] Zvi Kons and Hagai Aronowitz. 2013. Voice transformation-based spoofing of text-dependent speaker verification systems.. In *INTERSPEECH*.

[29] Bernd J Kröger, Georg Schröder, and Claudia Opgen-Rhein. 1995. A gesture-based dynamic model describing articulatory movement data. *The Journal of the Acoustical Society of America* 98, 4 (1995), 1878–1889.

[30] P Ladefoged. 2014. *A course in phonetics*. Hardcourt Brace Jovanovich Inc. NY.

[31] Adam Lammert, Louis Goldstein, Shrikanth Narayanan, and Khalil Iskarous. 2013. Statistical methods for estimation of direct and differential kinematics of the vocal tract. *Speech communication* 55, 1 (2013), 147–161.

[32] Eleanor Lawson, Jane Stuart-Smith, James M Scobbie, Satsuki Nakai, David Beavan, Fiona Edmonds, Iain Edmonds, Alice Turk, Claire Timmins, J Beck, et al. 2015. Dynamic Dialects: an articulatory web resource for the study of accents.

(2015).

[33] Su Jun Leow, Eng Siong Chng, and Chin-hui Lee. 2016. Zero resource anti-spoofing detection for unit selection based synthetic speech using image spectrogram artifacts. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific*. IEEE, 1–6.

[34] Jian Liu, Yan Wang, Gorkem Kar, Yingying Chen, Jie Yang, and Marco Gruteser. 2015. Snooping keystrokes with mm-level audio ranging on a single phone. In *ACM MobiCom*.

[35] Robert Allen Meyers et al. 1987. *Encyclopedia of physical science and technology.* Academic Press.

[36] Joseph P Olive, Alice Greenwood, and John Coleman. 1993. *Acoustics of American English speech: a dynamic approach.* Springer Science & Business Media.

[37] Saurabh Panjwani and Achintya Prakash. 2014. Crowdsourcing attacks on biometric systems. In *Symposium On Usable Privacy and Security (SOUPS 2014)*. USENIX Association, 257–269.

[38] Mosur K Ravishankar. 1996. *Efficient Algorithms for Speech Recognition.* Technical Report. DTIC Document.

[39] Sylvain Sardy, Paul Tseng, and Andrew Bruce. 2001. Robust wavelet denoising. *IEEE Transactions on Signal Processing* 49, 6 (2001), 1146–1152.

[40] Sayaka Shiota, Fernando Villavicencio, Junichi Yamagishi, Nobutaka Ono, Isao Echizen, and Tomoko Matsui. 2016. Voice liveness detection for speaker verification based on a tandem single/double-channel pop noise detector. *Odyssey 2016* (2016), 259–263.

[41] Adrian P Simpson. 2001. Dynamic consequences of differences in male and female vocal tract dimensions. *The journal of the Acoustical society of America* 109, 5 (2001), 2153–2164.

[42] Aleksandr Sizov, Elie Khoury, Tomi Kinnunen, Zhizheng Wu, and Sébastien Marcel. 2015. Joint Speaker Verification and Antispoofing in the i-Vector Space. *IEEE Transactions on Information Forensics and Security* 10, 4 (2015), 821–832.

[43] Maciej Smiatacz. 2017. Playback Attack Detection: The Search for the Ultimate Set of Antispoof Features. In *International Conference on Computer Recognition Systems*. Springer, 120–129.

[44] Kaavya Sriskandaraja, Vidhyasaharan Sethu, Eliathamby Ambikairajah, and Haizhou Li. 2016. Front-End for Anti-Spoofing Countermeasures in Speaker Verification: Scattering Spectral Decomposition. *IEEE Journal of Selected Topics in Signal Processing* (2016).

[45] Chun Wang, Yuexian Zou, Shihan Liu, Wei Shi, and Weiqiao Zheng. 2016. An Efficient Learning Based Smartphone Playback Attack Detection Using GMM Supervector. In *Multimedia Big Data (BigMM), 2016 IEEE Second International Conference on.* IEEE, 385–389.

[46] Zhi-Feng Wang, Gang Wei, and Qian-Hua He. 2011. Channel pattern noise based playback attack detection algorithm for speaker recognition. In *IEEE ICMLC*.

[47] James R Williams. 1998. Guidelines for the use of multimedia in instruction. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 42. SAGE Publications Sage CA: Los Angeles, CA, 1447–1451.

[48] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li. 2015. Spoofing and countermeasures for speaker verification: a survey. *Speech Communication* 66 (2015), 130–153.

[49] Zhizheng Wu, Sheng Gao, Eng Siong Cling, and Haizhou Li. 2014. A study on replay attack and anti-spoofing for text-dependent speaker verification. In *Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)*. IEEE, 1–5.

[50] Zhizheng Wu, T Kinnunen, ES Chng, and H Li. 2012. A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case. In *IEEE APSIPA ASC*.

[51] Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi, Md Sahidullah, and Aleksandr Sizov. 2015. ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. *Training* 10, 15 (2015), 3750.

[52] Zhizheng Wu and Haizhou Li. 2016. On the study of replay and voice conversion attacks to text-dependent speaker verification. *Multimedia Tools and Applications* 75, 9 (2016), 5311–5327.

[53] Jie Yang, Yingying Chen, and Wade Trappe. 2009. Detecting spoofing attacks in mobile wireless environments. In *SECON*.

[54] David D Zhang. 2012. *Biometric solutions: For authentication in an e-world.* Vol. 697. Springer Science & Business Media.

[55] Linghan Zhang, Sheng Tan, Jie Yang, and Yingying Chen. 2016. VoiceLive: A Phoneme Localization based Liveness Detection for Voice Authentication on Smartphones. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 1080–1091.