This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2022.3199995

A Continuous Articulatory Gesture Based Liveness Detection for Voice Authentication on Smart Devices

Linghan Zhang, Sheng Tan, Yingying Chen, Jie Yang

***Abstract*—Voice biometrics is drawing increasing attention to user authentication on smart devices. However, voice biometrics is vulnerable to replay attacks, where adversaries try to spoof voice authentication systems using pre-recorded voice samples collected from genuine users. To this end, we propose VoiceGesture, a liveness detection solution for voice authentication on smart devices such as smartphones and smart speakers. With audio hardware advances on smart devices, VoiceGesture leverages built-in speaker and microphone pairs on smart devices as Doppler Radar to sense articulatory gestures for liveness detection during voice authentication. The experiments with 21 participants and different smart devices show that VoiceGesture achieves over 99% and around 98% detection accuracy for text-dependent and text-independent liveness detection, respectively. Moreover, VoiceGesture is robust to different device placements, low audio sampling frequency, and supports medium range liveness detection on smart speakers in various use scenarios, including smart homes and smart vehicles.**

***Index Terms*—Voice authentication, continuous liveness detection, IoT, articulatory gesture.**

I. Introduction

VOICE biometrics has been widely used as an alternative to passwords on smartphones for user authentication. For example, Google developed "Trusted Voice" for Android device access [1], whereas Saypay supports voice biometrics secured online transactions on mobile devices [2]. Recently, voice biometrics is drawing increasing attention as it enables secure and convenient interactions between users and smart devices for various application and services. For instance, voice biometrics has been used for access control on smart speakers, locks, vacuums, and thermostats in the smart home hub [3]. The Enterprise Bank deploys voice biometrics to secure complex personalized banking operations, such as checking balances, making transfers, and sending bank statements [4]. Moreover, automakers such as BMW, Audi, etc. support in-car voice assistants [5] and provide voice-controlled skills like navigation, whereas self-driving vehicles makers like Tesla [6] have gone a further step by allowing voice biometrics-based car controls.

However, a growing body of research has demonstrated the vulnerability of voice authentication systems to replay attacks [7]. An attacker could easily collect the genuine users' voice samples from social media or record those via common digital devices for replay attacks. Indeed, such low-cost and low-effort attacks are highly effective in spoofing voice authentication systems. For instance, simply replaying a pre-recorded voice command of a user could unlock her/his mobile devices that have the voice-unlock feature (e.g., Android devices) [1]. 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% [7].

To defend against replay attacks, traditional liveness detection methods distinguish acoustic characteristics of live users' voices and replayed ones. Such methods cannot detect replayed high-quality voice recordings [7]. Commercial voice authentication service providers like Nuance [8] mainly rely on challenge-response for liveness detection. These methods require extra user cooperation besides standard authentication procedures. Moreover, many researchers target detecting human vital signs during the human speak for liveness detection. For example, Zhang *et al.* measure the time-difference-of-arrival(TDoA) of different phonemes to the two microphones of the smartphone [9]. Wang *et al.* [10] examine human breathing patterns, while Shang *et al.* [11] detect human body vibrations. These solutions require users to hold the device close to or even in touch with their body when they speak. Besides, some liveness detection systems necessitate extra devices. For instance, WiVo [12] and REVOLT [13] use Wi-Fi signals to detect mouth motions and breathing. VAuth [14] and VocalPrint [15] leverage wearable sensors and mmWave to measure body vibrations, respectively. Furthermore, most solutions only support text-dependent liveness detection [9]–[11], [16] and hence could not protect real-time conversations between users and smart devices.

In this paper, we introduce VoiceGesture, a continuous 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. Specifically, VoiceGesture measures a user's articulatory gestures 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

L. Zhang is with the Department of Cyber Security Engineering, George Mason University, Fairfax, VA 20030.

S. Tan is with the Department of Computer Science, Trinity University, San Antonio, TX 78212.

Y. Chen is with the Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ, 08854.

J. Yang (Corresponding Author) is with the Department of Computer Science, Florida State University, Tallahassee, FL 32306.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2022.3199995

involves multidimensional movements of multiple articulators. Unlike the human, the loudspeaker produces sound relying solely on the diaphragm that moves in one dimension (i.e., forward and backward). Thus, by sensing the articulatory motions, VoiceGesture enables to distinguish a human speaker from a loudspeaker. Moreover, there exist minute differences in articulatory gestures among people due to individual diversity in the human vocal tract (e.g., shape and size) and the habitual way of pronouncing phoneme sounds [17]. VoiceGesture leverages such minute differences to detect mimicry attacks.

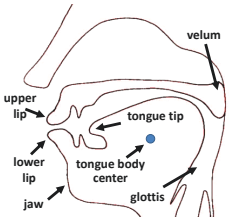
Specifically, VoiceGesture re-uses a pair of built-in speaker and microphone on a smart device as a Doppler radar to sense user-specific features of his/her articulatory gestures while the user speaks. In particular, the built-in speaker transmits an ultrasound probe signal at 20kHz to the moving articulators, which reflect the probe signal and cause Doppler frequency shifts at around 20kHz. Meanwhile, the built-in microphone keeps listening and recording the audio reflections and the user's voice samples. Next, VoiceGesture separates the voice samples for conventional voice authentication and extracts articulatory gesture features from the frequency shifts for liveness detection. In particular, we extract both the frequency and the energy features to reveal the velocity and the location information of the articulatory gestures, respectively. Eventually, we evaluate VoiceGesture with 21 participants, three different types of phones, and a smart speaker under various experimental settings. The results show that VoiceGesture is highly effective in detecting live users and works with users' habitual ways of interacting with the device. The contributions of our work are summarized as follows.

- We leverage smart devices' audio hardware to sense the unique articulatory gestures of a user when he/she speaks a passphrase. We also show that it is feasible to capture the minute differences in different people's articulatory gestures when they speak the same phoneme sounds.
- We develop VoiceGesture, a practical liveness detection system secures both text-dependent and text-independent voice authentication. VoiceGesture requires neither cumbersome operations nor additional hardware other than a pair of speaker and microphone that are commonly available on the latest smart devices.
- The experimental results show that VoiceGesture achieves over 99% detection accuracy at around 1% EER. Moreover, VoiceGesture works on different smart devices with distinct sampling frequencies. Especially, VoiceGesture supports medium-range liveness detection in various use scenarios like smart vehicles and smart homes.

II. PRELIMINARIES

A. System and Attack Model

Voice authentication is the process of verifying a user's claimed identity by extracting the acoustic features that reflect both behavioral and physiological characteristics of a user. This work primarily focuses on the text-dependent systems, where a user-chosen or system prompted passphrase is used for authentication. We also extend this solution to text-independent liveness detection. Text-dependent offers high accuracy and thus has been used for user authentication [18]. However, text-dependent authentication only provides one-time identity verification with enrolled passphrases. In comparison, text-independent authentication can enroll and verify the user's identity transparently and continuously regardless of the speech contents. This feature is especially critical to popular conversational voice assistant systems like Google Assistant.



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

Fig. 1: Articulators, phonemes and the corresponding articulatory gestures.

For the attack models, we consider replay attacks in our work as they are easy to implement and highly effective in spoofing the voice authentication systems. 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 VoiceGesture, they could conduct more sophisticated mimicry attacks, where an adversary tries to mimic the articulatory gestures of a genuine user while replay the pre-reocrded passphrase with a far-field speaker.

B. Articulatory Gesture

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

Figure 1 illustrates various articulators and their locations, common English 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 articulatory gestures, their movement speeds and intensities could be different. For example, both *[d]* and *[z]* require tongue tip constrictions. However, they differ in terms of the exact tongue tip radial and angular position.

C. Sensing the Articulatory Gesture

We leverage the Doppler effect to sense the articulatory gestures. When a user speaks a passphrase to a phone held

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2022.3199995

by his/her ear, the built-in speaker of the phone emits a high-frequency probe signal to the users' articulators. The moving articulators reflect the probe signal and the reflections are recorded by the built-in microphone. In our context, the articulators reflecting the probe signal can be considered 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 slower. In particular, an articulator moving at a speed of v with an angle of α from the microphone results in a Doppler shift (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.

Equation (1) shows 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 ultrasound probe signals at 20kHz, which are close to the limit of the built-in speaker/microphone of smart devices. Such a high-frequency signal maximizes the Doppler shifts caused by the articulatory gesture and is also inaudible to the human ear. Moreover, the Doppler shifts are vectors decided by the moving directions of the articulators (i.e., α). An articulator moving away from the microphone results in negative Doppler shifts, while an articulator moving towards the microphone leads to positive Doppler shifts. In addition, faster speeds (i.e., v) result in larger Doppler shifts. Furthermore, the reflections from the articulators that are closer to the microphone result in stronger energy due to signal attenuation. Therefore, we could analyze the Doppler shifts' magnitudes and energy distribution to distinguish different articulatory gestures or people who speak the same phonemes with different articulatory gestures.

D. Loudspeaker

Unlike the human, loudspeakers solely rely on diaphragms' one dimensional movements to produce sound waves. Specifically, the loudspeaker diaphragms move forward and backward to increase and decrease the air pressure in front of it, thus creating sound waves. Such movements are controlled by the frequency and intensity of the input audio signals. For instance, high-pitch input sounds result in fast movement of the diaphragm, while with higher volume, the diaphragm pushes harder to produce a higher pressure in the air.

Therefore, a loudspeaker could be distinguished from a live speaker with sound production mechanisms. First, they differ in terms of movement complexity and the number of articulators. Second, the movements of human articulators do not always produce sound, whereas the movements of diaphragms certainly result in sound waves. Figure 2 shows the Doppler shifts sensed by the probe signal at 20kHz when a loudspeaker replays and a live user speaks the same phrase, respectively. The frequency distribution inside each pair of vertical bars in the figure corresponds to the Doppler shifts

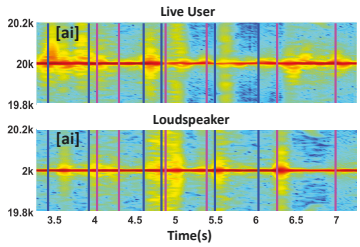


Fig. 2: Doppler shifts of a live user and a speaker replay.

resulted from the same phoneme sound. We could observe that the Doppler shifts of the loudspeaker look relatively clean due to the simple diaphragm movements, whereas the Doppler shifts caused by a live user's complex articulatory gestures spread out over a much larger volume of space.

E. Individual Diversity of Articulator Gesture

There exist minute differences in articulatory gestures 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 people adopt different movement trajectories of articulators to produce the same utterances [19]. Moreover, physiological features of vocal tracts, such as the sizes and shapes of the lips and tongues, vary among people [20]. Furthermore, there are diverse articulatory strategies for sound production. For instance, some speakers' jaw movements are closely connected with tongue body gestures, while others are not [21].

According to research on five individuals [22] articulatory gestures, the averaged speed differences of their upper lips and jaws are 0.04m/s and 0.06m/s, respectively. Given the duration of a phoneme sound is around 250ms and most smart devices support 192kHz sampling frequency, we could achieve 1Hz frequency resolution when calculating each phoneme's frequency shifts. Moreover, with the 20kHz probe signal, 1Hz Doppler shift corresponds to the articulator speed of 0.017m/s, which provides a 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 enable to differentiate different people even if they are pronouncing the same phoneme sound. The differences in articulatory gestures are expected to be much smaller under the mimicry attacks, where an adversary mimics the articulatory gestures of a genuine user. Nevertheless, each articulatory gesture involves movements of multiple articulators. In addition, a passphrase consists of a sequence of phoneme sounds, which dramatically increase the possibility to distinguish between a genuine user and an attacker.

III. SYSTEM DESIGN

A. Approach Overview

The key idea underlying our liveness detection system is to leverage the smart devices' audio hardware to sense the articulatory gesture of a sequence of phoneme sounds when a user speaks to the devices. Taking the smartphone use case in Figure 3 as an example, the built-in speaker at the

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2022.3199995

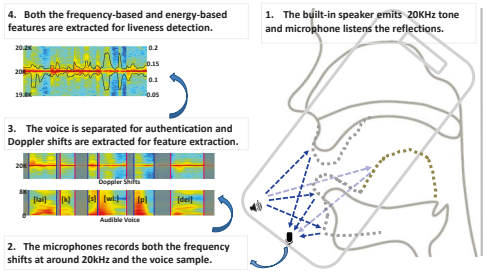


Fig. 3: Illustration of the articulatory gesture based liveness detection on smartphone.

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. Then we extract features based on both the frequency shift and energy distribution in the observed Doppler shifts around 20kHz, and compared those against the passphrase-based or phoneme-based user profile obtained during user enrollment for text-dependent or text-independent liveness detection. A live user is declared if the similarity score exceeds a predefined threshold. Under playback attacks, the extracted Doppler shift features are different from the user profile due to the fundamental difference between the human speech and the loudspeaker sound production systems. Under mimicry attacks, the extracted features show minute differences from the user profile given individual diversity of human vocal tract and the habitual ways of pronunciation.

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. Moreover, our system support medium-range (up to 1m) text-independent liveness detection for smart device use cases in IoT environments like smart vehicles and smart homes. Comparing with the commercially used challenge-response based solutions, our system does not require any cumbersome operations besides the standard authentication process. Once it integrated with a voice authentication system, the liveness detection is totally transparent to the users.

B. 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 4, the acoustic signal captured by the built-in microphone first passes through the Doppler Shifts Extraction process, which extracts the Doppler shifts for each phoneme sound in the spoken utterance. We rely on the user's audible voice samples to separate each phoneme and the corresponding Doppler shifts. Then, we map the segmentation to the inaudible frequency range at around 20kHz frequency to extract the Doppler shifts of each phoneme. Next, the Feature Extraction component extracts both energy-band and frequency-band features from the Doppler shifts. Then we utilize wavelet-based denoising technique to further remove the mixed noises.

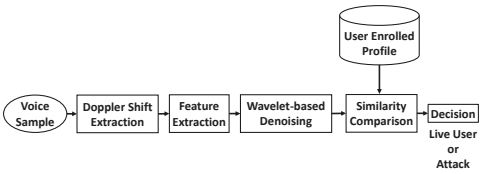


Fig. 4: The flow of our liveness detection system.

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.

C. Doppler Shifts Extraction

Once finish recording, our system first separates the user's voice samples (i.e., below 10kHz) for conventional voice authentication. Then, we rely on the audible voice samples to separate each phoneme and the corresponding Doppler shifts at around 20 kHz. Specifically, we convert the recorded signal from the time domain to the frequency domain by performing Short-Time Fourier Transform (STFT) with a window size of 250ms. Since voice samples are below 10 kHz whereas Doppler shifts are around 20 kHz, we could separate these two signals with low-pass and band-pass filters. Next, given the spectrogram of the recorded signal, we aim to extract the Doppler shifts for each individual phoneme. For text-dependent liveness detection, we further remove the pauses due to transaction between phoneme sounds and also the transactions between words. Then we perform phoneme segmentation [9] to obtain segmented and labeled phonemes for each word. Finally, our system matches the time stamp of each phoneme segmentation to 20kHz frequency range to extract the corresponding Doppler shifts.

D. Feature Extraction

After we obtain the Doppler shifts of all the phonemes, we first normalize them as the same length as those in the user profile. Such normalization is used to mitigate the user's different speech speeds when performing voice authentication. Then, we re-splice the normalized Doppler shifts of each phoneme together for text-dependent liveness detection. 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 of the whole passphrase (text-dependent systems) or single phonemes (text-independent systems): energy-band frequency features and frequency-band energy features.

The first type of feature quantifies the relative movement speeds among multiple articulations. By dividing the energy level of all the frequency shifts into several different bands, we can separate different parts of articulators based on their distances to the microphone. With higher energy of the captured

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2022.3199995

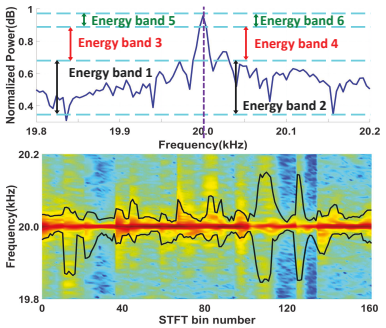


Fig. 5: An example of energy sub-band and energy-based frequency contours.

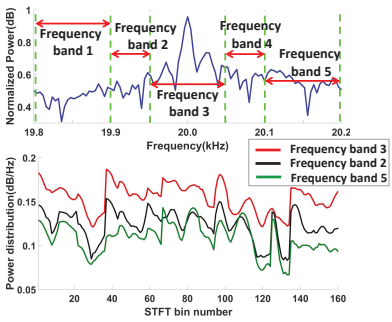


Fig. 6: An example of frequency sub-band and frequency-based energy contours.

Doppler shifts, a closer movement occurred with respect to the microphone. Before energy band partition, we first normalize each segmented phoneme's energy level into the same scale (i.e., from 0 to 1). Such normalization can mitigate the energy shifts caused by the inconsistency of a user speaking the same utterance to the device. We partition the energy into three levels based on the energy distribution, resulting in 6 sub-bands. Each energy level includes both positive and negative Doppler shifts, as shown in the top graph of Figure 5. Sub-band 5 and 6 with power levels between 0.95 to 0.99 capture the strongest Doppler shifts that resulted from closest articulators like lips. Sub-band 3 and 4 include the power level from 0.7 to 0.9. They catch the Doppler shifts caused by closer articulators like jaws. Whereas sub-band 1 and 2 with lowest energy level between 0.4 to 0.7 contain Doppler shifts of the farthest articulators like the tongue. Given each sub-band, we use the centroid frequency as the feature and combine all the centroid frequencies of each phoneme, resulting in one frequency contour for each band. The bottom part of Figure 5 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 a longer distance to the microphone.

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 results in a larger magnitude of Doppler shift, we can compare the energy levels of different articulators moving at similar velocities. In particular, we divide the frequency shifts into 5 major sub-bands considering three levels of velocities in both positive and negative directions, as shown in the upper part of Figure 6. Sub-band 3 covers frequency shifts from -50Hz to 50Hz, sub-band 2 and 4 include frequency shifts from 50Hz to 100Hz and -100Hz to -50Hz, and sub-band 1 and 5 correspond to 100Hz to 200Hz and -200Hz to -100Hz. Similar to the frequency contour, we calculate the average energy level at each frequency sub-band, and then splice the resulted energy levels together to form an energy contour. The lower part of Figure 6 demonstrates three frequency-band energy contours at the band 2, 3, and 5. We observe that the frequency band 3 contour has the highest energy level. It is because while speaking an utterance, the lower facial region of a user moves slowly. Nevertheless, the large size of the lower facial region leads to strong signal reflections. The frequency band 5 contour demonstrates the lowest energy level caused by articulators farthest from the microphone, such as the tongue.

E. Wavelet-based Denoising

We adopt Discrete Wavelet Transform (DWT) to remove the noisy components mixed in the extracted features. Those components could be caused by hardware imperfections or surrounding environment interferences and noises. DWT decomposes the input signal into two components: approximation coefficients and detailed coefficients, which depict the signal overall trend and fine details, respectively. VoiceGesture first decomposes each extracted contour into approximation and detailed coefficients by going through low pass and high pass filters. VoiceGesture runs 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, VoiceGesture combines the original approximation coefficients with the filtered detail coefficients. After that, VoiceGesture use the inverse DWT to reconstruct the denoised contour.

F. Template Building

For text-dependent liveness detection, we build and compare passphrase-based contour features. While for text-independent liveness detection, we establish a set of phoneme-based templates for each individual user. We notice that the contour features of some phonemes are more stable than those of others. For example, a short-sound phoneme tends to provide more consistent contour features than a long-sound phoneme. The reason is that when pronouncing a short-sound phoneme like a consonant or a monophthong, the individual's articulator movements are monotonous. However, when a human pronouncing a long-sound phoneme, especially a diphthong, the articulator movements could be changeable. This observation may vary depending on the speakers' accent.

Therefore, we improve the system stability by assigning different weights to phoneme templates according to their consistency. This could enhance the impact of the phonemes with stable features, whereas lower the influence of phonemes with unstable features. Specifically, we align the beginning of

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2022.3199995

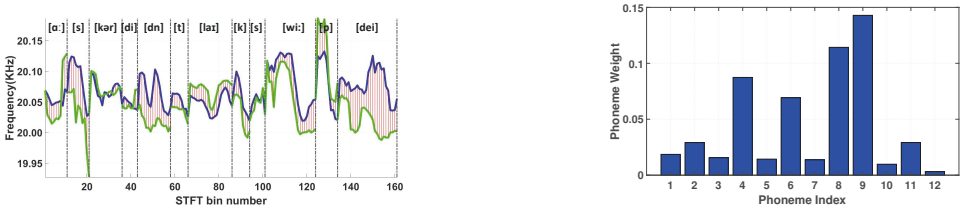


Fig. 7: Weighted Phoneme-based Templates Building.

each phoneme contour and then adopt the following equation to calculate the weight of a phoneme-based template:

$$w = \frac{\sum_{i=1}^{n} L_i}{\sum_{i,j=1}^{n}(A_i - A_j))} \tag{2}$$

where A is any one of the same type of contour features, and L is the length of the phoneme. The denominator of this equation calculates the aggregate areas between the contours of any two trials of the same phoneme. We remove the impact of the phoneme length by introducing the numerator that computes the total lengths of n trials of pronunciations. Figure 7b displays the weights of 12 phonemes in Figure 7a. As we could note that consonants or monophthongs like $[s]$ $[di]$ yield more consistent contour than diphthongs like $[dei]$.

G. Similarity Comparison

Text-dependent Similarity Comparison. For the text-dependent liveness detection system, 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. 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 $i = 1, 2, ..., n$. The Pearson correlation coefficient can be calcualted as:

$$r_{AB} = \frac{\sum_{i=1}^{n}(A_i - \bar{A})(B_i - \bar{B})}{(n-1)\delta_A \delta_B}, \tag{3}$$

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.

Text-independent Similarity Comparison. For the text-independent liveness detection, while the user speaks, VoiceGesture keeps listening while searching for the phoneme-based templates for each phoneme in the speech. After collecting all the templates for the current sentence, we compare the similarity between the contour features of each phoneme in this sentence and the phoneme-based weighted templates with the following equation:

$$\rho_{xy} = \frac{\sum_{i,j=1}^{n}[w_i(A_i - \bar{A}_i)(B_i - \bar{B}_i)]]}{\sqrt{\sum_{i=1}^{n}(w_i(A_i - \bar{A}))^2 \sum_{i=1}^{n}(w_i(B_i - \bar{B}))^2}} \tag{4}$$

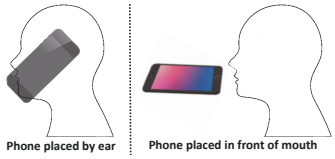


Fig. 8: Two different phone placements diagram.

where A_i and B_i are the corresponding contours of the i^{th} phoneme, and w_i is the weight of this phoneme. To be noticed, both A_i and B_i are sequences, and $\bar{A}_i$ and $\bar{B}_i$ are the averages of A_i and B_i respectively. Before comparison, we normalize A_i and B_i to become the same length, then we apply w_i to each point in A_i and B_i.

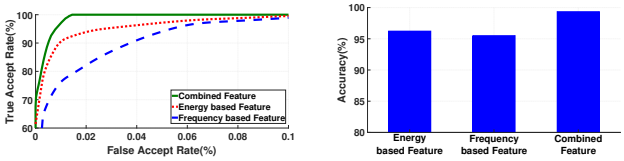
These procedures enable text-independent liveness detection by comparing the current speech with the weighted, phoneme-based templates, rather than the passphrase-based templates. Therefore, VoiceGesture could protect the whole communication session continuously in the IoT environments. To detect a live 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.

IV. PERFORMANCE EVALUATION

A. Experiment Methodology

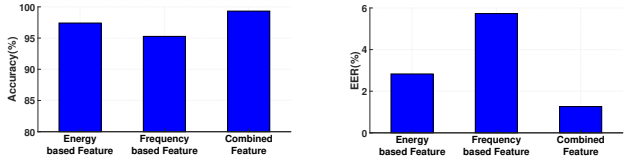
Phones and Placements. We employ three types of phones including Galaxy S5, Galaxy Note3, and Galaxy Note5 for our evaluation. These phones differ in terms of sizes and audio chipsets. Specifically, the lengths of S5, Note3 and Note5 are 14.1cm, 15.1cm and 15.5cm respectively, whereas the chipsets are 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 8 that peo-

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2022.3199995



(a) ROC (b) Accuracy

Fig. 9: Overall Performance.



(a) Accuracy (b) EER

Fig. 10: Performance under Replay Attacks

ple usually used to talk on the phone, i.e., holding smartphones closely by the ear or in front of the mouth.

Smart Speaker Setup We adopt MiniDSP UMA-8 [23] for experiments on smart speakers as it deploys circular arranged microphone arrays like many state-of-the-art popular smart speakers (e.g., Google Home and Amazon Echo). Moreover, it grants users access to the raw recording data. In particular, MiniDSP UMA-8 has a microphone array composed of 7 MEMS microphones. One of the microphones is located at the center and the other 6 microphones are uniformly configured around a circular board with a radius of 0.43 m. To extend the effective range of our liveness detection solution, we utilize the Delay-and-Sum beamforming technique on the microphone array. The Delay-and-Sum beamforming is based on the fact that the signals received by these microphones are similar, nevertheless, they have different delays and phases. Therefore, it calculates the time difference of arrival of the signals received by the 6 microphones and that recorded by the microphone in the center, and then shifts the signals by corresponding phases and sums them up. During experiments, the smart speaker locates on a typical desk around 1 meter tall. The participants stand in front of the smart speaker and face the speaker when they speak.

Data Collection. Our experiments involve 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 text-dependent liveness detection, they repeat each passphrase three times to enroll in the authentication system and use the averaged features to establish the user profiles. Whereas for text-independent liveness detection, the participants read the "Rainbow Passage" [24] that contains all English phonemes to create the phoneme-based frequency shift templates. Each participant tries 10 times for each passphrase to perform legitimate authentication, 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.

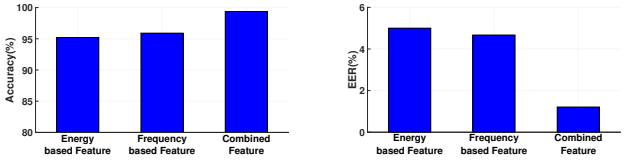
Attacks. We evaluate our system under two types of replay attack: *playback attacks* and *mimicry attacks*. Both forms of attacks are considered in our evaluation sections unless claimed otherwise. The playback attacks are conducted with loudspeakers including the standalone speakers, the built-in speakers 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 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.

B. Overall Performance

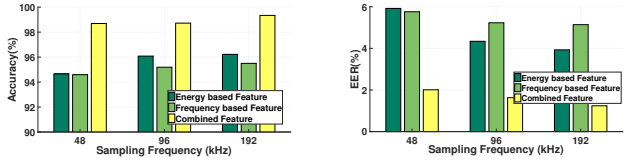
We first present the overall performance of our system in detecting live users under both playback and mimicry attacks. Figure 9a depicts the ROC curves of our system under both types of attacks. We observe that with 1% FAR, the detection rate is as high as 98% when using the combined features. Such 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

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2022.3199995



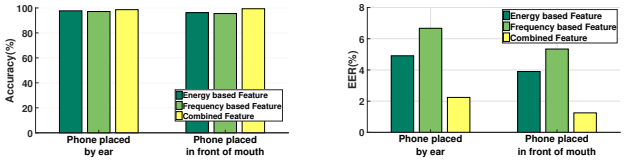
(a) Accuracy (b) EER

Fig. 11: Performance under Mimicry Attacks



(a) Accuracy (b) EER

Fig. 13: Performance under different sampling frequencies



(a) Accuracy (b) EER

Fig. 12: Performance under different phone placements

movements generate higher false accept rate. Additionally, Figure 9b 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 10 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.

Mimicry Attack. Next, we study the detailed performance under mimicry attacks. Figure 11 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.

C. 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 12a presents the performance comparison of the accuracy, whereas Figure 12b 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.

D. 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 13a 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 13b 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.

E. 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

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2022.3199995

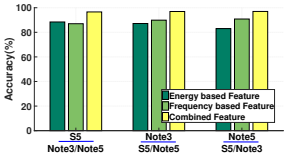


Fig. 14: Accuracy of using a phone for enrollment and another for authentication.

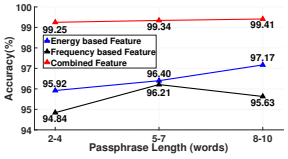


Fig. 15: Accuracy under different length of passphrase.

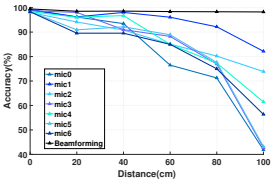


Fig. 17: Accuracy with smart speaker at different distances.

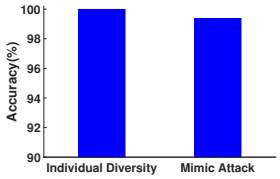


Fig. 18: Individual diversity v.s. Mimicry attacks.

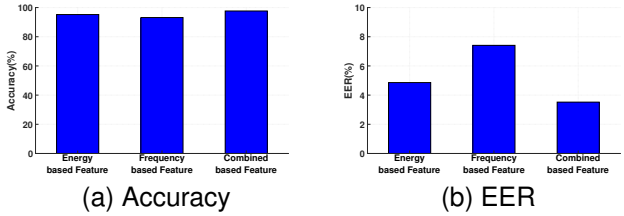


Fig. 16: Text-independent Liveness Detection Performance

two phones for online voice authentication. The performance of our system is in Figure 14. Results show that our system works well under such scenarios. In particular, the combined feature provides an accuracy of 96.58%, 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.

F. 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. 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 15 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.

G. Overall Performance of Text-independent Liveness Detection

To secure conversational voice assistants and continuous voice authentication, we build a text-independent liveness detection solution based on phoneme templates. Figure 16 present the overall accuracy and EER of the text-independent liveness system. We could observe that, similar to the performance of the text-dependent liveness detection, the combined features result in the best performance with 97.65% accuracy and 3.52% EER. Whereas the energy-based features yield better performance than the frequency-based features, which, nevertheless, both achieve accuracy around 95%. Comparing with the text-dependent liveness detection, whose best accuracy is around 99%, our system realize text-independent liveness detection at a small cost of around 2% accuracy loss. Such slight degradation is normal as it could protect the whole conversation from replay attacks.

H. Overall Performance of Medium-range Liveness Detection on Smart Speakers

With the microphone array on smart speakers and the beamforming technique, we realize medium-range liveness detection to support various use cases in IoT environments. For example, in smartphone use cases, the users may hold their device within 30cm to themselves. In a smart vehicle, a driver could activate the auto-drive function by talking with the in-car Audio System around 50cm away. Whereas in smart home environments, a user may sit on the couch while interacting with a smart speaker on the end table locating at a distance of 100cm. To start the experiment, we connect the UMA-8 to a laptop in an office environment. We ask one participant to stand in front of the microphone array with distances of 0 cm, 20 cm, 40 cm, 60 cm, 80 cm, and 100 cm. The participant is asked to speak 3 sentences at each distance, and to repeat each sentence for 10 times. During the experiments, we use a smartphone (Note5) to emit the 20 kHz probe signal, and record the reflected probe signal, as well as the voices with the microphone array at a sampling frequency of 48 kHz. Figure 17 shows the liveness detection accuracies of our solution on the UMA-8 smart speaker with and without beamforming. We could observe that when the user is 20 cm away, the accuracy resulted from single microphone recordings drops to as low as 89.55%. Moreover, with increasing distance, the accuracy keeps dropping and jumps to 76.49% at 60 cm, 71.26% at 80 cm, and 41.87% at 100 cm. In comparison, the beamformed audio recording results in stable high accuracy above 98% for all test distances.

V. 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, a piezoelectric speaker relies on a ceramic disc that interacts when it feels a certain voltage difference. A higher signal amplitude VPP (Voltage peak to peak) results in a larger piezo deformation and leads to a larger volume. Such a mechanism is fundamentally different from human speech production system. 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 stators instead of voice coil. As our liveness detection system relies on the movements of articulators for live user detection. Playing back with such loudspeakers 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 18 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. Nevertheless, we believe updating user profile periodically could potentially mitigate such a limitation. At last, the system expects users to hold smartphones with same placements and distances for enrollment and authentication. VoiceGesture leverages smart devices as Doppler radars and compares Doppler shifts caused by articulatory gestures for liveness detection. These Doppler shifts are decided by articulatory gestures' speeds and directions, and relative locations between the smart devices and users' articulators. This limits the system's applicability since it does not support cross-factor usage, where the user holds the device in one placement/distance for enrollment and changes to another for authentication.

Other Attacks. The evaluation focuses on defending replay attacks that are effective and require no expertise from attackers. Recently, researchers propose several new modalities of attacks on voice authentication systems in recent years. For example, Zhang *et al.* modulate malicious voice commands with inaudible carriers such that only microphones can detect the commands due to non-linearity [25]. Moreover, a list of adversarial attacks creates imperceptible malicious voice commands by adding crafted perturbations to environmental sounds like music and noises [26]–[30]. Although those attacks are inaudible or imperceptible to human ears, attackers must play the malicious voice commands via speakers. Therefore, VoiceGesture potentially enables to defend both attack modalities since it examines behavioral and physiological features that only exist in live human users when they speak voice commands.

VI. RELATED WORK

Biometrics authentication relies on the physiological and/or behavioral characteristics of a user. The human physiological characteristics, such as fingerprint [31], facial features [32], ear canal shapes [33], and toothprint [34] have been widely used for mobile authentication. Moreover, human behavioral characteristics, such as voice [35], gait [36], signature [37], vital signs [38], finger gestures [39], human activities [40], and user locations [41], have been intensively investigated as well. Among these biometrics, voice biometrics has been gaining increasing popularity on smart devices or for mobile applications.

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 [7], [42]. Acoustic feature based methods for attack detection have wide applicability. For example, Zhou and Liu detect replay attacks by analyzing acoustic parameters that reflect audio recordings' overall quality [43]. Nevertheless, such methods pose limited effectiveness if the attackers record and replay the voices with advanced audio hardware in quiet acoustic environments [44]. 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. Many smartphone based solutions require the user to hold or move the phone in some predefined manners. For example, Zhang *et al.* measure the phonemes' time-difference-of-arrival (TDoA) dynamics to the two microphones of the phone when a live user speaks for liveness detection [9]. Though effective, it necessitate the users to hold the smartphone in front of their mouths. Moreover, many liveness detection solutions are only effective when the user and the device are in close proximity. For instance, Wang *et al.* [10] require the users to hold the phone close to their mouth and detect the human speakers' featured breathing sounds for liveness detection. Whereas Shang *et al.* [11] and Wang *et al.* [45] ask the users to hold the smartphone against their throats and chests to examine the throat vibrations and heart beats respectively for liveness detection. Furthermore, some researchers resort to extra devices for liveness detection. One example is the WiVo system that quantifies the Wi-Fi signals' CSI (Channel State Information) changes caused by mouth motions [12]. Similarly, REVOLT uses Wi-Fi to measure the breathing rate [13], whereas VocalPrint employs mmWave to sense vocal vibrations [15]. Besides, VAuth collects body vibrations with wearable sensors for liveness detection [14]. 2MA asks the users to prove their presence with similar recordings collected from multiple devices carried or close to the user [46]. Yan *et al.* require two spaced microphones to measure the field prints. The effectiveness of this method could be largely affected by the size of the device (distance

between the microphones) [47]. Recently, Meng *et al.* leverage the circular layout of microphone array on smart speakers to differentiate sound propagation features of live users and speakers for liveness detection [48]. This method does not work for smart devices without microphone arrays. Moreover, Zhao *et al.* leverage various sensing channels, including inaudible sound, RFID, and Wi-Fi, and introduce random noises to these sensing signals to enhance mouth movement sensing-based liveness detection [49]. This system could be disturbed by environmental noises.

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.

VII. Conclusions

In this paper, we proposed a voice liveness detection system requiring only a speaker and a microphone that are commonly available on smart devices. Our system, VoiceGesture, is practical as no cumbersome operations are required besides the conventional 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 Doppler shift features caused by the articulatory gesture when a user speaks. Extensive experimental evaluation demonstrates the effectiveness of our system under various conditions, such as with different device types, placements and sampling rates. Moreover, VoiceGesture supports medium range liveness detection in various smart speaker use cases in smart homes and smart vehicles. Overall, VoiceGesture can achieve over 99% accuracy, with the EER at around 1% for text-dependent liveness detection, whereas around 98% accuracy and 3% EER for text-independent liveness detection.

Acknowledgments

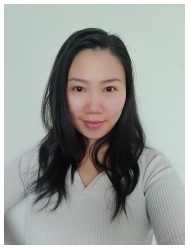
We thank the anonymous reviewers for their insightful feedback. The major part of this work was done when Linghan Zhang was a Ph.D. student at Florida State University. This work was partially supported by the NSF Grants: CNS 1910519, CNS 2131143, DGE 1565215, CNS 2120396, and CNS 1801630.

References

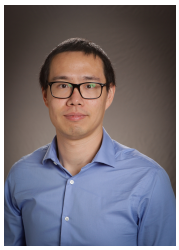
[1] "Google smart lock," https://get.google.com/smartlock/, 2017.
[2] "Saypay technologies," http://saypaytechnologies.com/, 2015.
[3] "20 smart voice recognition and voice activated products for the home," https://www.homestratosphere.com/smart-voice-recognition-for-home/.
[4] "Enacomm launches amazon alexa voice banking skill for enterprise bank," https://www.paymentsjournal.com/enacomm-launches-amazon-alexa-voice-banking-skill-for-enterprise-bank/.
[5] "Nuance's dragon drive powers automotive assistant in toyota concept-i user experience concept vehicle," https://www.globenewswire.com/news-release/2018/01/09/1286293/0/en/Nuance-s-Dragon-Drive-Powers-Automotive-Assistant-in-Toyota-Concept-i-User-Experience-Concept-Vehicle.html.
[6] "Support voice commands," https://www.tesla.com/support/voice-commands.
[7] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," 2017.
[8] "Vocalpassword," http://www.nuance.com/ucmprod/groups/enterprise/@web-enus/documents/collateral/nc_015226.pdf, 2015.
[9] L. Zhang, S. Tan, J. Yang, and Y. Chen, "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*, 2016, pp. 1080–1091.
[10] Q. Wang, X. Lin, M. Zhou, Y. Chen, C. Wang, Q. Li, and X. Luo, "Voicepop: A pop noise based anti-spoofing system for voice authentication on smartphones," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2019, pp. 2062–2070.
[11] J. Shang, S. Chen, and J. Wu, "Defending against voice spoofing: A robust software-based liveness detection system," in *2018 IEEE 15th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*. IEEE, 2018, pp. 28–36.
[12] Y. Meng, Z. Wang, W. Zhang, P. Wu, H. Zhu, X. Liang, and Y. Liu, "Wivo: Enhancing the security of voice control system via wireless signal in iot environment," in *Proceedings of the Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, 2018, pp. 81–90.
[13] S. Pradhan, W. Sun, G. Baig, and L. Qiu, "Combating replay attacks against voice assistants," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, pp. 1–26, 2019.
[14] H. Feng, K. Fawaz, and K. G. Shin, "Continuous authentication for voice assistants," in *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*, 2017, pp. 343–355.
[15] H. Li, C. Xu, A. S. Rathore, Z. Li, H. Zhang, C. Song, K. Wang, L. Su, F. Lin, K. Ren *et al.*, "Vocalprint: exploring a resilient and secure voice authentication via mmwave biometric interrogation," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, 2020, pp. 312–325.
[16] L. Zhang, S. Tan, and J. Yang, "Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 57–71.
[17] J. P. Olive, A. Greenwood, and J. Coleman, *Acoustics of American English speech: a dynamic approach*. Springer Science & Business Media, 1993.
[18] Z. Wu, S. Gao, E. S. Cling, and H. Li, "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, 2014, pp. 1–5.
[19] E. Lawson, J. Stuart-Smith, J. M. Scobbie, S. Nakai, D. Beavan, F. Edmonds, I. Edmonds, A. Turk, C. Timmins, J. Beck *et al.*, "Dynamic dialects: an articulatory web resource for the study of accents," 2015.
[20] A. P. Simpson, "Dynamic consequences of differences in male and female vocal tract dimensions," *The journal of the Acoustical society of America*, vol. 109, no. 5, pp. 2153–2164, 2001.
[21] K. Johnson, P. Ladefoged, and M. Lindau, "Individual differences in vowel production," *The Journal of the Acoustical Society of America*, vol. 94, no. 2, pp. 701–714, 1993.
[22] H. B. Kollia, V. L. Gracco, and K. S. Harris, "Articulatory organization of mandibular, labial, and velar movements during speech," *The Journal of the Acoustical Society of America*, vol. 98, no. 3, pp. 1313–1324, 1995.
[23] "Uma-8 usb mic array - v2.0," https://www.minidsp.com/products/usb-audio-interface/uma-8-microphone-array.
[24] "The rainbow passage," https://www.dialectsarchive.com/the-rainbow-passage.
[25] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, "Dolphinattack: Inaudible voice commands," in *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, 2017, pp. 103–117.
[26] N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou, "Hidden voice commands," in *25th USENIX security symposium (USENIX security 16)*, 2016, pp. 513–530.
[27] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *2018 IEEE security and privacy workshops (SPW)*. IEEE, 2018, pp. 1–7.
[28] X. Yuan, Y. Chen, Y. Zhao, Y. Long, X. Liu, K. Chen, S. Zhang, H. Huang, X. Wang, and C. A. Gunter, "{CommanderSong}: A systematic approach for practical adversarial voice recognition," in *27th USENIX security symposium (USENIX security 18)*, 2018, pp. 49–64.

[29] L. Schönherr, K. Kohls, S. Zeiler, T. Holz, and D. Kolossa, "Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding," *arXiv preprint arXiv:1808.05665*, 2018.
[30] Y. Chen, X. Yuan, J. Zhang, Y. Zhao, S. Zhang, K. Chen, and X. Wang, "{Devil's} whisper: A general approach for physical adversarial attacks against commercial black-box speech recognition devices," in *29th USENIX Security Symposium (USENIX Security 20)*, 2020, pp. 2667–2684.
[31] W. Yang, S. Wang, J. Hu, G. Zheng, and C. Valli, "Security and accuracy of fingerprint-based biometrics: A review," *Symmetry*, vol. 11, no. 2, p. 141, 2019.
[32] E. Bagherian and R. W. O. Rahmat, "Facial feature extraction for face recognition: a review," in *2008 International Symposium on Information Technology*, vol. 2. IEEE, 2008, pp. 1–9.
[33] Z. Wang, S. Tan, L. Zhang, Y. Ren, Z. Wang, and J. Yang, "Eardynamic: An ear canal deformation based continuous user authentication using in-ear wearables," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 1, pp. 1–27, 2021.
[34] Z. Wang, Y. Ren, Y. Chen, and J. Yang, "Earable authentication via acoustic toothprint," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 2390–2392.
[35] L. Zhang, S. Tan, Y. Chen, and J. Yang, "A phoneme localization based liveness detection for text-independent speaker verification," *IEEE Transactions on Mobile Computing*, 2022.
[36] Y. Ren, Y. Chen, M. C. Chuah, and J. Yang, "User verification leveraging gait recognition for smartphone enabled mobile healthcare systems," *IEEE Transactions on Mobile Computing*, vol. 14, no. 9, pp. 1961–1974, 2014.
[37] Y. Ren, C. Wang, Y. Chen, M. C. Chuah, and J. Yang, "Signature verification using critical segments for securing mobile transactions," *IEEE Transactions on Mobile Computing*, vol. 19, no. 3, pp. 724–739, 2019.
[38] J. Liu, Y. Wang, Y. Chen, J. Yang, X. Chen, and J. Cheng, "Tracking vital signs during sleep leveraging off-the-shelf wifi," in *Proceedings of the 16th ACM international symposium on mobile ad hoc networking and computing*, 2015, pp. 267–276.
[39] S. Tan and J. Yang, "Wifinger: Leveraging commodity wifi for fine-grained finger gesture recognition," in *Proceedings of the 17th ACM international symposium on mobile ad hoc networking and computing*, 2016, pp. 201–210.
[40] S. Tan, L. Zhang, Z. Wang, and J. Yang, "Multitrack: Multi-user tracking and activity recognition using commodity wifi," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–12.
[41] H. Liu, Y. Wang, J. Liu, J. Yang, Y. Chen, and H. V. Poor, "Authenticating users through fine-grained channel information," *IEEE Transactions on Mobile Computing*, vol. 17, no. 2, pp. 251–264, 2017.
[42] L. Zhang, S. Tan, Z. Wang, Y. Ren, Z. Wang, and J. Yang, "Viblive: a continuous liveness detection for secure voice user interface in iot environment," in *Annual Computer Security Applications Conference*, 2020, pp. 884–896.
[43] Y. Zhou and Y. Liu, "Replay attack anaysis based on acoustic parameters of overall voice quality," in *2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP)*. IEEE, 2021, pp. 599–604.
[44] C. Wang, Y. Zou, S. Liu, W. Shi, and W. Zheng, "An efficient learning based smartphone playback attack detection using gmm supervector," in *Multimedia Big Data (BigMM), 2016 IEEE Second International Conference on*. IEEE, 2016, pp. 385–389.
[45] L. Wang, K. Huang, K. Sun, W. Wang, C. Tian, L. Xie, and Q. Gu, "Unlock with your heart: Heartbeat-based authentication on commercial mobile phones," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 2, no. 3, pp. 1–22, 2018.
[46] L. Blue, H. Abdullah, L. Vargas, and P. Traynor, "2ma: Verifying voice commands via two microphone authentication," in *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, 2018, pp. 89–100.
[47] C. Yan, Y. Long, X. Ji, and W. Xu, "The catcher in the field: A fieldprint based spoofing detection for text-independent speaker verification," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019, pp. 1215–1229.
[48] Y. Meng, J. Li, M. Pillari, A. Deopujari, L. Brennan, H. Shamsie, H. Zhu, and Y. Tian, "Your microphone array retains your identity: A robust voice liveness detection system for smart speaker," in *USENIX Security*, 2022.
[49] C. Zhao, Z. Li, H. Ding, W. Xi, G. Wang, and J. Zhao, "Anti-spoofing voice commands: A generic wireless assisted design," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 3, pp. 1–22, 2021.

BIOGRAPHY SECTION

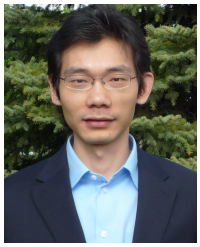


Linghan Zhang received her Ph.D. degree in the Department of Computer Science at Florida State University. She is currently an Assistant Professor in the Department of Cyber Security Engineering at George Mason University. Her research interests include biometrics, user authentication, cyber security and privacy, mobile computing, and human-computer interaction.



Sheng Tan received his Ph.D. degree in Computer Science from Florida State University in 2019. He is currently an Assistant Professor at the Department of Computer Science, Trinity University. His research interests include mobile computing, cyber security and human computer interaction.

Yingying (Jennifer) Chen is the Department Chair and Professor of Electrical and Computer Engineering at Rutgers University. She is a Peter Cherasia Endowed Faculty Scholar at Rutgers. She is an IEEE Fellow and NAI Fellow. She is also named as an ACM Distinguished Scientist. Her research interests include mobile sensing and computing, cyber security and privacy, Internet of Things, and smart healthcare. She is a pioneer in the area of RF/WiFi sensing, location systems, and mobile security. She had extensive industry experience at Nokia previously. She has published 3 books, 4 book chapters and 240+ journal articles and refereed conference papers. She is the recipient of seven Best Paper Awards in top ACM and IEEE conferences. Her research has been reported by numerous media outlets. She has been serving/served on the editorial boards of IEEE/ACM Transactions on Networking (IEEE/ACM ToN), IEEE Transactions on Mobile Computing (IEEE TMC), IEEE Transactions on Wireless Communications (IEEE TWireless), and ACM Transactions on Privacy and Security.



Jie Yang is an Associate Professor in the Department of Computer Science at Florida State University. He directs research in mobile computing and cybersecurity, is a pioneer in the area of WiFi sensing and mobile authentication. His work has been regularly featured in the media, including MIT Technology Review, New Scientist, Yahoo News, NPR, the New York Times, and The Wall Street Journal. He has published one book and three book chapters and 100+ research papers in prestigious journals and conferences. His recognitions include the FSU Developing Scholar Award, FSU CS Faculty Research Award, Google Faculty Research Award, the Stevens Francis T. Boesch Award, as well as Best Paper Awards at IEEE CNS 2014, IEEE CNS 2013, and ACM MobiCom 2011.