Liveness Detection for Voice User Interface via Wireless Signals in IoT Environment

Yan Meng, Haojin Zhu, *Senior Member, IEEE*, Jinlei Li, Jin Li, and Yao Liu, *Senior Member, IEEE*

Abstract—Voice interface has been a dominant User Interface (UI) channel in the popular smart home environment. Although Voice Control System (VCS) brings users conveniences, it is extremely vulnerable to spoofing attacks (e.g., hidden/inaudible command attack) due to its broadcast nature. In this study, to thwart spoofing attacks, we propose WSVA, a device-free voice liveness detection system based on the prevalent wireless signals generated by IoT devices without requiring user to carry any additional sensor or device. The basic insight of WSVA to distinguish the authentic voice command from a spoofed one is checking the consistency between the voice signal and its corresponding mouth motions, which can be captured by wireless signals. To achieve this goal, WSVA builds a theoretical model to describe the correlations among the wireless signal changes, the mouth motions, and the syllables in the voice command. Then, WSVA selects appropriate features from both voice and wireless signals, and calculates the consistency between these two types of signals to determine whether the VCS is suffering from the spoofing attack. To demonstrate the feasibility of WSVA, we conduct a case study on Samsung SmartThings platform and include WSVA as a new application, which is expected to significantly enhance the security of the existing VCS. We evaluate WSVA with various voice commands in different scenarios. Experimental results demonstrate that WSVA achieves the overall 99 percent true accept rate with 1 percent false accept rate with a good scalability and low latency.

Index Terms—Liveness detection, voice control system, wireless side channels

1 Introduction

With the rapid growing of Internet of Things (IoT) technology, smart home or home automation is gaining an increasing popularity due to its great benefits of allowing the users to control their domestic appliances (e.g., lights, temperature controller, electronic switch, microwave, refrigerator) via a variety of user interfaces such as image sensing, wireless communication and voice controller commands. According to the list of Top 10 Consumer IoT Trends in 2017 published by Parks Associates, voice controller is predicted to become the primary user interface for the smart home and intelligent lifestyle [2]. Currently, the typical IoT voice control systems include Amazon Alexa [3], Samsung SmartThings [4], Google Home [5] and other interactive voice interfaces. According to Grand View Research's report, the market share of voice recognition was $9.12 billion in 2017 with an increasing rate of 17.2 percent during the forecast period [6]. Besides, the household penetration of Voice Control System (VCS) is expected to reach 47 percent in the USA by 2022 [7].

Although voice controller is regarded as the most promising user interface in smart home, it also introduces some emerging security concerns due to the inherent broadcast nature of voice channel, which makes it extremely vulnerable for spoofing attacks including *the replay attacks*, *the hidden command attacks* and *the inaudible command attacks*. The replay attack means that the adversary could fool the VCS using the pre-recorded voice samples of the legitimate user [8]. In the hidden command attack, a falsified speech signal mixed with noise samples is used as the input of VCS [9]. As an extreme case of spoofing attacks, recent studies [10], [11] show that it is feasible to inject some hidden or even inaudible voice commands which cannot be understood/heard by human but can still be interpreted by the VCS. This kind of spoofing attacks opens a concealed door for the adversary to query the user's sensitive information from VCS, or force the smart devices to perform misbehaving behaviors (e.g., unlocking the door when user leaves home), which poses a serious security threat to the smart home systems.

Existing solutions to defend against the spoofing attacks mainly fall into two categories: *voice password based access control* and *two-factor based liveness detection*. In the password based access control, the user is required to speak a special password before inputting the voice controller commands [12]. However, speaking a password is either inconvenient for user or vulnerable to eavesdropping attack. On the other hand, the two-factor based liveness detection exploits the information (e.g., image/video collected by camera [13], magnetic filed emitted from loudspeaker [14], time-difference-of-arrival changes from different microphones of smartphone [15], acceleration data of user's wearable devices [16] and the

- *Y. Meng, H. Zhu, and J. Li are with Shanghai Key Laboratory of Scalable Computing and Systems, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China. E-mail: {yan_meng, ricardolee}@sjtu.edu.cn, zhu-hj@cs.sjtu.edu.cn.*
- *J. Li is with the School of Computer Science, Guangzhou University, Guangzhou 510006, China. E-mail: jinli71@gmail.com.*
- *Y. Liu is with the Department of Computer Science and Engineering, University of South Florida, Tempa, FL 02125 USA. E-mail: yliu@cse.usf.edu.*

Manuscript received 4 Apr. 2019; revised 21 Nov. 2019; accepted 6 Feb. 2020. Date of publication 13 Feb. 2020; date of current version 11 Nov. 2021. (Corresponding author: Haojin Zhu.)
Digital Object Identifier no. 10.1109/TDSC.2020.2973620

Doppler shift of ultrasonic caused by user's mouth motion [17]) that are closely correlated with the operations of VCS as the user's liveness features to differentiate between the voice samples generated by legitimate user and adversary. However, the existing two-factor based liveness detection schemes require the user to either carry specialized sensing devices or perform specific actions to collect the liveness information, thus their practicalities are limited. More seriously, some of these schemes pose unacceptable privacy risks, since the user's daily behaviors may be leaked from the collected information (e.g., image or video data in [13]).

In this study, we present WSVA, a wireless signal based voice authentication system to thwart the spoofing attacks aiming at VCS. Unlike prior liveness detection schemes, WSVA is a device-free system without requiring the user to carry any additional device or sensor, and it leverages the prevalent wireless signals generated by Wi-Fi devices in IoT environment. WSVA is motivated from the following observations. First, inspired by the widely application of lip-reading technology, it is feasible to understand the speech by sensing the movements of the lips, face and tongue. In other words, voice command can be cross-checked by the user's mouth motions. Second, the prior researches show that the indoor object movement will disturb the multiple-path of wireless signals and can be reflected on Channel State Information (CSI) of Wi-Fi signals. Thus a variety of human activities can be identified by using the CSI based wireless sensing techniques. Therefore, it is natural to raise the following question: *is it feasible to build the correlation between the user's mouth motion and the environmental CSI change, and leverage this correlation to verify the liveness of voice commands received by VCS?*

The answer for the above questions is not straightforward. WSVA faces three major challenges: i) The impact of mouth motion on wireless signals is subtle. Although previous works utilize sophisticated methods such as MIMO beamforming or Frequency-Modulated Carrier Waves (FMCW) [18], [19] to improve the wireless sensing capability, they may not be suitable for our problem because the commercial IoT devices are resource-constrained and cannot implement these sophisticated wireless techniques. ii) According to our experimental result, only the jaw and tongue movements can be recognized by wireless signals while the vocal vibration which contributes a lot to voice signal could not be distinguished. Besides, prior works [19], [20] pointed that not all voice syllables can be recognized by lip-reading techniques. iii) To correlate the voice and CSI signals, how to select appropriate features from these two-dimensional signals still remains a big challenge.

This study shows how WSVA addresses the above challenges and achieves liveness detection. First, we build a new model to describe the correlation among the CSI changes, the mouth motions, and the syllables of the received voice signals. Then, WSVA proposes a novel signal processing method to filter the noises of collected voice and CSI signals, and to extract syllables and mouth motions within the voice command. Further, WSVA utilizes a novel method to extract both time-domain and frequency-domain of two types of signals and performs the liveness detection. We conduct experiments to evaluate the liveness detection performance of WSVA and give a case study on Samsung SmartThings platform to demonstrate its feasibility in IoT environment. The contributions of this work are summarized as follows:

- We present WSVA, a two-factor liveness detection system to thwart the various voice spoofing attacks aiming at VCS. By utilizing the existing wireless signals in IoT environment, WSVA shows its advantages of device-free, feasible deployment and privacy preservation.
- We study the correlation between voice samples and wireless signals. Specifically, we build a mapping model to correlate the syllables within voice command, the user mouth motions and their corresponding CSI change patterns.
- We devise the architecture and algorithms of WSVA. We exploit some effective technical mechanisms to process voice samples and CSI data, design novel algorithms to extract the features from these different types of signals, and propose the liveness decision algorithm.
- We design and implement a testbed on Samsung SmartThings platform to demonstrate the practicility of WSVA. We evaluate the impact of various factors on WSVA and our experimental results on 6 volunteers show that WSVA achieves 99 percent liveness detection accuracy with 1 percent false accept rate.

In this paper (which is an extended version of the work in [1]), we re-devise the signal processing and feature selection method of WSVA to improve the liveness detection performance. Besides, more factors are evaluated and discussed. We point out that this paper does not propose to use wireless signals for lip reading, since the existing works ([19], [21]) have shown that the lip reading accuracy is limited. Instead, this paper aims to utilize the consistency between voice and CSI signals to authenticate the voice commands.

The remainder of this paper is organized as follows. In Section 2, we introduce the preliminaries of this work. In Section 3, we introduce the research motivation by showing the consistency between voice and wireless signal changes during user's voice commanding. We elaborate the detailed design of WSVA in Section 4, which is followed by evaluation, discussion and related work in Sections 5, 6 and 7 respectively. Finally, we conclude this paper in Section 8.

2 Preliminaries

2.1 Attack Model

In this study, we consider the spoofing attack, which is defined as that the adversary tries to fool the VCS by injecting some pre-collected or forged voice commands as illustrated in Fig. 1. The existing studies show that there are three major types of spoofing attacks.

- Replay attack. The adversary can deploy an audio recorder to obtain the authentic user's voice samples, and then utilize a loudspeaker to play the voice commands sythenized from pre-collected voice to spoof the VCS [8].
- Hidden voice attack. Most of the VCSs leverage the Mel-frequency cepstral coefficient (MFCC) extracted

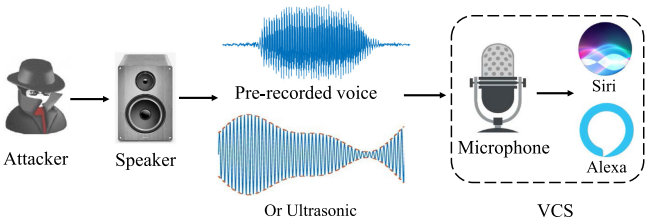


Fig. 1. Illustrations of the attack to VCS.

from human voice to perform speech recognition. Thus the adversary can generate voice commands which are heard as noise by human ears but contain the user's MFCC features to spoof the VCS [9].

- Inaudible attack. Recent studies show that many microphones have drawbacks on their system frequency responses. The adversary thus can utilize ultrasonic signals to synthesize voice commands which can not be heard by human to spoof the VCS [10], [11].

Without loss of the generality, in the remainder of this paper, we use spoofing attacks to represent the above-mentioned three kinds of attacks. Our proposed defense scheme is based on the fact that, in the spoofing attacks, the fake voice commands are generated by the machine rather than the human, which means that there are no corresponding mouth motions for these voice commands. This inconsistency can be leveraged for performing liveness detection. Note that, there exists another attack type—insider attack, which means the adversary can break into the home and impersonate a real user to inject fake voice command. However, this attack model has a very strong assumption and is less practical in smart home environment. Therefore, this attack type is not considered in our proposed liveness detection scheme and we will discuss its defend strategy in Section 6.

2.2 Channel State Information

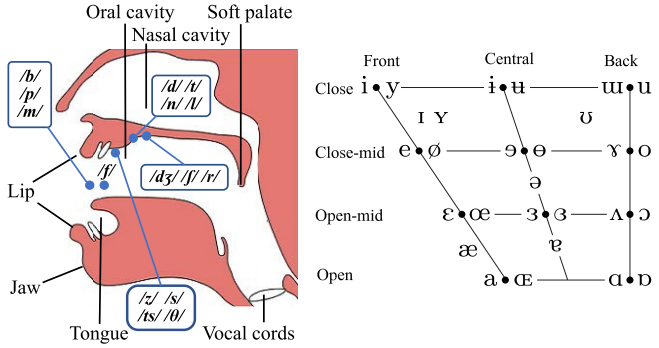
In this paper, we consider the Wi-Fi wireless communication protocol which is widely applied by many IoT devices (e.g., smart camera and smart alarm) [22]. Wi-Fi standards like IEEE 802.11n/ac support Orthogonal Frequency Division Multiplexing (OFDM), which is designed to significantly improve the channel capacity of the wireless system [23]. In a wireless communication system with N_{TX} transmitter antennas, N_{RX} receiver antennas and N_s OFDM subcarriers, $N_{TX} \times N_{RX} \times N_s$ subcarriers will be utilized to transmit signal at the same time.

CSI characterizes Channel Frequency Response (CFR) H in different subcarriers. In this paper, we only consider the system with only single antenna pair, and thus CSI data extracted from a packet could be represented by N_s dimension vector. And for the ith subcarrier, CSI value H_i can be defined as

$$H_i = |H_i|e^{j\angle H_i} = \alpha e^{-j2\pi f\tau}, \tag{1}$$

where α is the signal magnitude attenuation, f is the frequency and τ is the time-of-light. Given the length of signal propagation path d, the signal wavelength λ and the speed-of-light c, τ can be calculated as $\tau = d/c$ and Eqn. (1) can be rewritten as

$$H_i = \alpha e^{-j2\pi c\tau/\lambda} = \alpha e^{-j2\pi d/\lambda}. \tag{2}$$



(a) Vocal organs and consonant pronunciation [26].

(b) Vowel pronunciation.

Fig. 2. Articulatory gestures for voice pronunciation.

According to Eqns. (1) and (2), when the user speaks a voice command, the movements of the lips and the jaw will change the d and α of the wireless signal. These constructive and destructive interference of several multi-path signals will be reflected by a unique pattern in the time-series of CSI values, which can be related to the presence of the legitimate voice command. In this study, CSI extraction is quite easy: we can deploy Universal Software Radio Peripheral (USRP) [24] and COTS device (e.g., Intel 5300 NIC [25]) to extract CSI with all subcarrier values and 30 subcarrier values respectively.

2.3 Articulatory Gesture

It is well known that the articulation is related to human organs (e.g., vocal cords, tongue, lips, jaw), as shown in Fig. 2a. The voice differences depend on the motions of organs, which could affect the vibration frequency of the air (i.e., the timbre). According to the air vibration position, the procedure of voice generation can be divided into the following three stages:

i) Voice generation procedure starts when the air is sent out from the thorax. The air passes through the vocal cords comprising of cartilages and muscles, whose different shapes and positions have a significant effect on the air propagation. ii) The air arrives at the soft palate after passing through the pharynx. The soft palate controls the direction and speed of the airflow and decides whether it could enter into nasal cavity. iii) The voice wave is about to leave the mouth when the air arrives at the oral cavity, after which the voice is spread in the air. In this period, the user can produce different phonemes with different motions of tongue, lips and jaw, which is known as articulatory gesture. According the the International Phonetic Alphabet [27], as shown in Fig. 2, the users pronounce different phonemes with different mouth shapes. For instance, as shown in Fig. 2b, the position of the jaw can be half way opened and fully opened when the user pronounce /e/ and /a/ respectively.

3 Motivation

In this section, we elaborate the rationale behind WSVA by answering the following questions: first, do the mouth motions really have the correlation with the change of Wi-Fi signals? Second, how can we model this correlation between

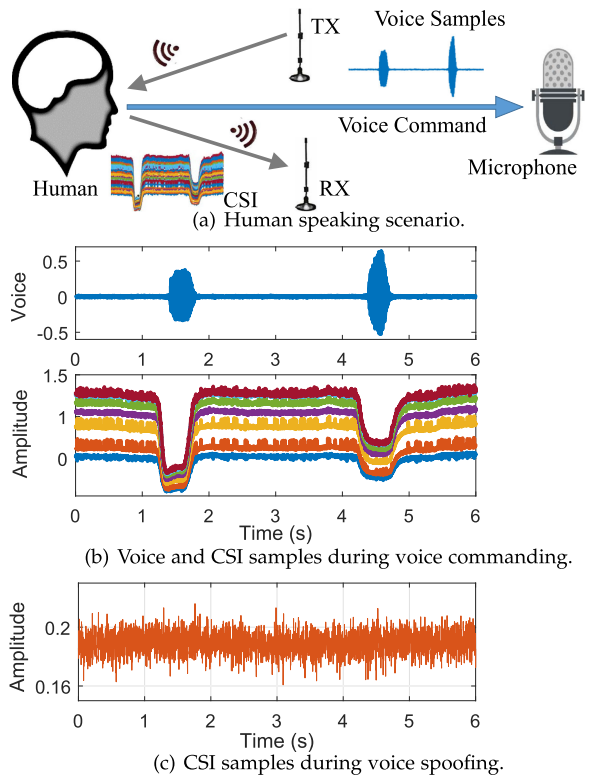


Fig. 3. Illustrations of the basic idea of WSVA.

the mouth motions and the CSI vibration? We answer these two questions via performing a series of experiments.

3.1 The Influence of Mouth Motion on CSI

Fig. 3a demonstrates the typical scenario of human voice commanding in VCS environment such as SmartThings or Amazon Alexa platform. When a user interacts with VCS, WSVA exploits a pair of antennas of the IoT devices in the proximity to collect the CSI data from Wi-Fi packets, and leverage a microphone to record the voice samples simultaneously. Generally speaking, since CSI reflects the environmental constructive and destructive interference on several multi-path signals, the change of multi-path propagation caused by the mouth motions during the voice speaking can generate a unique pattern in the time-series of CSI values. In this case, we investigate the influence of the mouth motions on the CSI, which can be regarded as a liveness pattern of the user. As shown in Fig. 3b, the dramatic fluctuations of CSI waveforms happen with the occurrence of human voice command. However, as shown in Fig. 3b, if an adversary launches the spoofing attack described in Section 2.1, in which the fake voice command is injected without any corresponding mouth motion, the attacks can be easily detected due to the lack of the corresponding changes in CSI data. Therefore, our experimental results validate our intuition that it is feasible to leverage the consistency of fluctuations between voice samples and CSI streams to detect the spoofing attacks.

3.2 Modeling the Correlation Among CSI Vibrations, Voice Syllables and User Mouth Motions

The previous works have demonstrated that human movements can be sensed via wireless signals [28], [29], [30], [31].

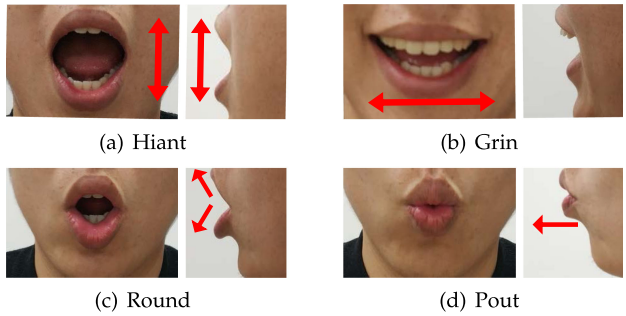


Fig. 4. Four types of mouth motion shapes.

However, in IoT environment, achieving very precise speech recognition is less possible since it may beyond the sensing capability of Wi-Fi signal. As shown in Eqn. (2), the sensing capability of wireless signal depends on the wavelength of the signals. In practice, the Wi-Fi signal (e.g., 12.5 cm wavelength for 2.4 GHz) based sensing mechanisms cannot accurately capture the tiny motion of human mouth. To make matters worse, in addition to the motion of the tongue, lips and jaw, Wi-Fi can hardly recognize the impact of other vocal organs. According to the study of Dodd *et al.* [32], only 40 percent words in English can be recognized by only considering mouth motions.

Although it's not feasible to achieve accurate lip reading via Wi-Fi signals, WSVA is devised to authenticate the voice commands by checking the consistency between voice and CSI signals rather than accurately identifying each syllable. Therefore, in this paper, by analyzing the International Phonetic Alphabet, we classify the mouth motions into four categories, including hiant, grin, round, and pout, which correspond to Figs. 4a, 4b, 4c and 4d, respectively. With the exception of a few syllables (e.g., /ə/) with non-significant mouth motions, most phonetic syllables can be categorized into one of these types. As shown in Table 1, the hiant, the motion of opening the mouth largely, can pronounce the phonemes like /ɑ:/ and /æ/, which can be heard in words "*bar*" and "*cat*". The grin, the motion of grinning human mouth like Fig. 4b, can pronounce the phonemes like /e/ and /ei/, which can be heard in "*A*" and "*base*". The round, rounding lips at ease, can generate the phonemes like /ɔ:/, which can be heard in "*lot*" and "*saw*". Finally, the pout, the motion pouting the lips, can send out the phonemes like /u:/, which can be heard in words "*root*" and "*shoe*". After such a classification, different types of mouth motions can be correlated with different CSI features according to relevant voice syllables, as mentioned in the following sections.

4 System Design

4.1 System Overview

The basic strategy of WSVA to detect if a voice command is an authentic one is checking the consistency between the voice samples and its corresponding CSI data introduced by mouth motions. The CSI data can be collected via a specialized device (e.g., USRP) or the COTS device. In the context of s mart home, with the prevalent of IoT platforms such as Samsung SmartThings, which controls the smart devices with wireless signals, it is technically feasible to

TABLE 1
Four Categories of Mouth Motions and Their Corresponding Syllables

Mouth motion	Syllables	Words
Hiant	/a:/ /æ/ /ai/	bar, cat
Grin	/e/ /ei/	A, base
Round	/ɔ/ /ɔ:/	lot, saw
Pout	/u:/ /ʊ/	root, shoe
Non-significant	/ə/ /iə/	sir, here

take advantage of these existing wireless infrastructures to collect the voice samples and their corresponding CSI data simultaneously.

As shown in Fig. 5, WSVA consists of the following four modules. In *Data Collection Module*, when human voice is detected by the VCS, WSVA collects the voice samples and its corresponding CSI data. In *Data Cleansing and Preprocessing Module*, WSVA exploits wavelet based method to remove the noise in CSI, and segments the collected voice samples. *Feature Extraction Module* enables WSVA to select appropriate features from macro-level and mouth motion respectively. Finally, *Feature Matching Module* utilizes a classification mechanism to determine whether the received voice command is an authentic one or suffering from spoofing attacks.

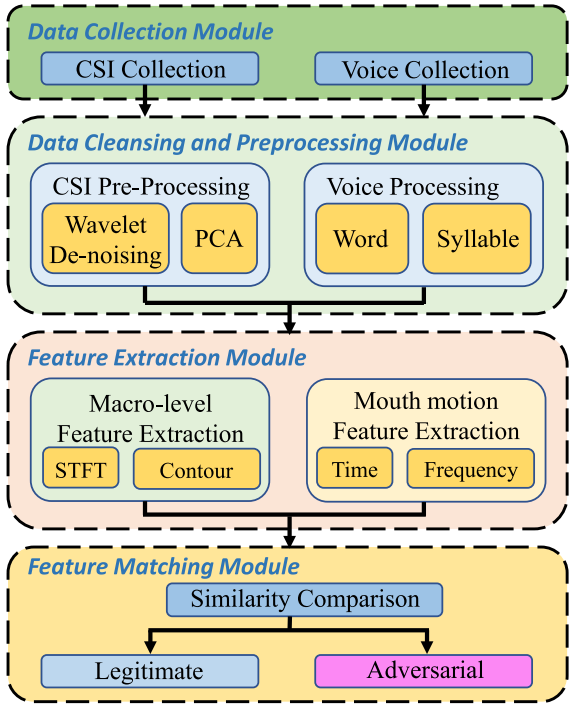


Fig. 5. Workflow of WSVA.

4.2 Voice Samples and CSI Data Collection

In this subsection, we introduce how to collect voice samples and the corresponding CSI data. For most of the VCSs (e.g., Google Now and Amazon Alexa), they require the user to speak a predefined magic word as a trigger. For instance, Apple iPhone needs "Hey, Siri" and Amazon Alexa needs "Alexa" to initialize their voice assistants. Only when the voice trigger is recognized by the VCS, WSVA will be activated and start to collect voice samples V and CSI data H by utilizing microphone and antenna pair respectively. To collect CSI, the antennas can be equipped by the different devices, or incorporated in the same IoT device. One of these antennas acts as a transmitter to continuously send wireless packets (e.g., broadcast packets) and the other receives packets and extracts CSI data from the preamble sequences of these wireless packets.

4.3 Data Cleansing and Preprocessing

4.3.1 CSI Denoising

The original collected CSI data contain a large amount of background noises which should be removed for future liveness detection. In this paper, for each subcarrier data from collected CSI data H, WSVA leverages wavelet-based denoising to eliminate high frequency noises as the following three steps.

Discrete Wavelet Transform (DWT). Generally speaking, let $H_i[n]$ be the ith CSI subcarrier $H(:,i)$, and this one-dimension discrete signal can be expressed in terms of the wavelet function by the following equation:

$$H_i[n] = \frac{1}{\sqrt{L}}\sum_{k} W_{\phi}[j_0,k]\phi_{j_0,k}[n] + \frac{1}{\sqrt{L}}\sum_{j=j_0}^{\infty}\sum_{k} W_{\psi}[j,k]\psi_{j,k}[n], \quad (3)$$

where L represents the length of $H_i[n]$. The functions $\phi_{j_0,k}[n]$ refer to scaling functions and the corresponding coefficients $W_{\phi}[j_0,k]$ refer to the approximation coefficients. Similarly, functions $\psi_{j,k}[n]$ refer to wavelet functions and coefficients $W_{\psi}[j,k]$ refer to detail coefficients. During the decomposition process, the origin signal is first divided into the approximation coefficients which depict the trend of origin signal and detail coefficients which retain the small scale characteristics. Then the approximation coefficients are iteratively divided into the approximation and detail coefficients of next level.

Threshold Selection. After recursive DWT decomposition, the raw signal is broken into detail coefficients W_{ψ} (high-frequency) and approximation coefficients W_{ϕ} (low-frequency) at different frequency levels. Then, the threshold is applied to the detail coefficients to remove their noisy components and obtain new coefficients W'_{ψ}. The threshold selection is important because a small threshold will remain the noisy components while a large threshold will lose the major information of signals. In this study, we empirically choose an adaptive minimax threshold based on the experimental results.

Wavelet Reconstruction. After the above two steps, we reconstruct the signal to achieve noise removal by combining the coefficients of the last approximation level W_{ϕ} with all thresholded details W'_{ψ}. In this study, we choose Daubechies D4 wavelet [33] and perform 4-level DWT decomposition in wavelet denoising. As shown in Fig. 6c, after wavelet-based denoising, most of the burst noises in H_i can be removed.

4.3.2 Voice Samples Segmentation

After performing wavelet-based denoising, it is observed that the CSI waveform shows a strong correlation with

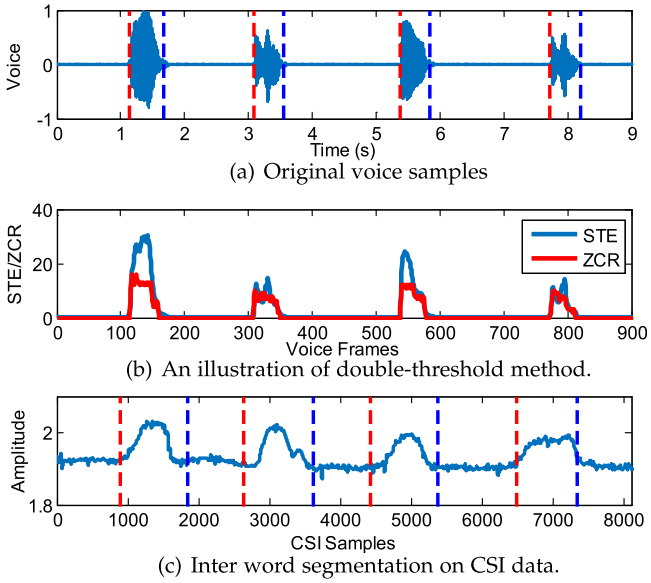


Fig. 6. An example of word level segmentation.

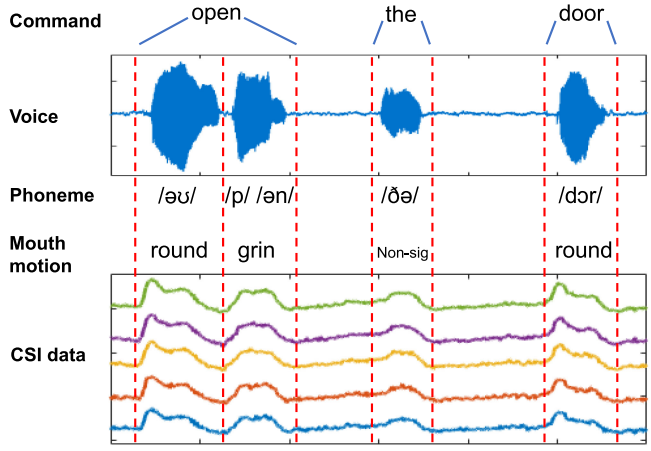


Fig. 7. An example of the mouth motion detection.

mouth motions. In order to verify the consistency between CSI and voice samples, it is critical to detect the start and end points of each mouth motion samples. However, since the CSI waveform has many break points, applying the burst detection method [34] directly on CSI data does not work well in this case. Therefore, we first perform word level segmentation and phoneme level segmentation on voice samples, and then extract the corresponding CSI data according to the timestamps.

Word Level Segmentation. When the user speaks a command, there is a short interval (e.g., 200 ms) between the pronunciation of two successive words. Therefore, the interval between two word samples can be utilized to segment voice command into different word samples. WSVA exploits double-threshold detection method in this paper. Specifically, WSVA splits the voice samples V into frames of N_v points length, with shifting N_s points each time. In this study, N_v and N_s are set to 512 and 256 respectively. For totally N frames, WSVA calculates their short term energy $STE[n]$ and zero-crossing rate $ZCR[n]$, and selects two adaptive thresholds for $STE[n]$ and $ZCR[n]$ to detect the start and end points $s_{v,i}$ and $e_{v,i}$ of the ith word W_i. Then, according to the timestamps, we can also divide the CSI data into several word waveforms. Fig. 6 illustrates the proceeding of inter word segmentation. For the kth CSI subcarrier $H(:,k)$, its corresponding ith word's CSI waveform W_i can be represented as follows.

$$H_{W,i} = H(s_{c,i} : e_{c,i}, k), \tag{4}$$

where $s_{c,i}$ and $e_{c,i}$ are the start and end CSI indexes of the ith word W_i which are converted from the timestamps $s_{v,i}$ and $e_{v,i}$ on voice samples. Note that, s_i and e_i are extended on both sides by 200 CSI indexes respectively, due to the fact that the CSI change introduced by the mouth motion can be occured a little bit earlier or later than the speech can be heard.

Phoneme Level Segmentation and Mouth Motion Inference. For a specific word, pronouncing it may involve more than one mouth motion. For instance, speaking the word "open" needs the mouth motions of "round" and "grin". Besides, as mentioned in Section 3.2, the correlation between different categories of mouth motion and CSI vibration types is a key factor which can be leveraged in liveness detection. Therefore, the next step of WSVA is dividing the given CSI word waveforms into multiple CSI mouth motion waveforms, and then calculates the similarity between the collected CSI mouth motion waveforms and pre-trained CSI motions data.

Similar to word level segmentation, WSVA processes the voice samples of the user and infers the start and end points of each mouth motion. In particular, WSVA first utilizes automatic speech recognition to identify each word of a voice command. The state-of-the-art system DeepSpeech [35] is adopted to perform such a task automatically. After identifying existing words, WSVA then utilizes Munich Automatic Segmentation System (MAUS), a widely adopted phonetic segmentation system [36]. MAUS is based on the Hidden Markov Model method, and it can label the phonemes of voice signals by analyzing the sound file and text description of the voice. Specifically, based on standard pronunciation model, the identified text will be transformed into expected pronunciation. Then, a probabilistic graph will be generated by combining the canonical pronunciation with millions of different accents, which contains all possible phoneme combinations and the corresponding probabilities. MAUS finally adopts Hidden Markov Model to perform path search and find the combination of phonetic units with the highest probability.

After combining phonemes into syllables, and inferring the mouth motions according to International Phonetic Alphabet, we can obtain the segmented and labeled mouth motions of the inputting voice command. WSVA matches the timestamps of each segmented motion to the CSI samples to extract CSI mouth motion waveforms as the method defined in Eqn. (4). One example is illustrated in Fig. 7, which segments a voice command ("Open the door") into several phonemes and extracts the CSI mouth motion waveforms. It is worth mentioning that since the number of voice commands commonly used in VCS is limited, the performance of speech recognition can be improved according to pre-defined common commands set. In addition, WSVA also utilizes inter word segmentation result to improve the

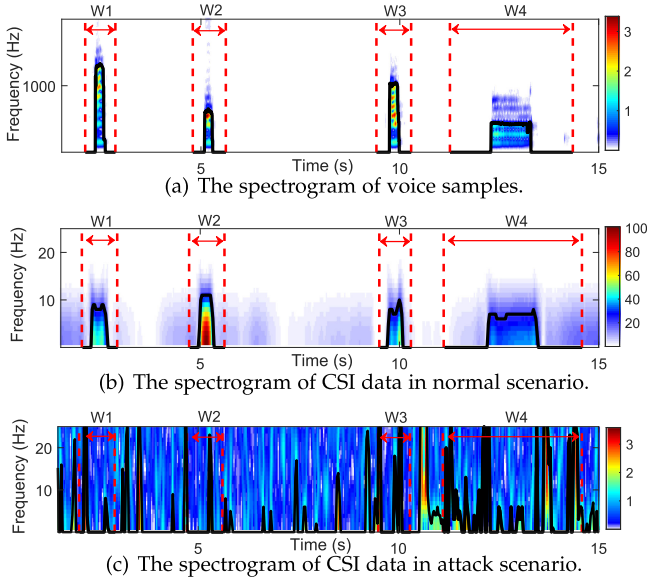


Fig. 8. Illustration of the macro-level feature extraction.

phoneme segmentation performance. After these steps, we obtain the start and end points of all N_m mouth motions $\mathbf{M} = \{M_1, M_2, \ldots, M_{N_m}\}$ in a voice signal, and then extract the CSI data H_{M_i} fot the ith mouth motion M_i.

4.4 Feature Extraction

After data cleansing and pre-processing on CSI and voice samples, WSVA selects the appropriate features to characterize the consistency between these two types of signals. As mentioned in Section 3.1, in the macro level, it is observed that CSI variation occurs along with the human pronunciation. Besides, the CSI data of different mouth motion types show different features, which is another criterion to describe the consistency. Therefore, WSVA extracts features from both macro-level and mouth motion level to determine whether the voice command and the mouth motion are consistent.

4.4.1 Macro-Level Feature Extraction

In particular, after performing wavelet denoising, CSI data still compose of N_s subcarriers (e.g., N_s = 52 in this study). To remove the DC components in all subcarriers and extract the strongest correlation component with mouth motions, WSVA adopts PCA to extract the first principle component H_{PCA} of all CSI subcarriers H. Then, WSVA adopts Short Time Fourier Transform (STFT) on both CSI data H_{PCA} and voice samples V to obtain their two-dimensional frequency spectrograms. Figs. 8a and 8b show the frequency shifts on the voice samples and the corresponding CSI data spectrograms in non-attack scenarios. It is observed that in a non-attack scenario, the contours (marked by black lines) of CSI and voice samples have similar variation trends. However, as shown in Figs. 8a and 8c, in the spoofing scenario, since the voice samples are injected by the adversary without any user's mouth motion, the CSI contour is disordered and not in consistent with that of voice samples. Therefore, to measure the consistency between voice and CSI samples, an intuitive solution is to calculate the similarity between the spectrogram contours of these signals.

However, directly calculating the similarity between the spectrogram contours of V and H_{PCA} is inappropriate, since the frequency shifts of these signals are affected by different factors (i.e., voice tunes on voice and mouth movements on CSI) which are not necessarily related. Instead, for the N_w words $\mathbf{W} = \{W_1, W_2, \ldots, W_{N_w}\}$ in the command, WSVA calculates the similarity between contours of voice and CSI signals for each word W_i, and then combines these similarities to obtain the macro-level similarity S_{Macro} . For the ith word W_i, to calculate its similarity, we first extract the CSI and voice samples H_{W_i} and V_{W_i} which are represented as

$$H_{W_i} = H_{PCA}(s_{c,i} - L_{c,i} : e_{c,i} + L_{c,i}), \tag{5}$$

$$V_{W_i} = V(s_{v,i} - L_{v,i} : e_{v,i} + L_{v,i}), \tag{6}$$

where $s_{v,i}$, $s_{c,i}$, $e_{v,i}$, $e_{c,i}$ are the begin and ending indexes of ith word W_i on voice and CSI samples respectively. $L_{v,i}$ and $L_{c,i}$ are the spans of the voice and CSI samples of W_i, in which $L_{v,i} = e_{v,i} - s_{v,i} + 1$, and $L_{c,i} = e_{c,i} - s_{c,i} + 1$. Note that, instead of directly using the Eqn. (4), we extend both sides of H_{W_i} and V_{W_i} to obtain more details about the W_i.

Then, we extract the contour C_{CSI,W_i} from the frequency spectrogram of ith word's CSI data. First, we resize the CSI spectrogram with frequency from 0 to 30 Hz into a m-by-n matrix $M_{CSI}(j,k)$ and normalize the $M_{CSI}(j,k)$ to a range between 0 and 1. Note that, in $M_{CSI}(j,k)$, each column represents the normalized frequency shifts during the jth time slide. Then, we choose a pre-defined $threshold$ and get the contour $C_{CSI,W_i}(j)$, where $j = 1 \ldots n$. $C_{CSI,W_i}(j)$ is the maximum value k which satisfies that $M_{CSI}(j,k) \geq threshold$. The process of calculating contours C_{V,W_i} for the voice spectrogram is similar to calculating C_{CSI,W_i}. Besides, as mentioned in Section 4.3.2, we can set the value $C_{V,W_i}(j)$ to 0 to eliminate the interference of background noise, if the jth time slide is not within the word segments.

After obtaining C_{CSI,W_i} and C_{V,W_i} for W_i, we measure the correlation between these two contours by adopting Pearson correlation coefficient [37], which is defined as $Corr(W_i)$. $Corr(W_i)$ ranges from 0 to +1, where a higher value represents a higher level of similarity. To calculate $Corr(W_i)$, we first re-sample C_{CSI,W_i} and C_{V,W_i} into the sample length, and $Corr(W_i)$ can be represented as

$$Corr(W_i) = \left| \frac{\sum_{i=1}^{n} (C_{CSI,W_i}(i) - \overline{C_{CSI,W_i}})(C_{V,W_i}(i) - \overline{C_{V,W_i}})}{(n-1)\delta_{CSI}\delta_V} \right|, \tag{7}$$

where n is the length of re-sampled sequences C_{CSI,W_i} and C_{V,W_i}, δ_{CSI} and δ_{Voice} are the sample standard deviations of C_{CSI,W_i} and C_{V,W_i}, respectively. After calculating the similarity $Corr(W_i)$ for ith word W_i, for total words $\mathbf{W} = \{W_1, W_2, \ldots, W_{N_w}\}$ we could calculate the macro-level similarity S_{Macro} as follows:

$$S_{Macro} = \frac{\sum_{i=1}^{N_w} Corr(W_i)}{N_w}. \tag{8}$$

4.4.2 Mouth Motion Feature Extraction

In the previous section, we have discussed how to obtain the macro-level feature S_{Macro} from CSI data and voice

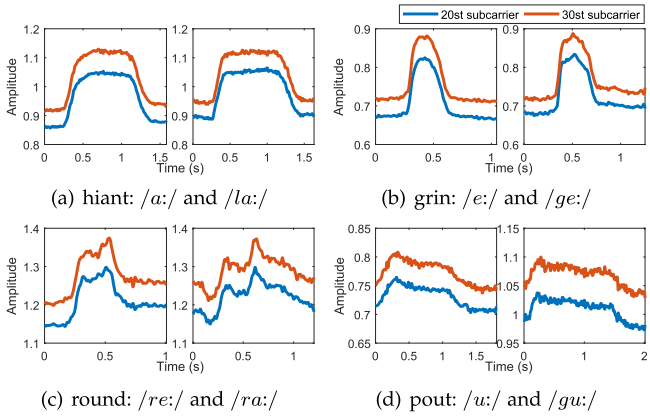


Fig. 9. Time domain of four mouth motion types.

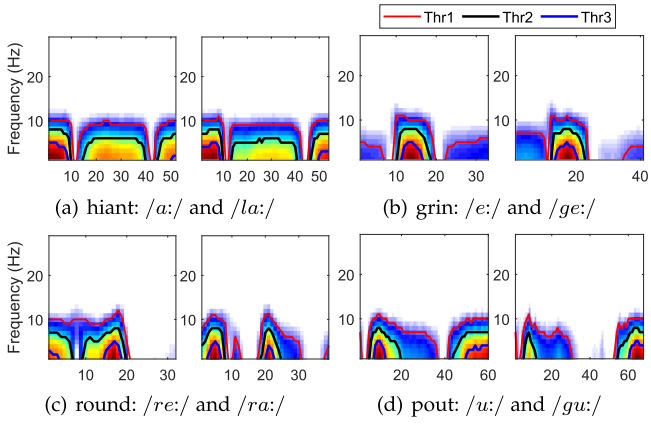


Fig. 10. Frequency domain of four mouth motion types.

samples during the voice command pronunciation. However, it may be not enough to perform liveness detection only relying on S_{Macro}. For example, the dramatic change of environment may generate the drastic vibrations of CSI data, which lead to a deviated contour C_{CSI,W_i} and a higher similarity $Corr(W_i)$ for detected word W_i. Therefore, to further improve detection performance, WSVA will extract the mouth motion level features from both time and frequency domain perspective in this subsection.

Time Domain Feature Extraction. Fig. 9 shows the amplitudes of CSI syllable data belonging to four mouth motion categories. It is observed that the CSI waveforms belonging to the same mouth motion category have the similar shapes. For instance, in Fig. 9a, the waveforms of syllable /a:/ and /la:/ which belong to the motion "hint" have the similar waveform shapes and amplitude vibrations. And it is also discovered that the ranges of CSI amplitudes from different mouth motion categories are quite different. For instance, as shown in Figs. 9a and 9d, the CSI amplitude ranges of syllables /a:/ and /la:/ are much larger than syllables /u:/ and /gu:/. Thus we can extract the ranges from the CSI waveforms as their time domain features. For a given CSI mouth motion M and its CSI data H_M, the CSI time domain feature $Range(M)$ can be calculated as

$$Range(M) = \sum_{i=1}^{N_s} \frac{Max(H_{M,i}) - Min(H_{M,i})}{N_s \times Mean(H_{M,i})}, \tag{9}$$

where N_s represents the number of CSI subcarriers and $H_{M,i}$ represents the ith subcarrier of H_M. Note that, the PCA processed data H_{PCA} is not utilized in this scenario, since the PCA process will distort the signal's range.

Frequency Domain Feature Extraction. In time domain feature extraction part, the CSI waveform shape changes over time. However, the experimental results show that the frequency shifts of CSI data caused by mouth movement have a relatively stable pattern. Fig. 10 shows the STFT spectrograms of syllables which are displayed in Fig. 9, and the contours of frequency spectrograms with three different thrresholds (i.e., $Thr1 = 0.2$, $Thr2 = 0.5$, $Thr3 = 0.8$) are marked as solid lines. It is observed that the CSI syllable data from different mouth motion categories have quite different contours. For instance, the contours of syllables /a:/ and /la:/ are more widely than that of /e:/ and /ge:/. Therefore, for a given mouth motion M and its CSI data H_M, we can utilize the contours of CSI spectrogram as the frequency features. In this study, we choose these three thresholds $\{Thr1, Thr2, Thr3\}$ and obtain the corresponding contours $C_{M,Thr1}, C_{M,Thr2}$ and $C_{M,Thr3}$. After that, each contour C_{H_M,Thr_i} is compressed to N_c points (e.g., N_c is 5 in this study), and WSVA merges these compressed contours into the feature vector F_M with $3 \times N_c$ elements.

4.5 Similarity Comparison and Liveness Identification

Before performing liveness detection, it is reasonable to assume that the user can provide totally J-by-N pre-collected CSI mouth motion data H_{Pre}, which contain J syllable categories (i.e., four motion categories proposed in Section 3.2) and each category contains N motions' CSI data $H_{Pre}(i,j)$, where $j = 1, 2, \ldots, N$. Then, for a given voice command input containing N_M mouth motions $\mathbf{M} = \{M_1, M_2, .., M_{N_M}\}$ which belong to four motion categories, WSVA processes the voice samples V and CSI data H using the above mentioned modules. After that, WSVA obtains the macro-level similarity S_{Macro} and the mouth motion feature $Range(M_i)$ and F_{M_i} for each motion M_i.

Mouth Motion Feature Combination. For a given motion M, WSVA first calculates the time domain range difference $SMR_{Time}(i)$ between its feature $Range(H_M)$ and pre-collected ith mouth motion category's features, which can be calculated as

$$SMR_{Time}(i) = \sum_{j=1}^{N} \left| \frac{Range(H_M) - Range(H_{Pre}(i,j))}{N} \right|. \tag{10}$$

Since the corresponding motion catrgory of M can be calculated from the voice processing module as described in Section 4.3.2, we can calculate the time domain similarity score between M and its corresponding motion type as follow:

$$S_{Time} = Min\left\{1 - \frac{SMR_{Time}(type)}{Max(SMR_{Time})} + \alpha, 1\right\}, \tag{11}$$

where $type$ represents the motion type of M, which ranges from 1 to J. The resulted S_{Time} ranges from 0 to 1, and the value closer to 1 indicates a high level of similarity. Note that, the function of adjustment factor α is to prevent S_{Time} from being zero, and we empirically set α to 0.1 in this study.

Then, WSVA compares the similarity between the frequency domain feature F_M of M and the J-by-N features $F_{Pre}(i,j)$ which are extracted from pre-collected motion data H_{Pre}. Different from the previous work [1] which utilizes Dynamic Time Warping with $O(N^2)$ time complexity, in this paper, to speed up the computation, WSVA exploits neural network based solution to characterize the similarity. WSVA utilizes the pre-collected CSI data H_{Pre} to train a forward propagation neural network net with 20 neurons in the hidden layer. For a given CSI data $H_{Pre}(i,j)$, the input for net is frequency domain feature $F_{Pre}(i,j)$ and the training label is the mouth category number. In this study, we set the category numbers for motion "hiant", "grin", "round" and "pout" are 1, 2, 3 and 4 respectively. After training, in the ideal case, net could map a specific motion feature F_M to its corresponding motion type. The similarity between M and the ith mouth motion category can be calculated as

$$SMR_{Freq}(i) = |net(F_M) - label(i)|, \tag{12}$$

where $label(i)$ is the label of ith motion category, and $net(F_M)$ is the prediction of net.

Similar to Eqn. (11), WSVA calculates the similarity score between F_M and its corresponding motion category $type$ as

$$S_{Freq} = Min\{1 - \frac{SMR_{Freq}(type)}{Max(SMR_{Freq})} + \alpha, 1\}, \tag{13}$$

where the adjustment factor α is set to 0.1, and the result S_{Freq} closer to 1 indicates a high level of similarity.

After obtaining the time domain similarity score S_{Time} and the frequency domain similarity score S_{Freq} of a given mouth motion M, we can calculate the combination mouth motion level similarity score S_{Motion} as

$$S_{Motion}(M) = S_{Time} \times S_{Freq}. \tag{14}$$

Liveness Detection. For a voice command which contains N_M mouth motions belonging to four motion categories, after performing mouth motion feature combination, we obtain its macro-level feature S_{Macro} and the mouth motion score $S_{Motion}(M_i)$ of each motion M_i, where $i = 1, 2, \ldots, N_M$. Then, we can calculate the finial decision score of the input, which is calculated as

$$Score = S_{Macro} \times \prod_{i=1}^{N_M} S_{Motion}(M_i). \tag{15}$$

We utilize threshold based mechanism to perform human liveness detection in this paper. For the given voice command input, if its $Score$ is larger than the pre-defined verification threshold, WSVA regards it as an authentic voice command. Otherwise, WSVA judges it as a fake command and refuses to execute it. In the next section, we will give a detailed experimental evaluation.

5 Performance Evaluation

In this section, we conduct a series of experiments to evaluate the performance of WSVA in different scenarios, and explore the implementation of WSVA in real-world IoT environment.

5.1 System Setup

Hardwares. WSVA consists of two hardwares: i) an Universal Software Radio Peripheral N210 device which connects two commercial Wi-Fi antennas, and ii) a microphone, responsible for collecting voice samples. In the experiment, the distance between antennas and human is 20 cm. When a user speaks a voice command, the USRP N210 collects CSI data at the rate of 1,000 packets/second in 2.4 GHz Wi-Fi frequency with the 1/2 BPSK modulation mechanism, and the microphone collects the voice samples simultaneously. We exploit USRP rather than COTS device (e.g., Intel 5300 NIC) to collect more stable CSI data, since some commercial devices change its power adaptively and result in unstable CSI measurements. However, USRP and COTS devices have the same wireelss functions in essential.

Data Collection. Our experiment totally recruits 6 volunteers. Before performing voice command, each volunteer was required to perform the four categories of mouth motions (i.e., the corresponding syllables) for 10 times as WSVA's pre-collected CSI profiles. Then, each volunteer performs voice commands and the adversary performs spoofing attacks using this volunteer's voice profiles. WSVA finally performs liveness detection by analyzing the collected CSI data and voice samples with the volunteer's mouth motion profiles.

Metrics. To assess the performance of WSVA, we choose the False Accept Rate (FAR) and the True Accept Rate (TAR) as evaluation metrics. TAR is the rate which WSVA detects the authentic user correctly, while FAR characterizes the rate which an attacker is wrongly accepted by the system and considered as an authentic user. Both FAR and TAR are influenced by varying the pre-defined verification threshold, and we show their relationship using Receiver Operating Characteristic (ROC) curve. In our experiment, we adjust the verification threshold value of WSVA to study more comprehensive results.

5.2 Thwarting Spoofing Attacks

In this subsection, we evaluate the effectiveness of WSVA to defend against spoofing attacks. First, for each volunteer, he/she is required to provide his/her pre-collected CSI profiles and speak 150 legitimate voice commands. After that, we perform spoofing attacks as described in Fig. 3 by using each volunteer's voice profiles for 750 times. We totally collect 5,400 voice commands, and in a given command, the numbers of mouth motions belonging to the four types as mentioned in Section 3.2 range from 4 to 8. The ROC curve of WSVA in detecting live users in non-attack scenarios and spoofing attack scenarios is depicted in Fig. 11. We can observe that with 1 percent FAR, the TAR is as high as 99.2 percent when WSVA exploits combined features S_{score}. Moreover, we find that the TAR relying on mouth motion feature is better than that relying on macro-level feature. More concretely, with 1 percent FAR, the TAR relying on mouth motion feature still keeps above 99 percent. However, the TAR relying on macro-level feature is reduced to 90.2 percent. The reason is that the macro-level features are more susceptible to the environment noise. After collecting voice and CSI data, the average time delay of performing each liveness detection is 0.26 seconds, which is acceptable in practice and is smaller than that in previous work (i.e.,

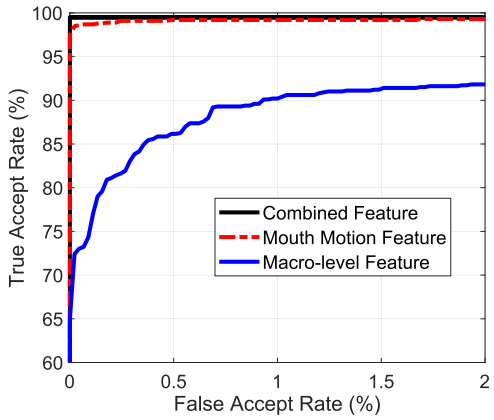


Fig. 11. Performance on thwarting spoofing attacks.

0.32 second in [1] with the same hardware condition). In summary, our experimental results well validate that WSVA is highly effective in defending against spoofing attack, while the macro-level feature and the mouth motion feature can complement each other to improve the detection performance.

5.3 Scale up to Multiple User's Scenario

In Section 5.2, for each user, WSVA performs liveness detection based on his/her pre-collected CSI profiles. However, in some smart home environments with multiple users, it is less likely to collect each user's mouth motion profiles. A more desirable design is to collect profiles from only one user but work for multiple users. In this section, we perform experiments to evaluate the scalability of WSVA. We first recruit a volunteer to provide WSVA with his/her mouth motion profiles which record his/her articulatory gesture. Then we recruit another two volunteers to perform voice commands for 300 times. After that, we also implement spoofing attacks for 600 times. Fig. 12 shows the evaluation result of WSVA, where WSVA achieves 97.6 percent TAR with 1 percent FAR, and 97.9 percent TAR with 2 percent FAR. Note that, the mouth motion feature based detection rate (i.e., 89.6 percent TAR with 2 percent FAR in Fig. 12) is less than that in Section 5.2. The reason is that the articulatory gesture of another volunteer is not the same as the user who provides the pre-collected CSI profiles. However, compared with spoofing attacks, WSVA can still achieve a high

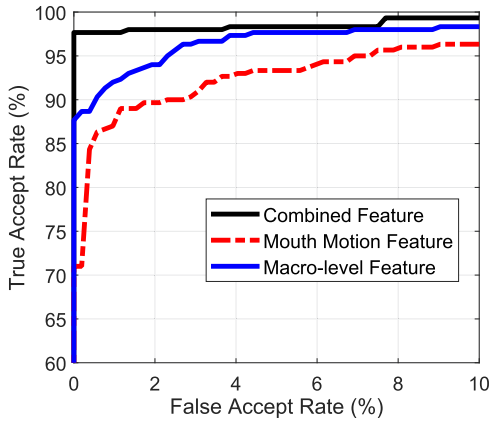


Fig. 12. Scaling up to multiple users.

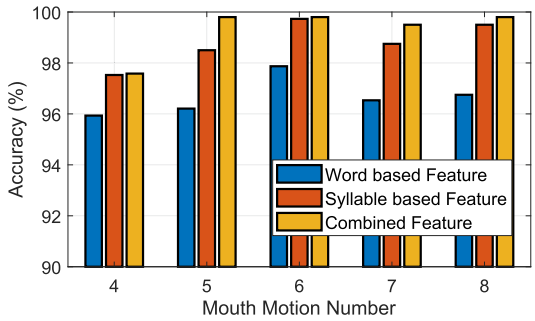


Fig. 13. The impact of syllable length.

detection accuracy, which demonstrates that it is also highly effective in multiple users scenario.

5.4 Impact of Mouth Motion Number

In this subsection, we investigate how different mouth motion numbers contained in the voice command affect the performance of WSVA. In our experiments, the motion numbers range from 4 to 8, and their corresponding accuracy are depicted in Fig. 13. The accuracy is the rate of successfully detecting authentic and spoofing commands among all commands under the 2 percent FAR. It is observed that with the increase of mouth motion number, the accuracy slightly rises from 97.5 to 99.8 percent. This result indicates that a higher number of mouth motion can reduce the impact of a single mouth motion misjudgment. In the experiment, the accuracy decreases slightly when the mouth motion number is greater than 6. This phenomenon is caused by unsatisfactory data quality during the process of data collection. Moreover, the accuracy exceeds 99 percent when mouth motion number is greater than 4, which means the features extracted from mouth motion by WSVA are accurate enough for liveness detection.

5.5 Impacts of Distance and LOS/NLOS

In above evaluations, the volunteer is located at the line of sight (LOS) places of antennas, and the distance between user's mouth and the receiver antenna is set to 20 cm. To evaluate the impact of distance on detecting voice spoofing, a volunteer is recruited to conduct experiment with distances of 50 cm, 100 cm, and 150 cm respectively. For each distance, the volunteer is required to provide CSI profiles and generate 150 voice commands, and then loudspeaker is deployed at the volunteer's location to perform spoofing attacks 300 times. As shown in Fig. 14, it is observed that the detection accuracy decreases when the distance becomes greater. By using the combined feature, WSVA achieves over 99 percent

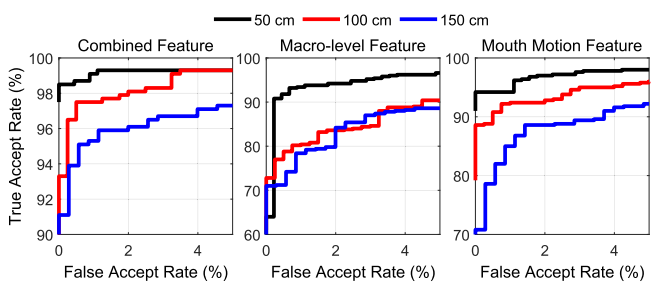
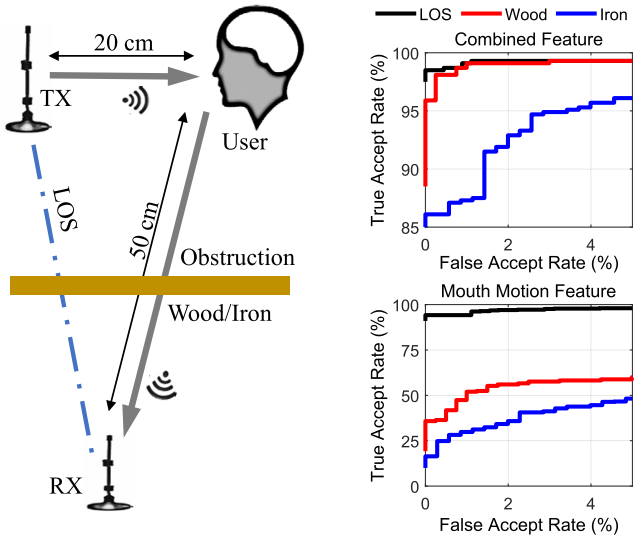


Fig. 14. The impact of distance on WSVA.



(a) The non-LOS scenarios.

(b) Detection results under different obstructions.

Fig. 15. Evaluations of Obstructions.

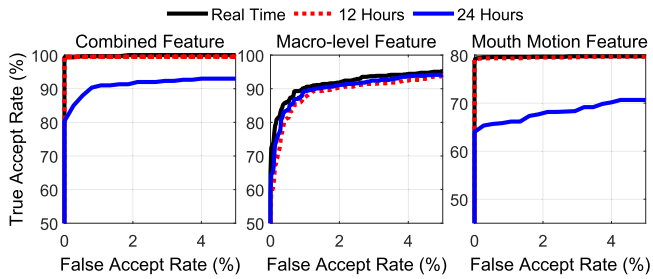


Fig. 16. The impact of time.

TAR with 2 percent FAR when the distance is 50 cm. However, the TAR is decreased to 98 and 96 percent when the distance is 100 cm and 150 cm respectively. Besides, The TAR under 2 percent FAR decreases dramatically when only utilizing macro-level feature (from 94 percent in 50 cm to 80 percent in 150 cm) and mouth motion feature (from 97 percent in 50 cm to 88 percent in 150 cm) individually. It means that when distance increases, the impact of mouth motion on multiple-path propagation of CSI becomes weaker and causes the degradation of WSVA's performance. However, when the distance is set to 1.5 m, WSVA could still achieve satisfactory accuracy (96 percent) using combined feature, which is acceptable in most cases.

To evaluate the performance of WSVA in the non-LOS scenarios, two additional experiments are conducted. As shown in Fig. 15a, the volunteer is required to stay out of the line of sight area. To further demonstrate this scenario, we consider a more extreme case in which we insert the obstruction board to separate the transmitter antennas from the receiver while the volunteer is at the same side of the transmitter. The dataset obtained with the distance of 50 cm as shown in Fig. 14 is chosen as the control group. The experimental results are shown in Fig. 15b. When WSVA utilizes combined feature, with 2 percent FAR, the TARs of WSVA under wood obstruction and control group are still over 99 percent. However, the TAR under iron obstruction is decreased to 92.7 percent. More specifically, when only exploiting mouth motion feature, the TARs under wood and iron obstructions are decreased to 56 and 36 percent. The results demonstrate that the obstruction in LOS could degrade the wireless sensing capability, especially only with the mouth motion feature extraction of WSVA. It is notable that WSVA's performance under iron obstruction is much weaker than that under wood, since iron material in LOS could cause greater multiple-path distortions. However, WSVA could still be effective under wood obstruction, and it is feasible for the users to keep them on LOS places in their own smart home.

5.6 Timeliness of Pre-Collected CSI Profiles

In ideal cases, the collected CSI patterns should be only related to mouth motion and do not change with time. However, as reported by previous research [38], [39], CSI patterns are changed over time in real-world scenario. To evaluate the timeliness of CSI profiles, we first recruit a volunteer to provide mouth motion profiles. Then, the volunteer and adversary are required to perform 150 voice commands and 150 spoofing attacks with the time step of 12 hours. Fig. 16 shows the performance of WSVA in real-time, 12 hours and 24 hours. It is observed that after 12 hours, WSVA achieves above 99 percent TAR with 1 percent FAR, which is similar to real-time performance. After 24 hours, WSVA can still achieve 90.6 percent TAR with 1 percent FAR by utilizing the combined feature. Note that after 24 hours, the performance of mouth motion based feature is decreased to 75.8 percent TAR with 2 percent FAR. The performance degradation may caused by the emotion changes of the user or the background environment changes. This is an inherent drawback of CSI based sensing, but it does not hinder the deployment of WSVA essentially. Adaptively updating the user's profile can effectively avoid the effects of environment changes [39]. The updating can be processed during the user's daily usage of voice commands and the cost is affordable for the user since we only utilize 40 mouth motion samples for training.

5.7 Real-World Case Study

In this subsection, we explore the implementation of WSVA in real-world IoT environment. In this study, we study a popular smart home platform Samsung SmartThings, which can work cooperatively with Amazon Alexa, a popular VCS around the world. We develop a SmartApp (an intelligent application which can control multiple devices and let them automatically work together) in SmartThings platform to implement the function of WSVA. As shown in Fig. 17a, the SmartThings hub interacts with the Amazon Alexa, WSVA, a smart light, a smart switch and a smart alert via wireless connections. Note that, the WSVA and Amazon Alexa communicate with SmartApp by generating a virtual device in the SmartApp and deploying a "skill" in Alexa backend respectively. When the Alexa receives the user's voice command such as "let there be light", its skill will send a message to the SmartApp, and at the same time, WSVA performs liveness detection by analyzing the collected voice and CSI data. SmartApp will execute the voice command if and only if the liveness detection is successful, and conduct the corresponding operation (e.g., open the smart light). As shown in

(a) SmartApp user interface

(b) Real case scenario

Fig. 17. SmartThings App UI and system testbed.

Fig. 17b, we ask the volunteers to speak three commands (Listed in Table 2) which could be recognized by Alexa.

In this case study, each command is launched 10 times, and we conduct attacks toward each command 50 times. In liveness detection, when the verification threshold is set to 10^{-3}, WSVA can achieve 100 percent TP with 3.3 percent FP. As shown in Table 3, when decreasing the threshold, the FP will rise while the TP remains 100 percent. In practice, the threshold of WSVA is generally set from 5×10^{-4} to 10^{-3}, which will provide good performance in most cases. For each voice command, the average processing delay of WSVA is 4.8s. We admit that this delay may affect the user experience with the VCS. However, the main causes of this delay are the speech recognition system DeepSpeech (1.9s, accounting for 39.6 percent) and the I/O process of MATLAB (1.7s, accounting for 35.4 percent). In order to reduce these negative impacts, a feasible approach is creating a smaller speech recognition model specifically for common speech commands and implementing WSVA with specialized hardware equipment and algorithm, which will significantly reduce the overall processing delay of WSVA.

6 Limitations and Discussions

6.1 Limitations

The performance evaluation part demonstrates the effectiveness of WSVA on thwarting spoofing attacks. However, there are some limitations that may degrade the detection accuracy of WSVA and leave possibilities for adversary to attack the VCS successfully.

Antenna and User Positions. In this study, the distance between the user and the antennas of WSVA affects the performance of WSVA. When the distance is too long (depending on the hardware condition), the collected CSI cannot reflect the mouth motion components and result in inaccurate judgment of WSVA. However, in smart home environment, many applications of voice control system leverage voice command to control home appliances (e.g., light bulbs and temperature), which also has the specific requirement on the environment factors (including the distance). For example, according to the CNET's report about Amazon Echo, it needs more than one Echo devices for full coverage of a large home [40]. In practice, we can deploy multiple antennas on smart home to make WSVA applicable in a larger area or distance. When the user interacts with VCS, WSVA could dynamically choose the antennas which are closest to the user to collect CSI data.

TABLE 2
The Voice Commands in Case Study

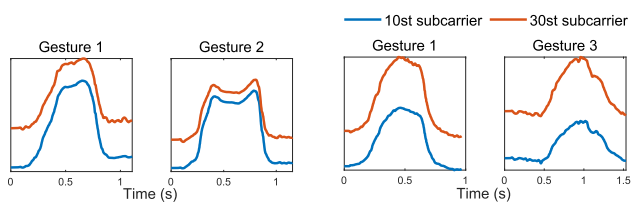
Command	Mouth motions
Make a **ca**ll **to my da**d.	grin, round, pout, hiant, grin.
Send a **me**ssage **to my pho**ne.	grin, grin, pout, hiant, round.
set a **two** o'**clo**ck a**la**rm.	grin, pout, round, hiant.

TABLE 3
The FP and TP Rate in Real-World Case Study

Threshold	10^{-2}	10^{-3}	5×10^{-4}	10^{-4}
Accuracy	20 (66.7%)	30 (100%)	30 (100%)	30 (100%)
TAR	0 (0)	5 (3.3%)	7 (4.7%)	35 (23.3%)

Pronunciation Behaviours. Currently, WSVA can only work for the situation that all users speak the voice commands in English strictly according to the International Phonetic Alphabet. However, in reality, for the same phoneme, different users may use different articulatory gestures [41], [42]. In the experiment, two volunteers are required to pronounce the phoneme /a:/ with standard articulatory gesture (gesture 1 of hiant) and strange gestures (gesture 2 and 3). As shown in Fig. 18, although different articulatory gestures will result in collecting quite different CSI waveforms, it is also observed that when users utilize the same articular gesture (e.g., the gesture 1 used by user 1 and user 2), the collected CSI still have similar patterns. In the family scenario with limited user numbers (generally 2-4 users), it is feasible for these users to agree on a common articulatory gesture. The detection accuracy is also improved by the utilization of macro-level feature. Therefore, WSVA still has high the practicality in multiple-user scenario.

Insider Attack. As described in Section 2.1, the adversary can launch a more serious attack (i.e., insider attack), which is not considered in this study. In an insider attack scenario, the adversary can approach the VCS physically and mimic the mouth motion of a benign user, therefore, it brings the consistency between vibrations of CSI data and voice samples, and decreases the performances of WSVA. To reduce this risk, a potential method is that the user makes special adjustments to the WSVA. For example, the user is required to perform some pre-defined and secret additional mouth motions after each voice commanding. As shown in Fig. 19, WSVA can distinguish between benign users and insider attackers by detecting the existence of this additional motions.



(a) phoneme /a:/ generated by user 1

(b) Detection results under different obstructions.

Fig. 18. CSI waveforms under different articulatory gestures.

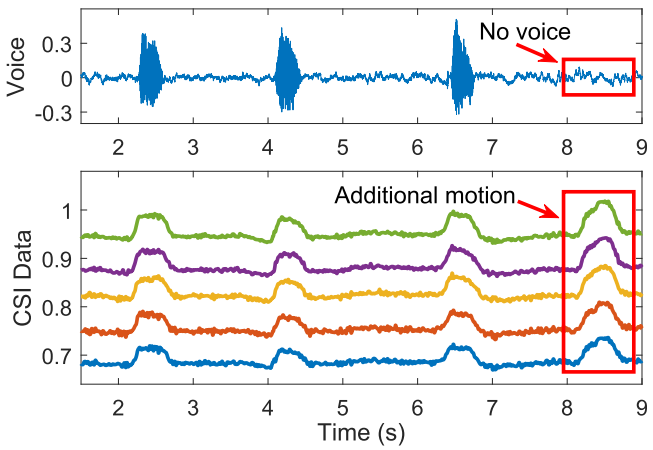


Fig. 19. An example of the strategy to defend insider attack.

Differentiating Multiple Users. With the existence of insider attack, it is reasonable to wonder whether WSVA can differentiate multiple users. In the experiment, the user differentiation is implemented based on the mouth motion feature. Four volunteers are required to provide their CSI profiles and generate 150 voice commands respectively. Then we use the four profiles to train four neural networks. When predicting, the commands are inputted into four neural networks separately. By which neural network provides the highest prediction probability, we can infer the command is issued by which user. Table 4 shows the confusion matrix of user differentiation. The mouth motion feature achieves high accuracy on user differentiation.

However, achieving high user identification accuracy with low FAR (e.g., 2 percent) is difficult. Note that user identification is quite different from multiple-user scenario which differentiates another user from the spoofing one. Fig. 20 and Table 5 show the detecting results when we regard other user's voice as insider attack (as mentioned above). We can see that, with 2 percent FAR, the average TAR is only 53.5 percent, which is not affordable for most users. Therefore we do not recommend using this method to defend against insider attack, and we leave it for future work.

6.2 Compared With Previous Works

In this subsection, we compare WSVA with existing wireless signal based VCS protection method, and discuss the difference between WSVA and lip reading methods.

Comparison Between WSVA and VSButton. Besides WSVA, previous research [43], [44] also proposed a voice liveness detection method named VSButton. As shown in Fig. 21, the basic assumption of VSButton is that a voice command should be generated when the user's physical presence is detected. Thus the main insight of VSButton is detecting the human's indoor motions in the room through CSI, and adding a running condition to the voice assistant system to enhance its security. By contrast, the insight of WSVA is distinguishing a fake voice command by checking the consistency between voice samples and CSI. Technically, WSVA is not limited to detecting human motions with CSI, but also involves matching voice samples with the corresponding mouth motion categories by utilizing phonetic-related knowledge.

TABLE 4
The Results and Accuracy of User Differentiation

Predicted / Actual	user 1	user 2	user 3	user 4
user1	**77.3%**	10%	0	16.7%
user1	19.3%	**66.7%**	0.7%	13.3%
user1	9.3%	7.3%	**80%**	3.3%
user1	4.7%	4.7%	0	**90.7%**

TABLE 5
The TAR and Accuracy of Thwarting Another User's Voice

	user 1	user 2	user 3	user 4
Accuracy	67%	64.8%	82.7%	88%
TAR	36%	31.6%	67.3%	78%

For the aspect of working scenario, since VSButton mainly detects the body movement which has large influence on CSI, the effective detection range could achieve 8 meters, which is much larger than that in WSVA (1.5 meters). However, the VSButton requires the user to move to activate the voice assistance, which is not convenient when the user is sitting or lying. By contrast, WSVA could still authenticate the user's voice commands in real time, as long as the user is located in its working area. We summarize the differences on Table 6, it can be found that while the methodologies of VSButton and WSVA are different, their application scenarios are quite complementary. Therefore the user can deploy VSButton and WSVA simultaneously, and choose either of them according to the actual situation to provide better protection for the VCS.

Comparison Between WSVA and Lip Reading. Previous research have proposed some CSI based lip reading methods such as WiHear [19] and WiTalk [21]. These methods attempt to infer the contents of voice samples only through the CSI information. However, in this study, we do not propose to use wireless signals for lip reading, instead, WSVA aims to utilize the consistency between voice and CSI signals to

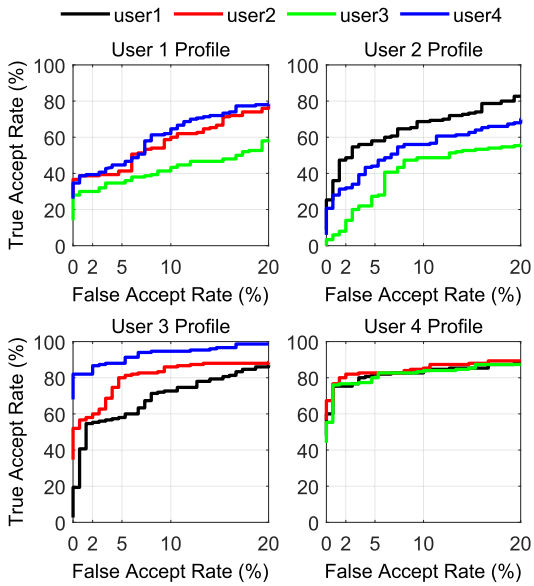


Fig. 20. The results when regarding another user's voice as attack.

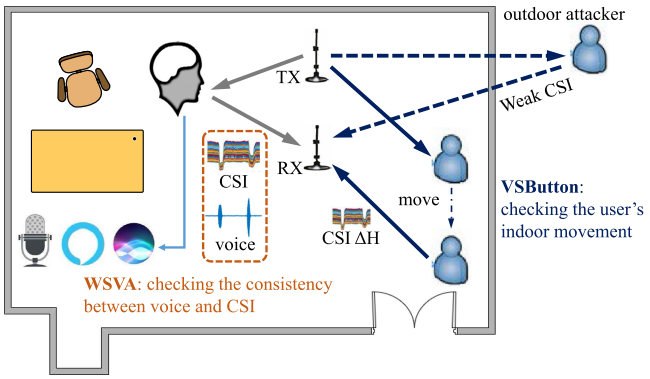


Fig. 21. The insights and scenarios of VSButton and WSVA.

authenticate the voice commands and defend voice spoofing attacks targeted at voice control system.

In addition, due to the limitation of Wi-Fi signals (e.g., only 12.5 cm wavelength for 2.4 GHz), achieving high accuracy detection in lip reading is inherent difficult. For instance, WiHear and WiTalk can only recognize 14 and 12 different syllables respectively, which means that many voice syllables cannot be identified by CSI. Furthermore, [20] shows not all voice syllable can be recognized by lip-reading techniques in theory. For instance, the SilentTalk [45] shows the ultrasonic based lip reading can only identify 12 basic mouth motions, even if ultrasonic (e.g., 17 mm wavelength for 20 kHz) has stronger sensing capability than CSI. However, in the application scenarios of WSVA, the contents of the voice samples are easy to obtain. So instead of recognizing syllables, the technical contribution of WSVA is modeling the consistency between the voice samples and the CSI information to determine whether a voice command is issued by a legitimate user.

7 Related Work

Attacks Towards VCS. Voice interface has been the dominant interface in consumer IoT environment [46], [47], [48]. However, its security issues have been proposed in recent researches [9], [10], [11], [49]. Besides traditional replay attacks, Carlini *et al.* [9] showed that the adversary could produce the voice signals that are difficult to understand by human but could be interpreted to valid commands by VCS. Based on the hardware limitations of VCS, Roy *et al.* [11] demonstrated that its practical to exploits two high-frequency waves to inject voice commands into VCS. Yuan *et al.* [49] found that the voice commands can be stealthily embedded into songs to control the target VCS, while not being noticed by a human listener. Zhang *et al.* [10] exploited inaudible ultrasonic waves to inject state-of-the-art VCS (e.g., Siri and Amazon Alexa) and this attack could achieve almost 100 percent attack success rate for Siri in office environment.

Defense Mechanisms Against VCS Attacks. To enhance the security of VCS against the above attacks, many researchers have proposed defense mechanisms [13], [14], [15], [16], [17], [50], [51]. Papernot *et al.* [51] introduced a defensive distillation based mechanism to reduce the effectiveness of adversarial samples on deep neural network, thus detecting malicious voice commands. Chen *et al.* [14] explored magnetic field emitted from loudspeakers as the essential characteristic for detecting VCS Attacks. Feng *et al.* [16] proposed a scheme which utilizes the acceleration data collected from the user's wearable devices to achieve two-factor based liveness detection. Zhang *et al.* [15], [17] utilized the Doppler effect of ultrasonic generated from the loudspeaker of smartphone to perform liveness detection. Lei *et al.* [43], [44] propose VSButton to thwart voice spoofing by detecting the presence of human via wireless signals, and we give a detailed comparison in Section 6.2.

TABLE 6
Comparison Between VSButton and WSVA

Aspects	VSButton	WSVA
Main assumption	Voice command happens with user's physical presence	Voice command happens with user's moun motion
Key method	CSI based human movement detection	Analyzing the correlation between voiie and CSI
Working range	Up to 8 m	Best within 1.5 m
Main attack scenario	Replay attacks	Replay and inaudible attacks
After authentication	Activate/Turn off the voice assistant	Allow/Deny VCS to conduct voice command
User requirment	Keep movement to activate VCS	Keep stationary except mouth

Wireless Sensing Technologies. Using wireless signals to sense human motion has the advantages of device-free and non-invasion, and recent studies [28], [29], [30], [31], [52] demonstrate its feasibility. Shi *et al.* [28] showed that existing Wi-Fi signals generated by indoor IoT devices can be utilized to achieve user authentication based on the daily activities. Ali *et al.* [52] proposed a keystroke inference systems called WiKey, which uses the CSI waveform pattern to distinguish keystrokes on a external keyboard. Tan *et al.* [29] developed WiFinger to capture subtle changes of finger movements for fine-grained gesture recognition. Qian *et al.* [30] and Wang *et al.* [31] demonstrated using Wi-Fi signals could achieve human localization and tracking with centimeter-level precisions.

8 Conclusion

In this paper, we propose WSVA, a device-free liveness detection system to thwart the spoofing attacks aiming at VCS. WSVA utilizes the prevalent wireless signals in IoT environment to sense the human mouth motion, and then verifies the liveness of voice command according to the consistency between voice samples and CSI data. WSVA does not require the user to carry any device or demand a large number of training data. We implement WSVA on SmartThings platform to demonstrate its feasibility and the results show that WSVA can achieve 99 percent detection accuracy with 1 percent false accept rate.

Acknowledgments

This work is supported in part by National Natural Science Foundation of China (No. 61672350, No. 61972453) and

China Scholarship Council (201906230162). The work of Jin Li is supported in part by the National Natural Science Foundation of China for Outstanding Youth Foundation (No. 61722203) and National Natural Science Foundation of China for Joint Fund Project (No. U1936218).

REFERENCES

[1] Y. Meng *et al.*, "WiVo: Enhancing the security of voice control system via wireless signal in IoT environment," in *Proc. 18th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, 2018, pp. 81–90.

[2] P. Associates, "Top 10 consumer IoT trends in 2017," 2017. [Online]. Available: http://www.parksassociates.com/whitepapers/top10-2017

[3] Amazon, "Amazon alexa developer," 2019. [Online]. Available: https://developer.amazon.com/alexa

[4] Samsung, "Smartthings," 2019. [Online]. Available: https://www.smartthings.com

[5] Google, "Google home," 2019. [Online]. Available: https://store.google.com/product/google_home

[6] G. V. Research, "Voice and speech recognition market size, share and trends analysis report," 2018. [Online]. Available: https://www.grandviewresearch.com/industry-analysis/voice-recognition-market

[7] P. Associates, "The impact of voice technologies on consumer entertainment," 2018. [Online]. Available: http://www.parksassociates.com/report/impact-voice-technologies

[8] W. Diao, X. Liu, Z. Zhou, and K. Zhang, "Your voice assistant is mine: How to abuse speakers to steal information and control your phone," in *Proc. 4th ACM Workshop Secur. Privacy Smartphones Mobile Devices*, 2014, pp. 63–74.

[9] N. Carlini *et al.*, "Hidden voice commands," in *Proc. USENIX Secur. Symp.*, 2016, pp. 513–530.

[10] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, "DolphinAttack: Inaudible voice commands," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2017, pp. 103–117.

[11] N. Roy, H. Hassanieh, and R. Roy Choudhury, "BackDoor: Making microphones hear inaudible sounds," in *Proc. 15th ACM Annu. Int. Conf. Mobile Syst. Appl. Services*, 2017, pp. 2–14.

[12] A. Aley-Raz, N. M. Krause, M. I. Salmon, and R. Y. Gazit, "Device, system, and method of liveness detection utilizing voice biometrics," U.S. Patent 8 442 824B2, May 14, 2013.

[13] Y. Chen, J. Sun, X. Jin, T. Li, R. Zhang, and Y. Zhang, "Your face your heart: Secure mobile face authentication with photoplethysmograms," in *Proc. IEEE Conf. Comput. Commun.*, May 2017, pp. 1–9.

[14] S. Chen *et al.*, "You can hear but you cannot steal: Defending against voice impersonation attacks on smartphones," in *Proc. IEEE 37th Int. Conf. Distrib. Comput. Syst.*, 2017, pp. 183–195.

[15] L. Zhang, S. Tan, J. Yang, and Y. Chen, "VoiceLive: A phoneme localization based liveness detection for voice authentication on smartphones," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 1080–1091.

[16] H. Feng, K. Fawaz, and K. G. Shin, "Continuous authentication for voice assistants," in *Proc. 23rd Annu. Int. Conf. Mobile Comput. Netw.*, 2017, pp. 343–355.

[17] L. Zhang, S. Tan, and J. Yang, "Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2017, pp. 57–71.

[18] T. Wei, S. Wang, A. Zhou, and X. Zhang, "Acoustic eavesdropping through wireless vibrometry," in *Proc. 21st Annu. Int. Conf. Mobile Comput. Netw.*, 2015, pp. 130–141.

[19] G. Wang, Y. Zou, Z. Zhou, K. Wu, and L. M. Ni, "We can hear you with Wi-Fi!" in *Proc. 20th Annu. Int. Conf. Mobile Comput. Netw.*, 2014, pp. 593–604.

[20] D. Ostry and J. Flanagan, "Human jaw movement in mastication and speech," *Archives Oral Biol.*, vol. 34, no. 9, pp. 685–693, 1989.

[21] C. Du, X. Yuan, W. Lou, and Y. T. Hou, "Context-free fine-grained motion sensing using WiFi," in *Proc. 15th Annu. IEEE Int. Conf. Sens. Commun. Netw.*, 2018, pp. 1–9.

[22] SmartThings, "Smartthings public," 2017. [Online]. Available: https://github.com/SmartThingsCommunity/SmartThingsPublic

[23] IEEE Std. 802.11n-2009: Enhancements for higher throughput, 2009. [Online]. Available: http://www.ieee802.org

[24] "Ettus research," 2017. [Online]. Available: https://www.ettus.com/

[25] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "Tool release: Gathering 802.11 n traces with channel state information," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 1, pp. 53–53, 2011.

[26] "Places of articulation," 2017. [Online]. Available: https://en.wikipedia.org/wiki/File:Places_of_articulation.svg

[27] Wikipedia, "International phonetic alphabet," 2018. [Online]. Available: https://en.wikipedia.org/wiki/International_Phonetic_Alphabet

[28] C. Shi, J. Liu, H. Liu, and Y. Chen, "Smart user authentication through actuation of daily activities leveraging WiFi-enabled IoT," in *Proc. 18th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, 2017, pp. 5:1–5:10.

[29] S. Tan and J. Yang, "WiFinger: Leveraging commodity WiFi for fine-grained finger gesture recognition," in *Proc. 17th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, 2016, pp. 201–210.

[30] K. Qian, C. Wu, Z. Yang, Y. Liu, and K. Jamieson, "Widar: Decimeter-level passive tracking via velocity monitoring with commodity Wi-Fi," in *Proc. 18th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, 2017, pp. 6:1–6:10.

[31] J. Wang *et al.*, "LiFS: Low human-effort, device-free localization with fine-grained subcarrier information," in *Proc. 22nd Annu. Int. Conf. Mobile Comput. Netw.*, 2016, pp. 243–256.

[32] B. Dodd and R. Campbell, "Hearing by eye: The psychology of lip-reading," *Amer. J. Psychol.*, vol. 72, no. 6, 1987, Art. no. 479.

[33] S. Sardy, P. Tseng, and A. Bruce, "Robust wavelet denoising," *IEEE Trans. Signal Process.*, vol. 49, no. 6, pp. 1146–1152, Jun. 2001.

[34] C. C. Holt, "Forecasting seasonals and trends by exponentially weighted moving averages," *Int. J. Forecasting*, vol. 20, no. 1, pp. 5–10, 2004.

[35] A. Hannun *et al.*, "Deep Speech: Scaling up end-to-end speech recognition," *arXiv*, vol. abs/1412.5567, pp. 1–12, 2014.

[36] T. Kisler, F. Schiel, and H. Sloetjes, "Signal processing via web services: the use case WebMAUS," in *Proc. Digit. Humanities*, 2012, pp. 30–34.

[37] L. I.-K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, 1989.

[38] K. Ali, A. X. Liu, W. Wang, and M. Shahzad, "Keystroke recognition using WiFi signals," in *Proc. 21st Annu. Int. Conf. Mobile Comput. Netw.*, 2015, pp. 90–102. [Online]. Available: http://doi.acm.org/10.1145/2789168.2790109

[39] J. Liu, H. Liu, Y. Chen, Y. Wang, and C. Wang, "Wireless sensing for human activity: A survey," *IEEE Commun. Surveys Tuts.*, to be published, doi: 10.1109/COMST.2019.2934489.

[40] CNET, "How to bring alexa into every room of your home," 2017. [Online]. Available: https://www.cnet.com/how-to/how-to-install-alexa-in-every-room-of-your-home/

[41] E. M. Mugler, M. C. Tate, K. Livescu, J. W. Templer, M. A. Goldrick, and M. W. Slutzky, "Differential representation of articulatory gestures and phonemes in precentral and inferior frontal gyri," *J. Neurosci.*, vol. 38, no. 46, pp. 9803–9813, 2018.

[42] C. P. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica*, vol. 49, no. 3/4, pp. 155–180, 1992.

[43] X. Lei, G. Tu, A. X. Liu, C. Li, and T. Xie, "The insecurity of home digital voice assistants - Amazon alexa as a case study," *CoRR*, 2017. [Online]. Available: http://arxiv.org/abs/1712.03327

[44] X. Lei, G. Tu, A. X. Liu, C. Li, and T. Xie, "The insecurity of home digital voice assistants - Vulnerabilities, attacks and countermeasures," in *Proc. IEEE Conf. Commun. Netw. Secur.*, 2018, pp. 1–9.

[45] J. Tan, C. Nguyen, and X. Wang, "SilentTalk: Lip reading through ultrasonic sensing on mobile phones," in *Proc. IEEE Conf. Comput. Commun.*, 2017, pp. 1–9.

[46] Y. Meng, W. Zhang, H. Zhu, and X. S. Shen, "Securing consumer IoT in the smart home: Architecture, challenges, and countermeasures," *IEEE Wireless Commun.*, vol. 25, no. 6, pp. 53–59, Dec. 2018.

[47] W. Zhang, Y. Meng, Y. Liu, X. Zhang, Y. Zhang, and H. Zhu, "HoMonit: Monitoring smart home apps from encrypted traffic," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2018, pp. 1074–1088.

[48] Y. Zhang, R. Deng, D. Zheng, J. Li, P. Wu, and J. Cao, "Efficient and robust certificateless signature for data crowdsensing in cloud-assisted industrial IoT," *IEEE Trans. Ind. Informat.*, vol. 15, no. 9, pp. 5099–5108, Sep. 2019.

[49] X. Yuan *et al.*, "Commandersong: A systematic approach for practical adversarial voice recognition," in *Proc. 27th USENIX Secur. Symp.*, 2018, pp. 49–64.
[50] Y. Chen, J. Sun, R. Zhang, and Y. Zhang, "Your song your way: Rhythm-based two-factor authentication for multi-touch mobile devices," in *Proc. IEEE Conf. Comput. Commun.*, 2015, pp. 2686–2694.
[51] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *Proc. IEEE Symp. Secur. Privacy*, 2016, pp. 582–597.
[52] K. Ali, A. X. Liu, W. Wang, and M. Shahzad, "Keystroke recognition using WiFi signals," in *Proc. 21st Annu. Int. Conf. Mobile Comput. Netw.*, 2015, pp. 90–102.

Yan Meng received the BS degree in electronic and information engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2016. He is working toward the PhD degree in the Department of Computer Science and Engineering, Shanghai Jiao Tong University, China. His research interests include wireless network security and IoT security.

Haojin Zhu (Senior Member, IEEE) received the BSc degree in computer science from Wuhan University, China, in 2002, the MSc degree in computer science from Shanghai Jiao Tong University, China, in 2005, and the PhD degree in electrical and computer engineering from the University of Waterloo, Canada, in 2009. Since 2017, he has been a full professor with the Computer Science Department, Shanghai Jiao Tong University, Shanghai, China. His current research interests include network security and privacy enhancing technologies. He published more than 40 international journal papers, including the *Journal of the American Society of Cytopathology*, *IEEE Transactions on Dependable and Secure Computing*, *IEEE Transactions on Parallel & Distributed Systems*, *IEEE Transactions on Mobile Computing*, *IEEE Transactions on Wireless Communications*, *IEEE Transactions on Vehicular Technology*, and 60 international conference papers, including the ACM CCS, ACM MOBICOM, ACM MOBIHOC, the IEEE INFOCOM, and the IEEE ICDCS. He received a number of awards, including the IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award, in 2014, Top 100 Most Cited Chinese Papers Published in International Journals, in 2014, Supervisor of Shanghai Excellent Master Thesis Award, in 2014, a distinguished member of the IEEE INFOCOM Technical Program Committee, in 2015, the Outstanding Youth Post Expert Award for Shanghai Jiao Tong University, in 2014, and the SMC Young Research Award of Shanghai Jiao Tong University, in 2011. He was a co-recipient of Best Paper Award at the IEEE ICC, in 2007 and Chinacom, in 2008 the IEEE GLOBECOM Best Paper Nomination, in 2014, and the WASA Best Paper Runner-up Award, in 2017. He received the Young Scholar Award of Changjiang Scholar Program from the Ministry of Education of P. R. China, in 2016.

Jinlei Li received the BS degree from the Department of Computer Science and Engineering, Shanghai Jiao Tong University of China, in 2018. He is working toward the master's degree with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, China. His research interests are wireless network security and network privacy.

Jin Li received the PhD degree in information security from Sun Yat-sen University, Guangzhou, China, 2007. He is currently a professor of School of Computer Science, Guangzhou University, Guangzhou, China. He served as a senior research associate at Korea Advanced Institute of Technology, Korea and Illinois Institute of Technology, Chicago, Illinois from 2008 to 2010. He has authored more than 50 papers in international conferences and journals. His research interests include design of secure protocols in cloud computing (secure cloud storage, encrypted keyword search, and outsourcing computation), and cryptographic protocols. He also served as a technical program committee member for many international conferences.

Yao Liu (Senior Member, IEEE) received the PhD degree in computer science from North Carolina State University, Raleigh, North Carolina, in 2012. She is currently an associate professor with the Department of Computer Science and Engineering, University of South Florida, Tampa, Florida. Her research interests include computer and network security, with an emphasis on designing and implementing defense approaches that protect emerging mobile and wireless technologies from being undermined by adversaries.