VibLive: A Continuous Liveness Detection for Secure Voice User Interface in IoT Environment

Linghan Zhang
Florida State University
Tallahassee, FL, USA
lzhang@cs.fsu.edu

Sheng Tan
Trinity University
San Antonio, TX, USA
stan@trinity.edu

Zi Wang
Florida State University
Tallahassee, FL, USA
ziwang@cs.fsu.edu

Yili Ren
Florida State University
Tallahassee, FL, USA
ren@cs.fsu.edu

Zhi Wang
Florida State University
Tallahassee, FL, USA
zwang@cs.fsu.edu

Jie Yang
Florida State University
Tallahassee, FL, USA
jie.yang@cs.fsu.edu

ABSTRACT

The voice user interface (VUI) has been progressively used to authenticate users to numerous devices and applications. Such massive adoption of VUIs in IoT environments like individual homes and businesses arises extensive privacy and security concerns. Latest VUIs adopting traditional voice authentication methods are vulnerable to spoofing attacks, where a malicious party spoofs the VUIs with pre-recorded or synthesized voice commands of the genuine user. In this paper, we design VibLive, a continuous liveness detection system for secure VUIs in IoT environments. The underlying principle of VibLive is to catch the dissimilarities between bone-conducted vibrations and air-conducted voices when human speaks for liveness detection. VibLive is a text-independent system that verifies live users and detects spoofing attacks without requiring users to enroll specific passphrases. Moreover, VibLive is practical and transparent as it requires neither additional operations nor extra hardwares, other than a loudspeaker and a microphone that are commonly equipped on VUIs. Our evaluation with 25 participants under different IoT intended experiment settings shows that VibLive is highly effective with over 97% detection accuracy. Results also show that VibLive is robust to various use scenarios.

CCS CONCEPTS

• **Security and privacy** → **Biometrics**; **Access control**; **Intrusion detection systems**.

KEYWORDS

voice user interface, liveness detection, bone-conducted vibrations

ACM Reference Format:
Linghan Zhang, Sheng Tan, Zi Wang, Yili Ren, Zhi Wang, and Jie Yang. 2020. VibLive: A Continuous Liveness Detection for Secure Voice User Interface in IoT Environment. In *Annual Computer Security Applications Conference (ACSAC 2020), December 7–11, 2020, Austin, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3427228.3427281

ACSAC 2020, December 7–11, 2020, Austin, USA

ACM ISBN 978-1-4503-8858-0/20/12…$15.00
https://doi.org/10.1145/3427228.3427281

1 INTRODUCTION

The voice user interface (VUI) is becoming a hitting trend as an enabler of natural voice interactions between the user and the IoT environments with considerable efficiency and convenience [13, 14, 17, 21]. For example, VUI has been engaged in automotive to support Smart Car environments. Automobile manufacturers, such as Mercedes-Benz, BMW and Toyota have recently introduced VUI to provide the users intelligent voice in-car control [24]. VUI has also been widely used in e-commerce. For instance, Google Express offers users voice ordering service from more than 70 retail partners including Costco, Target and Home Depot [23]. Not to mention the pervasive VUI based Smart Home applications including the Smart Locks, Smart Thermostats, Smart plugs and switches and so on [16]. Statistically, due to the booming market of VUIs, the estimated sales of VUIs will reach USD 31.82 billion by 2025 [29]. These technologies facilitate our lives enormously, whereas at the possible cost of invading individual privacy and security.

Indeed, recent studies show that VUI based IoT devices are vulnerable to spoofing attacks, where a malicious party attacks the VUIs with a pre-recorded, concatenated, synthesized, or even modulated inaudible ultrasound voice command from the genuine user [10, 45, 47, 59, 62]. Among these attacks, replay attacks are extremely accessible and highly effective in spoofing VUIs [56]. After hijacking these devices, an attacker may access confidential information like the users' bank accounts, or eavesdrop privacies via the VUIs located at private places such as living rooms or bedrooms. Furthermore, owing to the proliferation of VUI controllable IoT devices, spoofing attacks could arise more serious breaching. For example, an attacker could easily disarm the alarm system of a house, unlock the smart door lock, or even take over and mislead a smart car [43].

In this paper, we propose VibLive, a liveness detection system designed for VUI capable devices in the IoT environment with three unique features.

Continuity. VibLive supports text-independent, continuous liveness detection without requiring the user to enroll specific passphrases. Continuity is important, since as soon as passing the one-time authentication that simply depends on the enrolled wake-up words, the latest VUI equipped IoT devices like Google Home keep listening

and executing commands [15] without further liveness detection. Therefore, an attacker could take over the speech recognition session either remotely or physically. For example, a Burger King TV advertisement misled Google Home to search and read information about "Whopper hamburger" to users [9]. Unfortunately, most recent liveness or spoofing detection systems are text-dependent. For instance, VoiceLive [64] and VoiceGesture [63] examine human articulating characteristics when the user speaks enrolled passphrases, whereas VoicePop [57] times the user's breath noises in registered sentences. Although CaField [60] is text-independent, it still requires user enrollments with user-dependent sound field features.

Transparency. VibLive provides user-transparent liveness detection that requires no additional cumbersome operations or added hardware other than a speaker and a microphone that are widely equipped on VUI devices. By contrast, most previous liveness detection work needs to collect extra channels of information during standard voice authentication. For example, VAuth [36] requires additional contact sensors to capture on-body movements, whereas WiVo [46] demands WiFi devices to sense articulator motions. CaField [60] requests two spaced microphones to catch the unique gradient of a sound field. This method is well suited to devices like smartphones, whereas not applicable to VUI devices that only equip one microphone.

Applicability. VibLive is applicable to various use scenarios in IoT environments, including short-range authentications (e.g. user to smartphone or app authentication) or long-range voice control, for example, over smart vehicles and smart home devices. Furthermore, VibLive allows variable distances between the user and the VUI device, i.e., the users are free to change their locations. In comparison, many related researches necessitate the user to either stay close or at a fixed location to VUI devices. For example, Echoprint [65] asks the user to hold the smartphone closely to sense the facial landmarks while she speaks. Although CaField has little position constraints, it requires the user to hold the smartphone at a consistent position to extract similar field prints. Additionally, the aforementioned work like VoiceLive [64], VoiceGesture [63], and VoicePop [57] also require the user to locate the device closely at a fixed position to collect the subtle psychological information resulted from articulation.

In particular, VibLive exploits the dissimilarity of bone-conducted vibrations and air-conducted voices when human speaks for continuous liveness detection. Many may realize that our voices sound differently on a recording. This is because when we listen to our own voices, we hear stereo voices formed with both air-conducted voices and bone-conducted vibrations; while others or the recorder could only hear or record air-conducted voices [3]. When human speaks, the vocal folds play the role of the sound source, which oscillates to generate a fundamental laryngeal formant. On one hand, this formant is transited and modulated via the vocal track, and eventually emitted from the human mouth and nose to form air-conducted voices. On the other hand, the formant is modified by the complex human body organizations to build up bone-conducted vibrations. Although sharing the same sound source, the air-conducted voices and the bone-conducted vibrations exhibit distinct acoustic features as the result of different propagation paths and characteristic modulations. Unlike human, the loudspeaker replay input audios by vibrating the diaphragm at the same frequencies precisely. Therefore, the replayed audio wave and the vibrations of the loudspeaker are highly similar. We thus are inspired to compare the air-conducted voices and the corresponding bone-conducted vibrations for continuous liveness detection on VUI capable devices.

To collect both air-conducted voices and bone-conducted vibrations, VibLive leverages the built-in speaker of the VUI capable IoT device to emit inaudible probe signals, and exerts the built-in microphone to record the reverberant probe signals modulated by the bone-conducted vibrations. When the user speaking or the loudspeaker replaying voice commands, the vibrating human head or the loudspeaker lengthen or shorten the distances of the propagation paths accordingly, which results in different attenuations of the probe signals. We analyze the received signal strength (RSS) of the modulated probe signals to extract the vibration. Then we compare the recovered vibrations with the recorded voice by their Linear Prediction Code (LPC) spectra, which exposes the spectral features of the signal modulations for live user and spoofing attack detection. The main contributions of this work are summarized as follows:

- We design VibLive, a liveness detection system for secure VUI in IoT environments. VibLive provides continuous user authentication without cumbersome operations or additional hardware. Moreover, VibLive works for both short-range and long-range liveness detection.
- We develop an acoustic sensing approach to sense the bone-conducted or loudspeaker vibrations. We also design a demodulation method to recover the vibrations from the modulated probe signals. Furthermore, we extract spectral features of the record and the recovered signals for the live user and spoofing attack detection.
- The extensive experiments with 25 participants under a diverse IoT intended experiment settings show that VibLive achieves around 98% accuracy and less than 0.5% FPR. Moreover, the experiments also proved that VibLive is robust to different lengths of speech, different distances, different angles and several attacks.

2 RELATED WORK

Conventional voice authentication has been proved to be vulnerable to spoofing attacks including replay attacks, synthesize attacks, conversion attacks, and impersonation attacks [5, 30, 35, 38, 42, 49, 58]. Among those, replay attacks are extremely accessible and highly effective as the attacker may fool the system merely with a stealthily recording or a piece of concatenated voice from the victim. According to a replay attack detection report, the EER(Equal Error Rate) of the current voice authentication systems could be as high as 31% under replay attacks [41]. Especially, recent work affirms that the adversary could spoof the voice authentication system by replaying modulated voice samples that are inaudible to human ears but recordable and understandable to VUIs [33, 52, 62]. To defend against replay attacks, traditional commercial voice authentication services providers like Nuance [4] and Voice Vault [8] challenge their users to repeat additional passphrases for liveness detection.

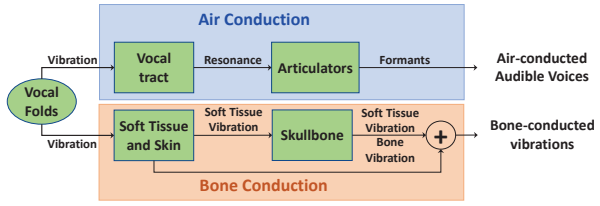

Figure 1: Air Conduction and Bone Conduction.

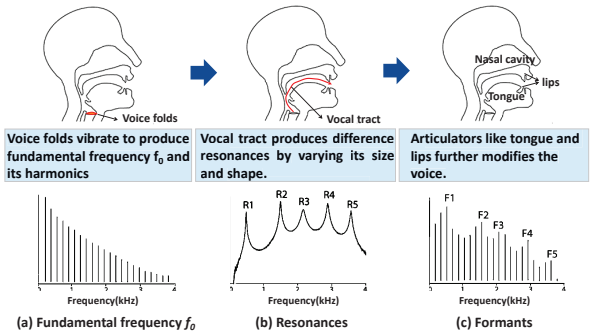

Figure 2: The Formation of Air-conducted Voices: The Source-filter Model.

These methods, however, could be cumbersome as the users are required to enroll a set of commands for the challenge-responses except for the ones for authentication. Whereas a new liveness detection patent filed recently by Amazon compares the recorded voice with the stored previous instances and spots replay attacks if the imbedded watermarks are used [20]. Though effective, this method consumes extensive storage for memorizing all the used watermarks.

Moreover, some smartphone-based voice liveness detection methods have been proposed. Specially, most of these methods try to exert the consequent physiological features of human vocalization that only exist on live user for liveness detection. For example, considering people articulate different phonemes at distinct locations with various manners, Zhang et al. [64] propose to detect the live user by the passphrases' TDoA dynamic to the two microphones of the smartphone; they also show that measuring ultrasound Doppler shifts caused by the articulatory gestures is effective for liveness detection [63]. Likewise, Wang et al. [57] propose to monitor the breathing sound and Zhou et al. [65] observe the facial landmarks movements during voice authentication for liveness detection. However, the effectiveness of these methods decreases with increasing distance between the smartphone and the user. Moreover, all these methods are text-dependent and require the system to enroll the user with specific passphrases in advance.

The aforementioned authentication or liveness detection methods are one-time actions designed to guarantee the security of a single access. However, as VUIs keep listening and executing voice commands as long as it's activated with the wake-up words, a continuous authentication or liveness detection method is imperative. However, such solutions are not well developed and there are only few related researches. For example, VAuth [36] examines the user body surface vibrations and matches them with the microphone collected speech signal for liveness detection on VUI. However, this method requires the user to wear objects like earbuds, eyeglasses or necklaces consistently to hold an accelerometer chip and a Bluetooth transmitter for signal collections and transmission. Whereas Wivo [46] detects the live user by comparing the articulator movements resulted Channel State Information (CSI) changes with the voice commands, and Chen et al. [34] evaluate the magnetic fields of the loudspeaker for spoofing attack detection. Unfortunately, the effectiveness of both methods could be heavily influenced by the location of the user or the attacker. Moreover, Yan et al. [60] propose to measure the unique sound fields of enrolled users for text-independent speaker verification. However, this method requires at least two spaced microphones to catch the sound field gradients when a user speaking at specific locations.

Furthermore, given the recent prevalence of VUIs in the IoT environment. Some researchers propose VUI oriented liveness detection methods. For example, 2MA [31] takes advantage of the ambient and personal devices operating in the same area to verify the presence of a live user. Specifically, it needs to collect the voices and ambient noises with a mobile device held by the user and compares the audio recorded by the mobile device and the one collected by VUIs for liveness detection. However, instead of proving the attendance of the live user, this method could only verify the presence of the assistant device, which also needs extra authorization procedures to decide its ownership. Besides, it could be inconvenient for the user to hold his mobile device whenever he uses the VUI. In addition, Gong et al. [37] suggest protecting the VUI using sound source identification to eliminate the replay attacks. They claim that the sound produced by a playback device usually contains unwanted effects resulted from high-pass filters. However, this method is less effective for high fidelity recorders and loudspeakers.

3 PRELIMINARIES

3.1 Air Conduction and Bone Conduction

During human vocalization, the vibrations of vocal folds propagate through two separated pathways, i.e., air conduction and bone conduction, result in air-conducted voices and bone-conducted vibrations respectively. As shown in Figure 1, with air conduction, the vocal folds vibrations pass the vocal tract and various articulators (e.g., tongue and lips) via air to form resonances and formants progressively. These formants are then emitted through the mouth and nose as air-conducted voices that can be recorded by normal microphones. By contrast, with bone conduction, the vocal folds vibrations transmit and be modulated through soft tissues, skins and skullbones as bone-conducted vibrations, which can only be picked up by attaching vibration sensors like contact microphones or accelerometers to the user.

Air-conducted Voice. The air-conducted voice production could be simplified as a source-filter model [11] like Figure 2. With this model, the air-conducted voice can be viewed as an ordered combination of the fundamental laryngeal sound f_0, resonances, and articulation effects. It characterizes the physiological function of each organ in the air conduction pathway [12]. Specifically, the lung pushes steady airflows through the periodically vibrating vocal folds to generate f_0 and its harmonics $n \cdot f_0$, as shown in Figure 2(a).

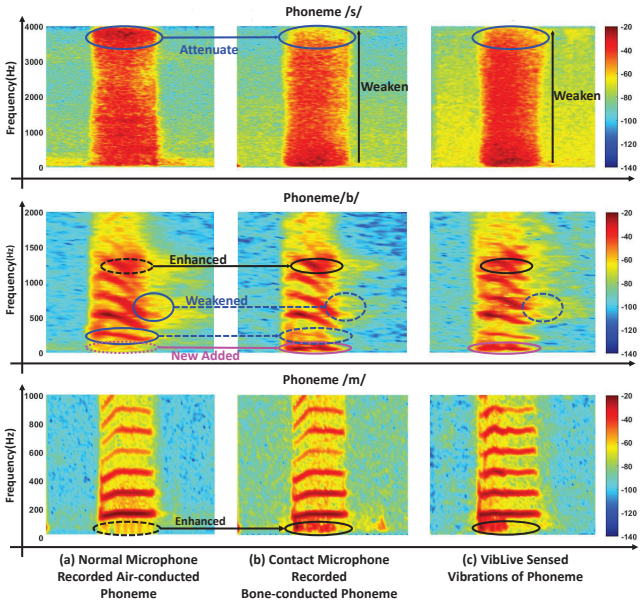

Figure 3: Spectrograms of Air-conducted Voices, Bone-conducted Vibrations, and VibLive Sensed Vibrations for the Same Phonemes.

Next, the harmonic sounds pass through the vocal tract composing of resonators such as the throat, mouth cavity, and nasal passages via air, result in characteristic resonances, as depicted in Figure 2(b). Then as described in Figure 2(c), the resonances are further modified while going through articulators including tongue, lips and so on successively and eventually generate recognizable voiced formants output by mouth and nose.

Bone-conducted Vibrations. Unlike air conduction, bone conduction involves complex pathways due to the considerable complexity of the human skull organizations [39]. Indeed, the human skull composes of various materials including air, aqueous humor, soft tissue, etc. All these materials contribute to bone conduction differently due to their distinct densities (1.2–1900 kg/m^{3}which result in a wide range of speeds of sounds (340–3100 m/s) [51]. In addition, the skull vibration also involves complicated mechanisms. With increasing frequencies, the skull bones vibrate asynchronously as two or more separate parts in different directions and mechanisms to form complex mechanical waves [39, 54]. Additionally, many researches discover that the bone-conducted vibrations are nonlinearly transmitted [40], which causes considerable harmonic distortions at low frequencies comparing with the air-conducted voices.

3.2 Air-conducted Voices, Bone-conducted Vibrations and VibLive sensed vibrations

Given the different propagation pathways and modulation mechanisms, bone-conducted vibrations distinguish themselves from air-conducted voices with several distinct features. For comparison, we record the same phoneme with a normal microphone, a contact microphone, and sensed and recovered with VibLive respectively.

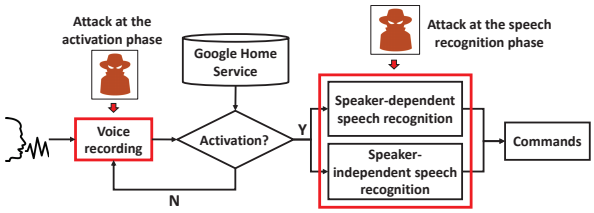

Figure 4: A Typical Case of VUI Workflow and Three Possible Places of Replay Attacks.

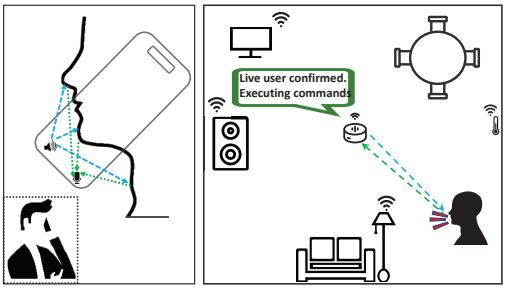

Figure 5: User Cases of Our Liveness Detection System.

During recording, the same user puts a normal microphone in front of her mouth, attaches the contact microphone on her neck, or holds the smartphone with the bottom microphone pointing to her face.

First, due to the attenuation effect of human tissue, bone-conducted vibrations usually miss high-frequency formants. Indeed, Munger and Thomson [50] claim that the signal intensity of bone-conducted vibrations decreases rapidly when the speech frequencies are above $1kHz$. This effect is especially evident for unvoiced consonants like /f/ and /s/, which consist of only noise-like high frequencies. In Figure 3, we could notice that the power of normal microphone recorded air-conducted voices increases uniformly between 0 to $2kHz$ and it surges after $3.5kHz$. By contrast, the power of the contact microphone recorded bone-conducted vibrations and the VibLive sensed vibrations are especially high for frequencies between 0 to $500Hz$. The power then slowly decreases with increasing frequency, and it shows obvious attenuation above $3kHz$.

Second, frequencies below 2 kHz are prone to be affected or even be modified by the bone conduction. Some researchers [61] suggest that, in comparison with air-conducted voices, bone-conducted vibrations are different in attenuated, enhanced, or new introduced formants. This phenomenon is especially obvious in plosive phonemes like /d/ and /g/. Figure 3 shows an example of such inconsistencies with phoneme /b/. Comparing with the normal microphone recorded air-conducted voice /b/, the spectrograms of the contact microphone recorded vibrations /b/ and the VibLive sensed one are highly similar. Specifically, the bone-conducted formants in the solid black circles are enhanced, whereas the formants in the dotted blue circles are weakened, and the formants in the solid magenta circles are new added.

Third, bone conduction could enhance low frequencies of nasal sounds like /m/ and /n/ [48]. We could observe in Figure 3 that the dotted and solid black circles mark noticeable power enhancements

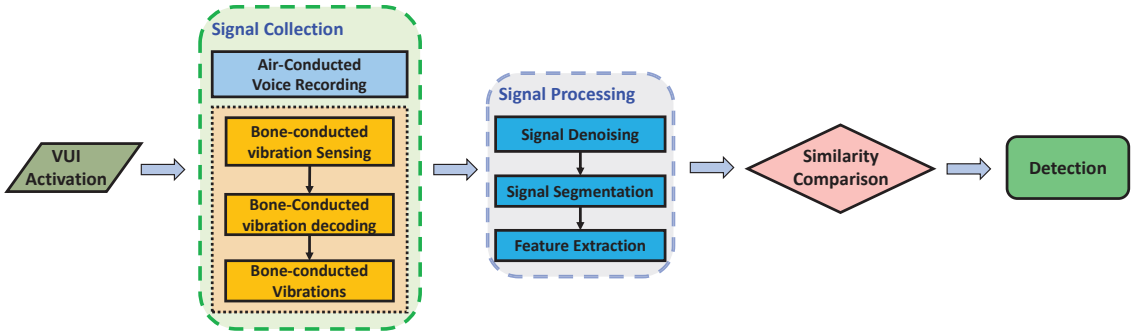

Figure 6: System Flow of VibLive: A Continuous Liveness Detection System.

between the normal microphone recorded voice /*m*/ and the contact microphone recorded vibrations /*m*/. Similarly, such enhancement shows in the VibLive sensed vibrations.

These observations inspire us to design VibLive, which catches the dissimilarities between bone-conducted vibrations and air-conducted voices when the human speak for liveness detection to secure the VUIs in IoT environments.

3.3 System and Attack Model

The state-of-the-art VUI capable devices normally work in two phases: the activation phase and the speech recognition phase. Figure 4 illustrates the main procedures of a typical VUI service. During the activation phase, VUI services like Alexa activates the device as long as someone speaks the right wake-up words like "Alexa". Whereas other services such as Google Home and Apple Siri support a more secure one-time voice authentication with the predefined wake-up words such as "Hey Google", "OK Google", and "Hey Siri". These systems only activate when the recorded voice biometric passes the voice authentication. Once activated, the VUIs either start speaker-dependent speech recognition or speaker-independent speech recognition to recognize the voice commands. The former only processes voice commands spoke by a specific authorized user while the latter accepts voice commands from any sound sources.

For attack models, we assume that the attacker could spoof the VUIs by replaying recorded, synthesized, or converted sounds via a loudspeaker, and we call them replay attacks collectively. Based on the VUI architecture illustrated in Figure 4, the replay attacks could take place under three scenarios. First, the attacker may use the recorded victim's wake-up words to activate the VUI device. After breaking in, the attacker could conduct replay attacks on the speaker-dependent recognition with recorded victim's voices, or attack the speaker-independent one with any malicious commands.

4 SYSTEM DESIGN

4.1 Approach Overview

The underlying principle of our liveness detection lies in the fact that when a live user speaking, the fundamental frequency generated by the vocal folds propagates via two distinct pathways, which results in air-conducted voices and the bone-conducted vibrations respectively with distinguishable features. Our system senses the bone-conducted vibrations with the built-in loudspeaker and the microphone on the VUI capable devices, and compares that with the corresponding microphone recorded air-conducted voices for liveness detection.

As illustrated in Figure 5, during the activation, the built-in speaker of the VUI system emits an inaudible probe signal at 20 kHz. When the user speaks, the built-in microphone records the user's air-conducted voices as well as the 20 kHz probe signal modulated by the bone-conducted vibrations. Despite of sharing the same sound source, the air-conducted voices and the corresponding bone-conducted vibrations are distinguishable due to different propagation modalities and modulations. Once finish recording the voice command, our system examines the amplitude variation of the modulated probe signal, and extracts the bone-conducted vibrations. Next, we compare the recovered vibrations with the recorded air-conducted voices by matching peak sequences of the LP spectrum.

When human speaks, our system detects a live user when the recorded audible voices and the sensed vibrations exhibit distinct features that are shown in Section 3.2. By contrast, we spot a replay attack if the received audio signals and the vibrations of the loudspeaker are highly similar. Our system utilizes inaudible acoustic signal to sense either the bone-conducted vibrations of the live user or the vibration of a loudspeaker, which allows VibLive to serve the VUI capable devices in the IoT environment for liveness detection from both short-range and long-range. Comparing with previous work, our method requires neither additional hardware nor on-body device other than one microphone and one loudspeaker on the VUI capable devices. Moreover, our system conducts text-independent continuous liveness detection, which could be applied to both the activation and speech recognition phase.

4.2 System Flow

The proposed system consists of four major components: *VUI Activation, Signal Collection, Signal Processing, Similarity Comparison.* Figure 6 presents the system flow of VibLive.

As soon as the VUI capable device is activated, our system starts *Signal Collection* with both the microphone and our bone-conducted vibration sensing mechanism simultaneously. The former records the air-conducted voices whereas the latter senses the corresponding bone-conducted vibrations of a live user or the loudspeaker vibrations from an attacker. The sensed signals then pass through

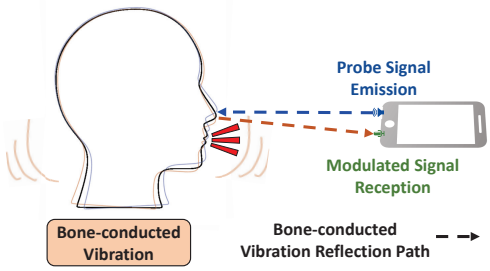

Figure 7: Illustration of Speech Modulation Procedures.

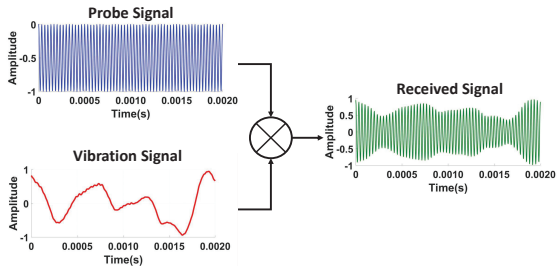

Figure 8: Illustration of the Probe Signal Modulating the Bone-conducted Vibrations.

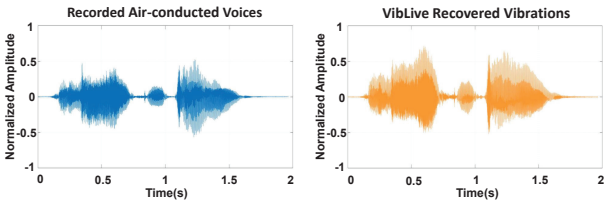

Figure 9: An Example of the Air-conducted Voices and the Corresponding Bone-conducted Vibrations in Time Domain.

the *Bone-conducted Vibration Decoding* to demodulate the bone-conducted vibrations from the 20 kHz probe signal.

For *Signal Processing*, we first apply denoising techniques on both recorded voices and recovered vibrations. Specifically, we adopt a high-pass filter to remove direct current components and disturbances caused by large-scale and slow movements such as the articulators' motions in the experimental environment. We also adopt the spectrum subtraction [32] to further remove background noises. Next, to remove the blank spaces and non-target sounds in the recorded voice, we examine the energy changes in the vibration signal for Signal Segmentation [53]. Then we match the segments' timing information onto the recorded voices to find the corresponding air-conducted voice segments. Afterwards, the *Feature Extraction* component is employed to extract the Linear Prediction(LP) spectrum of both the air-conducted voice and the bone-conducted vibration.

At last, we match the LP spectrums of the air-conducted voice and the recovered vibrations using the Pearson correlation coefficient. The generated similarity score is then compared with an empirical threshold τ. For the live user, the similarity score between the recovered vibrations and recorded voices is lower than τ due to their dissimilarities, whereas for the replay attacks, the similarity score is higher than τ since the recovered vibrations are always exactly the same as the loudspeaker replayed and normal microphone recorded voices.

4.3 Speech Modulation

Once activated, the VUI device collects both the audible air-conducted voices and the bone-conducted vibrations with the loudspeaker and microphone. Figure 7 illustrates the procedures of our system sensing the bone-conducted vibrations. When the user is speaking, the vocal folds' vibrations propagate through human organs like tissues, skins and skullbones. To sense such vibrations, we leverage the built-in loudspeaker of the VUI to emit an inaudible probe signal at 20 kHz, which is then reflected and modulated by the vibrating human vocal tract and head. The reflected signals as well as the corresponding voices are then received by the built-in microphone.

Assuming the user's head vibrates at a mono frequency f_0 with the amplitude A in a short period of time from t_0 to $t_0 + \Delta t$ while he speaks. In this time window, the bone-conducted vibration caused displacements Δs could be represented as:

$$\begin{aligned} \Delta s &= s(t_0 + \Delta t) - s(t_0) \\ &= A sin(2\pi f_0 \Delta t) \end{aligned} \tag{1}$$

Where $s(t_0 + \Delta t)$ and $s(t_0)$ are the displacements of human head at $t_0 + \Delta t$ and t_0 respectively. This equation omits the initial phase of the user's head at t_0.

Meanwhile, in the Line-of-Sight (LOS) scenario, once emitted from the built-in speaker, the probe signal attenuates while propagating through the air with increasing distance. Moreover, in our case, the transmission distances vary due to the bone-conducted vibrations resulted of human speeches, which causes different sound attenuations. According to [2], the corresponding sound pressure level (SPL) changes proportionally to the transmission distance, i.e. $p \propto \frac{1}{L}$. Therefore, consider during Δt, the probe signal transmits with distances $L(t_0)$ and $L(t_0 + \Delta t)$ from the sound source (built-in loudspeaker) to the reflector (human), which result in SPL $p(t_0)$ and $p(t_0 + \Delta t)$ respectively. We could express the Eq. (1) with the probe signal transmission distance:

$$\begin{aligned} \Delta s &= \frac{1}{2}(L(t_0 + \Delta t) - L(t_0)) \\ &\propto \frac{1}{p(t_0 + \Delta t)} - \frac{1}{p(t_0)} \\ &= \frac{1}{p(t_0 + \Delta t)p(t_0)} \Delta p \end{aligned} \tag{2}$$

where the $\Delta p = p(t_0 + \Delta t) - p(t_0)$, which is the SPL changes from t_0 to $t_0 + \Delta t$. Given Δt is extremely short, $\Delta p \ll p(t_0)$, thus in Eq. (2), $\frac{1}{p(t_0+\Delta t)p(t_0)} \approx \frac{1}{p(t_0)^2}$, and the equation could be denoted as:

$$\Delta s = \frac{1}{p(t_0)^2} \Delta p \tag{3}$$

To be noticed, the previous reasoning ignores the reflection loss factor and the angle between the human and the VUI device, which only introduce constants or change the results linearly. As $\frac{1}{p(t_0^2)}$ is a constant, we could draw the conclusion that bone-conducted

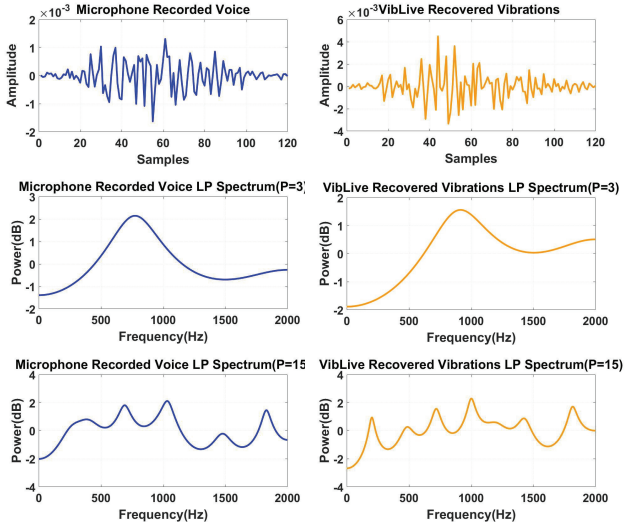

Figure 10: The LP Spectrum of the Air-conducted Voices and the Corresponding Bone-conducted Vibrations.

vibrations are proportional to the SPL changes of the probe signal. Therefore, we could consider that bone-conducted vibrations exert amplitude modulation on the probe signal. Figure 8 illustrates an example of such modulation. We could observe that with this assumption, the receiving signal's amplitude changes according to the vibrations, whereas its frequency is still the same as the probe signal.

4.4 Speech Demodulation

Once finish receiving the commands, we extract the bone-conducted vibrations from the modulated signal by examining the SPL changes of the 20 kHz probe signal along time.

Given T seconds of received signal $r(t)$ that sampled at frequency f_s, our system collects $M = f_s T$ samples. We divide these samples into a list of equal-length discrete signals with N points, and in total we generate $\lceil \frac{M}{N} \rceil$ segments. We assume that the size of these segments are small enough, thus within such short time duration, the received signal is constant. We calculate the Discrete Fourier Transform(DFT) of each segment as:

$$R_k = \sum_{n=0}^{N-1} r_n e^{-j2\pi kn/N} \tag{4}$$

Where r_n is a segment of the received signal and k is the frequency bin index. Next we extract the SPL of 20 kHz from each short segment at the DFT bin indexed by $k = \frac{20kHz}{f_s} N$. We concatenate the sequence of SPLs as the VibLive sensed vibrations. The size of the segments determines the frequency resolution of the DFT and the total frequency bin number. Moreover, in our case, it further determines the sampling frequency of the recovered bone-conducted vibration signal. Specifically, the larger the segment, the higher the DFT frequency resolution, however, the narrower the recovered frequency range. Especially when N is too small, the recovered audio shows signal aliasing distortion [18] since the sampling frequency is too low to cover the signal frequency range [26].

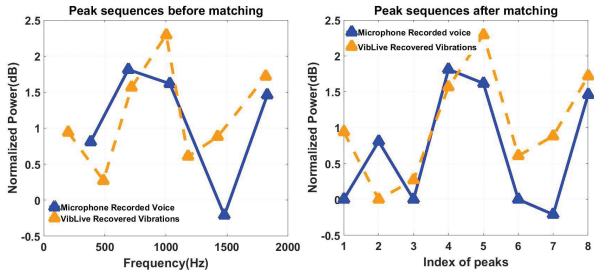

Figure 11: A LP Spectrum Peak Sequence Before and After Matching.

Considering the frequency of human voice is normally under 2 kHz, and the most important frequency range is between 0 to 2 kHz [7], we choose N equals to 48 samples to recover a 2 kHz frequency range. Figure 9 shows an example of the recovered vibrations and the air-conducted voices when the user speaks two words "Hello Google". We could observe that, despite of time-aligned, the recovered vibrations and the air-conducted voices show distinct features.

4.5 Feature Extraction

The differences between the microphone recorded air-conducted voices and the corresponding recovered vibrations are subtle but non-negligible. Unfortunately, the recovered signal from the vibrating loudspeaker replaying the same recorded voice show differences from the original recorded audio as well. These differences, however, are identified as noises that are introduced by algorithms or hardwares such as the lower sampling frequency and audio Analog-to-Digital(AD) conversion. Therefore, a fault-tolerant feature extraction method is critical. In summary, the purpose of our feature extraction method is to detect the distinct features between the recorded voice and the recovered bone-conducted vibrations, whereas neglect the minor differences between the recorded voice and the recovered loudspeaker vibrations.

We extract the spectral features of the signals using Linear Prediction (LP), which models the audio signal with an all-pole infinite impulse response(IIR) filter [55] and is widely adopted for audio coding(Linear Prediction Coding, LPC) and speech analysis. In particular, a P prediction order LP analysis predicts the n^{th} signal sample based on P past samples:

$$\begin{aligned} s(n) &= -a_1 s(n-1) - a_2 s(n-2) - \cdots - a_P s(n-P) + e(n) \\ &= \sum_{k=1}^{P} a_k s(n-k) + e(n) \\ &= \hat{s}_n + e(n) \end{aligned} \tag{5}$$

where a_k are the prediction coefficients, $\hat{s}_n$ is the prediction of the speech samples, and $e(n)$ is called the prediction error or LP residual. With the z-transform of the $e(n)$:

$$\begin{aligned} E(z) &= S(z) - \hat{S}_z \\ &= S(z)\,[1 - P(z)] \\ &= S(z)A(z) \end{aligned} \tag{6}$$

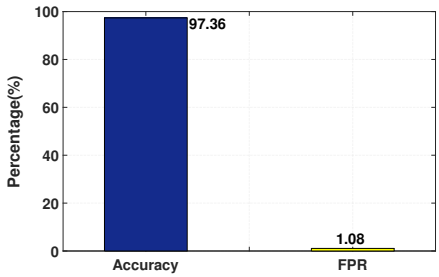

Figure 12: Overall Performance.

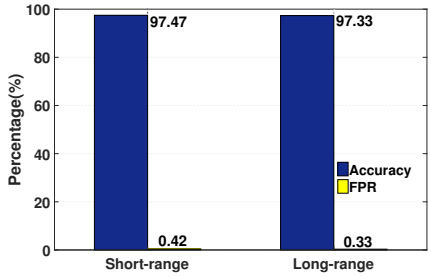

Figure 13: Different Ranges: Accuracy and FPR.

$P(z)$ is the prediction filter. The LPC uses the excitation $e(n)$ as the input to the all-pole IIR LPC filter:

$$H(z) = \frac{1}{1 - P(z)} \tag{7}$$

Thus we could calculate the estimated signal as:

$$S(z) = E(Z)H(Z) \tag{8}$$

The principle underlying LP audio processing is to consider the filter $H(z)$ as the time-varying vocal tract that decides the modulation mechanism whereas the $e(n)$ is the sound source. By choosing proper order of the LP analysis, VibLive enable to reveal enough differences between the air-conducted voices and the VibLive recovered vibrations of live users, nevertheless ignore the undesirable details.

Figure 10 presents one example of our LP analysis. We obtain two time-aligned frames from the microphone recorded voices and the VibLive recovered vibrations respectively. These frames are intercepted with Hamming window and all the LP spectrums are Z-score normalized [28]. We could notice that the low order LP spectrum only captures the coarse information of the speech spectrum, whereas the higher order one could capture more fine-grained details. Empirically, we choose 15^{th} order LP analysis [44] to reveal the different modulations between the air-conducted voice and the bone-conducted vibration.

4.6 Similarity Comparison

According to our assumptions, comparing with the air-conducted voice, the corresponding VibLive recovered vibration may vary by means of adding new formats or deleting, weakening or strengthening existing formants and so on. Therefore, a specific peak shows in the LP spectrum of the air-conducted voices may be larger, smaller or even disappear in the LP spectrum of the corresponding recovered vibrations. To compare the relative powers of the formants, our system extracts the peaks in the LP spectrum and concatenate them to be a peak sequence for similarity comparison.

We calculate the correlation coefficient to measure the linear relationship between the peak sequences. Although more sophisticated classification methods like the learning-based ones could be used, our evaluation targets on verifying the system methodology in this work. Before the calculation, our system searches for the matching peaks in both peak sequences. For a specific peak with the frequency of f in one LP spectrum, we only search for the matching peak within $[f - 50, f + 50]$ Hz frequency range in the other LP spectrum. When the system fails to find the peak in that range, it considers that the peak is missing, and will add a zero at that frequency. Figure 11 illustrates the original peak sequences and the ones after peak matching.

5 EXPERIMENTAL EVALUATION

5.1 Experimental Setup

Experiment Scenarios. We evaluate our system in both short and long-range to simulate the real-world application scenarios of VUI capable devices in the IoT environment. Specifically, we conduct short-range experiments based on the common use cases when people use the VUIs on their portable smart devices like smartphones for applications like secure device access or application login. In these cases, the user normally keep the device within 0.5m from himself, and puts the bottom microphone of the smartphone towards his face for voice recording. To test different VUIs, we employ three types of smartphones including a Samsung Galaxy S5, Samsung Galaxy Note5 and a Galaxy S8+ for the evaluation. These smartphones are different in audio chips and their operating systems, but they all support recording and replay up to 20 kHz.

Moreover, we design our long-range experiments based on VUI enabled IoT applications, such as in smart homes and smart vehicles. In these cases, the users are free to give voice commands from any locations in a typical size of a room or vehicle. We choose the distance above 0.5 m and up to 3 m as our long-range scenario test distance based on the smart home and smart vehicle applications. We conduct the long-range experiments with both smartphones and a same grade of microphone to the build-in one of Google Home due to lack of raw data from smart devices like Google Home.

Data Collection. We have 25 participants for the experiments. These participants are recruited by emails including 13 females and 12 males, both undergraduate and graduate students, native and non-native English Speakers, whose age range from 20 to 32. We inform these participants about the purpose of our experiments and ask them to act naturally as when they are talking to their own VUI devices. Since our system supports text-independent voice authentication, each of our participants is suggested to choose or design 10 pieces of speeches for authentication without enrollment. The lengths of these speeches vary from 5 to 25 words, among which, one third are 5 to 7 words, one third are 8 to 11 word, and one third are more than 12 words. To test our liveness detection system, a participant repeats 10 times for each piece of speech, which generates 2500 live user cases.

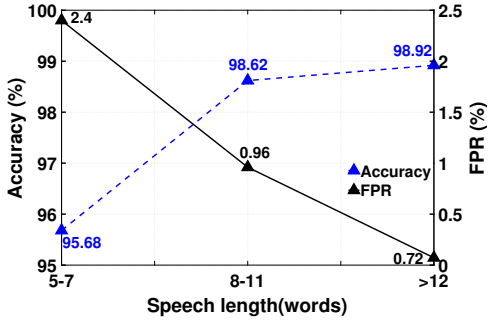

Figure 14: Impact of Speech Lengths.

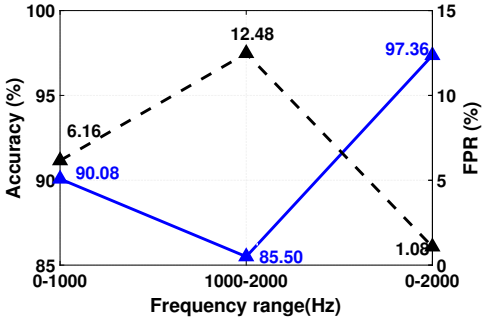

Figure 15: Impact of Frequency Ranges.

During the authentication, the speakers are allowed to choose their comfortable way of holding or locating the device with a short or long distance from them, and to face or not to the device. In particular, among the 25 participants, 19 of them choose short-range authentication whereas 6 of them adopt the long-range one. Thus, we collect 1900 short-range positive cases and 600 long-range positive cases. Additionally, our experiments are conducted in different environments such as the classroom, apartments, offices including noises like HVAC noises and people chatting. Specific, there also exists other movements such as people walking, typing, or cooking and so on in the environment.

Replay Attacks. We evaluate our system against replay attacks. The replay attacks are conducted with three kinds of loudspeakers including a DELL AC411 Speaker system, a Klipsch Groove Portable Bluetooth Speaker, and a Logitech Sound Z625 Speaker System. For each piece of speech, the three kinds of loudspeakers replay 3, 3 and 4 times respectively, which adds up to 2500 negative cases. The replay attacks are captured with the same device, at the same distance range as the participants used for authentication. Moreover, we consider the situations when the adversaries try to occlude the replay attacks by covering the loudspeaker, turning down the volume of the loudspeaker or changing the facing angle of the loudspeaker. In particular, we conduct 300 occlude experiments, 400 SPL experiments, and 300 different angle experiments.

Metrics. We present the experimental results with the following metrics. The *FPR*(False Positive Rate) is the chance that the system mistakenly detects the replay attacker as a live user. The *Accuracy* is the rate that our system makes the right decision about the replay attacks and the users.

5.2 Overall Performance

We first present the overall performance of our liveness detection system. Figure 12 shows the Accuracy and the FPR of our system with both short-range and long-range attacks. We could observe that the overall Accuracy of our system is 97.36% whereas the FPR is 1.08%, which shows that our liveness detection system is accurate at distinguishing live users and the replay attacks. Furthermore, the low FPR suggests that our system is especially effective at spotting the replay attacks.

Given the IoT environment where the users may communicate with the VUI capable devices in both short or long distances, we test our system with the user or the loudspeaker in different distances from the device. Figure 13 shows that the Accuracy of short-range and long-range experiments are 97.47% and 97.33% respectively, and the FPR are 0.42% and 0.33%. We could notice that the accuracy of the long-range liveness detection is slightly lower than that of the short-range one, while the corresponding FPR is smaller. This suggests that the system could misjudge the live user as replay attacks at a higher rate when the user is farther away. Nevertheless, the lower FPR shows that our system is more efficient at detecting replay attacks. One possible explanation is that the live users are more sensitive to the multipath effect of longer distances.

5.3 Impact of Speech Length

Our system detects replay attacks continuously as soon as the VUI is receiving speeches, thus we study how the length of the speech could affect the effectiveness of our liveness detection system. We categorize the speeches into 3 groups based on their lengths, i.e. 5 to 7 words, 8 to 11 words, and more than 12 words. Figure 14 shows that with the length of the speech increases, the accuracy rises from 96.58% to 98.92%, whereas the FPR drops from 2.40% to 0.72%. This is because that our system tries to extract distinct features from the air-conducted voices and recovered vibrations. However, given the same sound source, these two signals share a high similarity in general. Therefore, with a longer piece of speech, the system would be able to accumulate more differences for a more accurate liveness detection. Further, we could notice that with the speech length increases from 8 to 11 words to above 12 words, the improvement of the accuracy slows down, nevertheless, the FPR still drops at a promising rate. Thus as the user keeps giving voice commands continuously to the VUI device, the security of our system will keep getting enhanced.

5.4 Different frequency range

We extract the amplitude changes of the ultrasound for bone-conducted vibration measurement. After receiving the signal, we could choose different DFT sizes to recover different frequency ranges. In this section, we examine which frequency range contains more critical information for liveness detection. Due to the limitation of the device, the largest frequency range we could recover is from 0 to 2000Hz. As we could observe from Figure 15, the accuracy/FPR for frequency ranges 0 to 1000Hz, 1000 to 2000Hz, and 0 to 2000Hz are 90.08%/6.16%, 85.50%/12.48% and 97.36%/1.08% respectively. This result matches our preliminary study that the low frequency part of the bone-contacted vibration contains more

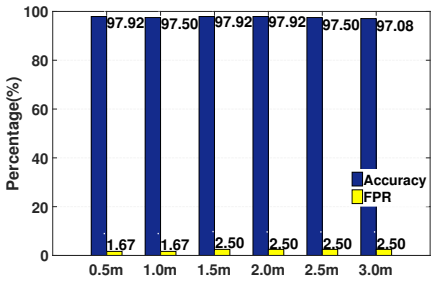

Figure 16: Impact of Different Distances.

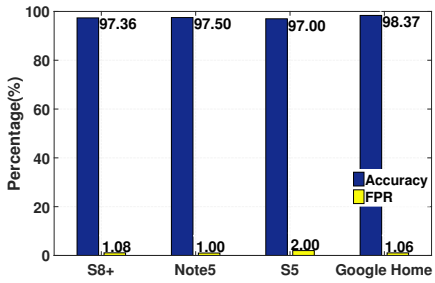

Figure 17: Impact of Different Devices.

differences from its corresponding air-contacted voices. Nevertheless, the higher frequency range could still provide complementary information. Therefore the largest frequency range achieves the best accuracy and FPR.

5.5 Different Distance between User and the Device

We conduct the liveness detection with different distances between the speaker and the VUI device. Specifically, we evaluate our system with 6 distances ranges 0.5m to 3.0m. We choose this distance range due to the application scenarios of the VUI devices in the IoT environment: when people use their smartphone as the VUI, we normally hold the smartphone within 0.5m from us; whereas when people interact with the VUI in their cars, the distance between the speaker and the VUI increases to the range around 0.5m to 1.5m; and for the case of IoT environment, a user could stay even further away. We choose 3m as the upper bound of these experiments due to the typical size of a living room. The results show that the performance of our system stays stable with increasing distance. Indeed, the accuracy stays around 97.50% for all the tested distances, until when the distance increases to 3m, and the accuracy drops slightly to 97.08%. However, the FPR has been around 1.67% when the distance is smaller than 1m, and increases to 2.50% for the distance from 1.5m to 3m. The results suggest that our system could be adopted in various VUI applications in the IoT environment.

5.6 Different Devices

Our system supports different types of VUI capable devices. Specifically, our experiments involve three types of smartphones including Note5, S5 and S8. Further, we simulate the user case of other popular IoT devices like Google Home on the same grade of equipments. Indeed, due to the lack of raw data, we use an external microphone with the same grade to the Google Home built-in microphone [6, 25, 27] to record the reflected signals. Specifically, these two kinds of microphones are similar in terms of the polar pattern (omnidirectional) and the frequency response (around 45Hz to > 20kHz). Indeed, the Google Home built-in microphone processes some better features such as the high SNR and enhanced sensitivity. Figure 17 shows the accuracy and FPR of using our method on these devices. We could observe that the liveness detection accuracy are 97.50%, 97.00%, 97.36%, and 98.37% whereas the FPR are 1.00%, 2.00%, 1.08% and 1.06%. According to the results, the accuracy of our system barely changes with different smartphones whereas the accuracy of our system is higher with the Google Home simulation. We assume that when replaying the probe signal with the loudspeaker, the signal power is much higher, thus the changes of the reflected signal are more recognizable.

5.7 Attack Scenarios

In this section, we consider situations that the attacker have knowledge to our liveness detection system, and try to tamper the vibrations of the loudspeaker so that the system would misjudge the similarity between the loudspeaker vibrations and the output audio, and take the replay attack as a live user. We come up with three possible methods to hide the vibrations from our sensing mechanism. For these sets of experiments, the distance between the loudspeaker and the VUI device is fixed at around $0.30m$ unless claimed otherwise.

Occlude. The adversary may occlude the loudspeaker, and thus disturb the propagation of the sensing signal. We experiment with three types of covers, including a standard A4 printing paper, a piece of common cardboard with the dimension of $170 \times 120 \times 2mm$, and a hardcover book measures $140 \times 100 \times 12mm$. We put these objects in front of the loudspeaker to fully cover the entire body of the loudspeaker, thus to block the Line-of-Sight(LOS) propagation path between the loudspeaker and the microphone. These objects are held stably by an assistant during the experiments. Additionally, we adjust the volume of the loudspeaker to around 60dB, which is the average SPL for normal conversation [1]. Figure 18 shows the results of the experiments. As we could observe, with the increasing thickness of the cover, the accuracy of our system decreases from 97.08% with paper, to 93.33% with the hardcover book. Indeed, comparing with the overall accuracy of the system at 97.36%, the system performance drops slightly when the loudspeaker is occluded by a paper. However, when using the cardboard for occlusion, our system accuracy drops around 3.00%. Nevertheless, the accuracy barely drops when the thickness of the cover increases 6 times from the cardboard to a hardcover book. The results indicate, with occlusions, our system performance could be influenced, however, the accuracy does not keep dropping when increasing the thickness of the occlusion object. Indeed, our liveness detection system could still achieve more than 93.00% accuracy and around 4.00% FPR with thick occlusion objects like a hardcover book.

SPL. Next, we evaluate our system performance against the changes of SPL. We have tested four SPLs, including 30dB, 50dB, 70dB, and 90dB. The typical sound sources of these levels could

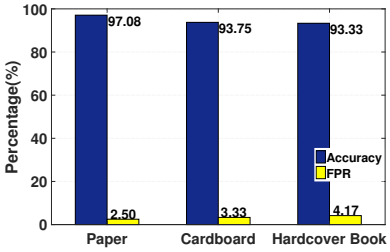

Figure 18: Attack: Occlude.

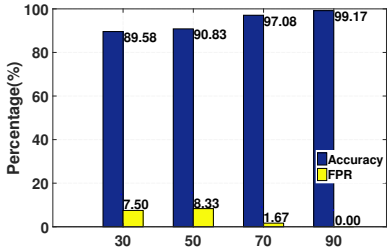

Figure 19: Attack: Low SPL.

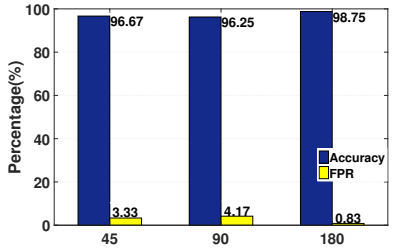

Figure 20: Attack: Angles.

be quite whisper, whispered speech, background conversation in a restaurant, and food blender respectively. We change the SPL by adjusting the volume of the loudspeaker and measure these SPL values with a decibel meter [22] by the side of the VUI device. As we could observe from Figure 19, increasing the SPL could improve our system accuracy dramatically from around 90% to above 99%, whereas the FPR drops from around 7.00% to 0%. This is because the SPL determines the vibration degree of the loudspeaker, and the loudspeaker vibrates at a larger degree when turning up the volume of the loudspeaker. Therefore the recovered signal gains higher Signal Noise Ratio (SNR) and matches better with the recorded audio signal. To be noticed, although the system performance is not as good with an SPL below 50dB, it's hardly the case that people will talk with the VUI by whispering. Especially, when the user speaks or the loudspeaker replays the command with a SPL lower than 50dB, it is highly possible that the built-in microphone of the VUI devices could not catch the air-conducted voice and thus fail to execute the voice command anyway.

Angle. We also examine the situation when the attackers change the angle of the loudspeaker to reduce the reflection surface of the probe signals, and thus sabotage the recovered vibrations. We consider when the loudspeaker diaphragm facing directly to the speaker's head, the angle between them is 0°. Then we set up the angle experiments by turning the loudspeaker by 45°, 90°, and 180°. The loudspeaker we utilize in the experiment is a satellite of the Logitech Z625 2.1 Speaker System. Figure 20 shows that for these three angles, the accuracy is 96.67%, 96.25% and 98.75% respectively. Given the vibrations of the loudspeaker are mainly drove by the diaphragm's one-dimensional movement, the results are impressive. When the angle is 45°, the vibration distance changes to $d \cdot sin(45^\circ)$, and the accuracy drops slightly; while when the angle is 90°, the vibration distance reduced to the minimum and so does the accuracy; whereas when the angle is 180°, the vibration distance is the same with 0°, and thus the performance is comparable with the results with that of the 0°. Nevertheless, our system still provides high accuracy at around 96% even with the worst angle.

6 DISCUSSION

First, VibLive works when the user or attacker stays meters away from the device. We do not consider the situation when the user is totally in another room in this work, but we would include this evaluation in our future work. Further, considering the density of the VUI devices in a modern smart home, the user and the VUI device in use are most likely located in the same room. Besides, for other scenarios like the smart car, the user could only use the VUI device within the car.

Second, we conduct our experiments in different environments with noises like HVAC noises, people chatting, walking, typing and cooking and so on. However, we do not evaluate our system in extremely noisy environments like factories. In these contexts, the VUI devices may not function normally as fail to pick up the voice commands of one specific user from the noisy background. Indeed, people normally communicate with contact microphones or other wearable devices in these scenarios.

Last but not least, our system does require ultrasound like 20kHz as the probe signal. The built-in loudspeakers of most state-of-the-art smartphones and few smart home speakers like the Amazon Echo Studio [19] support that. However, many other VUI devices only cover the frequencies that human could hear. Nevertheless, with the advanced technique, it is not expensive to upgrade the devices.

7 CONCLUSION

In this paper, we implement and evaluate a liveness detection system designed for VUI capable devices in the IoT environment. The insight is that the dissimilarity of bone-conducted vibrations and air-conducted voices when human speaks could be leveraged for liveness detection. We develop an acoustic sensing approach to sense and recover the bone-conducted vibrations without additional hardware other than a loudspeaker and a microphone that are commonly equipped on VUI capable devices. Our system is highly practical as it is transparent to the users and does not require any cumbersome operations. Furthermore, our system supports text-independent liveness detection, which could secure the whole human-VUI communication. The experimental results show that our system is effective under various setups including short-range and long-field, different lengths of speeches, and different distances. Moreover, we also considered a few types of attacks that might spoof our liveness detection system. Extensive experiment results show that our system could achieve over 97% accuracy in liveness detection.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their insightful feedback. This work was partially supported by the NSF Grants CNS-1835963, CNS-1910519, CNS-1514238, and DGE-1565215.

REFERENCES

[1] 2008. Personal Music Players and Hearing. https://ec.europa.eu/health/scientific_committees/opinions_layman/en/hearing-loss-personal-music-player-mp3/l-2/2-sound-measurement-decibel.htm.

[2] 2012. Dumping of sound level vs distance. http://www.sengpielaudio.com/calculator-distance.htm.

[3] 2013. Why does your voice sound different on a recording. http://www.bbc.com/future/story/20130913-why-we-hate-hearing-our-own-voice.

[4] 2015. VocalPassword. http://www.nuance.com/ucmprod/groups/enterprise/@webenus/documents/collateral/nc_015226.pdf.

[5] 2016. Adobe Voco 'Photoshop-for-voice' causes concern. https://www.bbc.com/news/technology-37899902.

[6] 2016. All the Technology Inside Your Google Home. https://electronics360.globalspec.com/article/7718/all-the-technology-inside-your-google-home.

[7] 2016. FACTS ABOUT SPEECH INTELLIGIBILITY. https://www.dpamicrophones.com/mic-university/facts-about-speech-intelligibility.

[8] 2016. VoiceVault. http://www.nuance.com/landingpages/products/voicebiometrics/vocalpassword.asp.

[9] 2017. Burger King's new ad forces Google Home to advertise the Whopper. https://www.theverge.com/2017/4/12/15259400/burger-king-google-home-ad-wikipedia.

[10] 2017. Google Home now recognizes your individual voice. https://money.cnn.com/2017/04/20/technology/google-home-voice-recognition/index.html.

[11] 2017. The Source Filter Theory. http://my.ilstu.edu/~jsawyer/resonancesoftchalk/resonancesoftchalk7.html.

[12] 2017. The Voice Foundation. https://voicefoundation.org/health-science/voice-disorders/anatomy-physiology-of-voice-production/understanding-voice-production/.

[13] 2018. Designing a VUI-Voice User Interface. https://uxplanet.org/designing-a-vui-voice-user-interface-c0b3b9b57ace.

[14] 2018. How Voice User Interface is taking over the world, and why you should care. https://medium.com/@goodrebels/how-voice-user-interface-is-taking-over-the-world-and-why-you-should-care-54474bd56f81.

[15] 2018. protecting privacy on VUI. https://medium.com/grandstudio/protecting-privacy-in-voice-user-interfaces-b800e47728.

[16] 2019. The 13 Best Smart Home Devices and Systems of 2019. https://blog.hubspot.com/marketing/smart-home-devices.

[17] 2019. 7 Key Predictions For The Future Of Voice Assistants And AI. https://clearbridgemobile.com/7-key-predictions-for-the-future-of-voice-assistants-and-ai/.

[18] 2019. Aliasing. https://en.wikipedia.org/wiki/Aliasing.

[19] 2019. Amazon Echo Studio: First impressions. https://www.techhive.com/article/3441216/amazon-echo-studio-first-impressions.html.

[20] 2019. Amazon files patent for replay attack detection method to protect voice authentication. https://www.biometricupdate.com/201901/amazon-files-patent-for-replay-attack-detection-method-to-protect-voice-authentication.

[21] 2019. Biometrics: authentication and identification (2019 review). https://www.gemalto.com/govt/inspired/biometrics.

[22] 2019. Decibel Meter. https://www.pce-instruments.com/us/measuring-instruments/test-meters/decibel-meter-kat_162375.htm.

[23] 2019. Google express store. https://express.google.com/stores.

[24] 2019. in-car voice assistant consumer adoption report. https://voicebot.ai/wp-content/uploads/2019/01/in-car_voice_assistant_consumer_adoption_report_2019_voicebot.pdf.

[25] 2019. INMP621 Wide Dynamic Range Microphone with PDM Digital Output. https://www.invensense.com/products/digital/inmp621/product-documentation.

[26] 2019. Nyquist frequency. https://en.wikipedia.org/wiki/Nyquist_frequency.

[27] 2019. smartLav+ Lavalier micropohne for smartphones. http://www.rode.com/microphones/smartlav-plus.

[28] 2019. Standard score. https://en.wikipedia.org/wiki/Standard_score.

[29] 2019. Voice Is New UI. What Does It Mean for Enterprises? https://www.linkedin.com/pulse/voice-new-ui-what-does-mean-enterprises-sujatha-visweswara.

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

[31] Logan Blue, Hadi Abdullah, Luis Vargas, and Patrick Traynor. 2018. 2ma: Verifying voice commands via two microphone authentication. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*. ACM, 89–100.

[32] Steven Boll. 1979. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing* 27, 2 (1979), 113–120.

[33] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. 2016. Hidden voice commands. In *25th {USENIX} Security Symposium ({USENIX} Security 16)*. 513–530.

[34] 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. In *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 183–195.

[35] 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 Transactions on Audio, Speech, and Language Processing* 20, 8 (2012), 2280–2290.

[36] Huan Feng, Kassem Fawaz, and Kang G Shin. 2017. Continuous authentication for voice assistants. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*. ACM, 343–355.

[37] Yuan Gong and Christian Poellabauer. 2018. Protecting voice controlled systems using sound source identification based on acoustic cues. In *2018 27th International Conference on Computer Communication and Networks (ICCCN)*. IEEE, 1–9.

[38] 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*. 930–934.

[39] Paula Henry and Tomasz R Letowski. 2007. *Bone conduction: Anatomy, physiology, and communication*. Technical Report. Army research lab aberdeen proving ground md human research and engineering.

[40] Shyam M Khanna, Juergen Tonndorf, and Judith E Queller. 1976. Mechanical parameters of hearing by bone conduction. *The Journal of the Acoustical Society of America* 60, 1 (1976), 139–154.

[41] 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).

[42] Tomi Kinnunen, Zhi-Zheng Wu, Kong Aik Lee, Filip Sedlak, Eng Siong Chng, and Haizhou Li. 2012. Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4401–4404.

[43] Xinyu Lei, Guan-Hua Tu, Alex X Liu, Chi-Yu Li, and Tian Xie. 2017. The insecurity of home digital voice assistants-amazon alexa as a case study. *arXiv preprint arXiv:1712.03327* (2017).

[44] John Makhoul. 1973. Spectral analysis of speech by linear prediction. *IEEE Transactions on Audio and Electroacoustics* 21, 3 (1973), 140–148.

[45] Khalid Mahmood Malik, Hafiz Malik, and Roland Baumann. 2019. Towards Vulnerability Analysis of Voice-Driven Interfaces and Countermeasures for Replay. *arXiv preprint arXiv:1904.06591* (2019).

[46] Yan Meng, Zichang Wang, Wei Zhang, Peilin Wu, Haojin Zhu, Xiaohui Liang, and Yao Liu. 2018. 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*. ACM, 81–90.

[47] Yan Meng, Wei Zhang, Haojin Zhu, and Xuemin Sherman Shen. 2018. Securing consumer IoT in the smart home: architecture, challenges, and countermeasures. *IEEE Wireless Communications* 25, 6 (2018), 53–59.

[48] Nafeesa Mubeen, A Shahina, A Nayeemulla Khan, and G Vinoth. 2012. Combining spectral features of standard and throat microphones for speaker identification. In *2012 International Conference on Recent Trends in Information Technology*. IEEE, 119–122.

[49] Dibya Mukhopadhyay, Maliheh Shirvanian, and Nitesh Saxena. 2015. All your voices are belong to us: Stealing voices to fool humans and machines. In *European Symposium on Research in Computer Security*. Springer, 599–621.

[50] Jacob B Munger and Scott L Thomson. 2008. Frequency response of the skin on the head and neck during production of selected speech sounds. *The Journal of the Acoustical Society of America* 124, 6 (2008), 4001–4012.

[51] William D O'Brien Jr. 2009. *Evaluation of acoustic propagation paths into the human head*. Technical Report. ILLINOIS UNIV AT URBANA BOARD OF TRUSTEES.

[52] Nirupam Roy, Haitham Hassanieh, and Romit Roy Choudhury. 2017. Backdoor: Making microphones hear inaudible sounds. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 2–14.

[53] B Sudhakar and R Bens Raj. 2013. Automatic speech segmentation to improve speech synthesis performance. In *2013 International Conference on Circuits, Power and Computing Technologies (ICCPCT)*. IEEE, 835–839.

[54] Georg v. Békésy. 1932. Zur theorie des hörens bei der schallaufnahme durch knochenleitung. *Annalen der Physik* 405, 1 (1932), 111–136.

[55] Palghat P Vaidyanathan. 2007. The theory of linear prediction. *Synthesis lectures on signal processing* 2, 1 (2007), 1–184.

[56] Jesús Villalba and Eduardo Lleida. 2011. Detecting replay attacks from far-field recordings on speaker verification systems. In *European Workshop on Biometrics and Identity Management*. Springer, 274–285.

[57] Qian Wang, Xiu Lin, Man Zhou, Yanjiao Chen, Cong Wang, Qi Li, and Xiangyang Luo. 2019. VoicePop: A Pop Noise based Anti-spoofing System for Voice Authentication on Smartphones. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2062–2070.

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

conference on machine learning and cybernetics, Vol. 4. IEEE, 1708–1713.

[59] 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.

[60] Chen Yan, Yan Long, Xiaoyu Ji, and Wenyuan Xu. 2019. 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*. 1215–1229.

[61] Bayya Yegnanarayana, A Shahina, and MR Kesheorey. 2004. Throat microphone signal for speaker recognition. In *Eighth International Conference on Spoken Language Processing*.

[62] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. 2017. Dolphinattack: Inaudible voice commands. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 103–117.

[63] Linghan Zhang, Sheng Tan, and Jie Yang. 2017. 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*. ACM, 57–71.

[64] 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.

[65] Bing Zhou, Jay Lohokare, Ruipeng Gao, and Fan Ye. 2018. EchoPrint: Two-factor Authentication using Acoustics and Vision on Smartphones. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. ACM, 321–336.