WiVo: Enhancing the Security of Voice Control System via Wireless Signal in IoT Environment

Yan Meng[1], Zichang Wang[1], Wei Zhang[1], Peilin Wu[1], Haojin Zhu[1], Xiaohui Liang[2] and Yao Liu[3]
[1]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University
[2]University of Massachusetts Boston
[3]University of South Florida
{yan_meng,wzc1214,zhang-wei,wplismit,zhu-hj}@sjtu.edu.cn
{xiaohui.liang}@umb.edu,{yliu}@cse.usf.edu

ABSTRACT

With the prevalent of smart devices and home automations, voice command has become a popular User Interface (UI) channel in the IoT environment. Although Voice Control System (VCS) has the advantages of great convenience, it is extremely vulnerable to the spoofing attack (e.g., replay attack, hidden/inaudible command attack) due to its broadcast nature. In this study, we present WiVo, a device-free voice liveness detection system based on the prevalent wireless signals generated by IoT devices without any additional devices or sensors carried by the users. The basic motivation of WiVo is to distinguish the authentic voice command from a spoofed one via its corresponding mouth motions, which can be captured and recognized by wireless signals. To achieve this goal, WiVo builds a theoretical model to characterize the correlation between wireless signal dynamics and the user's voice syllables. WiVo extracts the unique features from both voice and wireless signals, and then calculates the consistency between these different types of signals in order to determine whether the voice command is generated by the authentic user of VCS or an adversary. To evaluate the effectiveness of WiVo, we build a testbed based on Samsung SmartThings framework and include WiVo as a new application, which is expected to significantly enhance the security of the existing VCS. We have evaluated WiVo with 6 participants and different voice commands. Experimental evaluation results demonstrate that WiVo achieves the overall 99% detection rate with 1% false accept rate and has a low latency.

CCS CONCEPTS

• **Security and privacy → Mobile and wireless security**; **Privacy protections**;

KEYWORDS

Liveness Detection, Voice Control System, Wireless Side Channels

Haojin Zhu is the corresponding author.

Mobihoc '18, June 26–29, 2018, Los Angeles, CA, USA

ACM ISBN 978-1-4503-5770-8/18/06…$15.00
https://doi.org/10.1145/3209582.3209591

ACM Reference format:
Yan Meng[1], Zichang Wang[1], Wei Zhang[1], Peilin Wu[1], Haojin Zhu[1], Xiaohui Liang[2] and Yao Liu[3]. 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, Los Angeles, CA, USA, June 26–29, 2018 (Mobihoc '18),* 10 pages.
https://doi.org/10.1145/3209582.3209591

1 INTRODUCTION

Smart home or home automation systems are gaining an increasing popularity due to its great convenience of allowing the users to remotely control their domestic appliances (*e.g.*, LED lights, temperature controller, microwave, refrigerator) via a diversified range of Internet-of-Things (IoT) user interfaces such as wireless communications, voice control and image sensing. According to the Parks Associates' list of Top 10 Consumer IoT Trends in 2017, voice control is vying to become the primary user interface for the smart home and connected lifestyle [5]. The typical IoT voice controllers include Amazon Alexa, Samsung SmartThings, Google Home, and other interactive voice interfaces. The market share of Voice Control System (VCS) was $5.15 billion in 2016 and is expected to reach $18.3 billion by 2023 [4].

Though VCS is regarded as one of the most promising user interfaces, it also introduces new security risks due to its inherent broadcast nature, which makes it extremely vulnerable for spoofing attacks such as *the replay attacks*, *the hidden command attacks* and *the inaudible command attacks*. In the replay attacks, an adversary tries to fool the VCS by using the pre-recorded voice of the legitimate user[11]. The hidden command attacks use a falsified speech signal as the system input[7]. As an extreme case of spoofing attacks, the latest researches [17, 25] show that it is feasible to inject some hidden or even inaudible voice commands which cannot be understood/heard by the human but can still be understood by the VCS. This kind of spoofing attacks opens a new door for the adversary to query the user's sensitive information, and perform undesirable operations, which poses a serious threat on the security of smart home systems.

There are two types of approaches that have been proposed to defend against these attacks, including voice password based access control and two-factor based liveness detection. The password based access control requires the user to speak special password before giving the voice commands [6], which is vulnerable to eavesdropping attack. The two-factor based liveness detection leverages the information that is highly correlated to the VCS operations (e.g., image or video collected by camera [9], magnetic filed emitted from

loudspeakers [8], time-difference-of-arrival changes from different microphones [27], acceleration data of user's wearable devices [13] and the Doppler shift of ultrasonic caused by mouth motion [26]) as the user liveness features. However, the existing two-factor based liveness detection schemes require the users to carry the specialized sensing devices to collect the liveness information which seriously limits its practicality. Some of these schemes also impose unacceptable privacy risks, since image or video data can be utilized to infer the users' behaviors in their daily life.

In this study, we present WiVo, a device-free voice liveness detection system based on the prevalent wireless signals generated by WiFi devices without the users carrying any additional devices or sensors. WiVo starts from the following observations: firstly, according to Lip-reading, it is feasible to understand speech by interpreting the movements of the lips, face and tongue. In other words, voice command can be cross-checked by the mouth motions. Secondly, the existing researches show that a large set of in-door activities can be identified by using device-free Channel State Information (CSI) based sensing techniques. Therefore, it is natural to raise the following question: is it feasible to build the correlation between the CSI change and the mouth motion, and leverage this correlation to verify the liveness of voice command?

To answer the question above, WiVo should address three major challenges: i) The impact of mouth motion on wireless signals is subtle. Although previous works utilize sophisticated method such as MIMO beamforming or Frequency-Modulated Carrier Waves (FMCW) [22, 24] to improve the wireless sensing capability, they may not work for our problem since these sophisticated sensing techniques cannot be implemented in commercial IoT devices. Therefore, for signals collected by commercial off the shelf (COTS) devices, a novel signal processing method is highly desirable. 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. Therefore, we should build a new model to describe the correlation among the CSI changes, mouth motion, and the syllables of voice signal. iii) To correlate the voice and CSI signals, how to select appropriate features from these two-dimensional signals remains a big challenge. The contributions of this work are summarized as follows:

- We present WiVo, a two-factor liveness detection system to thwart the various attacks towards VCS by analyzing the consistency between voice and CSI signals. By utilizing the existing wireless signals in IoT environment, WiVo 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 between the voice signals of syllables and their corresponding CSI change patterns.
- We devise the architecture and algorithms of WiVo. 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.

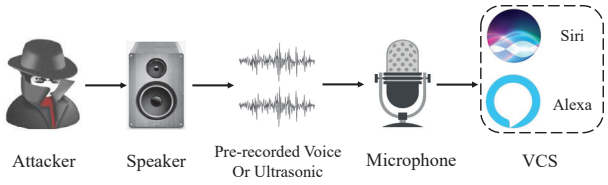


Figure 1: Illustrations of the attack to VCS.

- We design and implement a testbed on Samsung SmartThings platform to evaluate WiVo. Our extensive experimental results on 6 volunteers show that WiVo achieves 99% detection accuracy with 1% false accept rate, and demonstrate the effectiveness and flexibility of WiVo.

To the best of our knowledge, this is the first work to exploit wireless signals to perform liveness detection for VCS. 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 speaking. We elaborate the detailed design of WiVo in Section 4, which is followed by evaluation, discussion and related work in Section 5, 6 and 7 respectively. Finally, we conclude this paper in Section 8.

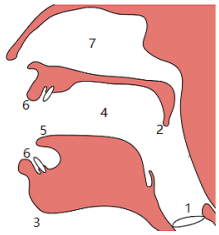
2 PRELIMINARIES

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 fake or outdated voice commands as illustrated in Fig. 1. The existing studies show that there are three variants for spoofing attack.

- Replay attack. The adversary can deploy a recorder to obtain the authentic user's voice samples, and then utilize a loudspeaker to play the pre-collected voice samples to spoof the VCS [11].
- Hidden voice attack. Most VCSs extract Mel-frequency cepstral coefficient (MFCC) from human voice to perform speech recognition. Thus the adversary can generate voice samples which are heard as noise by human ears but contain the user's MFCC features to spoof the VCS [7].
- 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 [17, 25].

Without loss of the generality, in the remainder of this work, we use spoofing attacks to represent the three kinds of attacks above. Our proposed defense technique 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 liveness detection. Our study does not consider the insider attack, which means the adversary can break into the home and impersonate a real user to inject fake voice command. This strong attack model is less practical in a smart home environment, which is out of the scope of this research.



Places of articulation:
1.Vocal cords
2. Soft palate
3. Jaw
4. Oral cavity
5. Tongue
6. Lip
7. Nasal cavity

Figure 2: Vocal organs[3].

2.2 Channel State Information

In this paper, we consider the WiFi wireless communication protocol which is popular in many VCSs [20]. WiFi standards like IEEE 802.11n/ac all support Orthogonal Frequency Division Multiplexing (OFDM), which are expected to significantly improve the channel capacity of the wireless system [1]. In a wireless communication system with transmitter antenna number N_{TX}, receiver antenna number N_{RX} and OFDM subcarriers number N_s, system will use $N_{TX} \times N_{RX} \times N_s$ subcarriers to transmit signal at the same time.

CSI measures 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 i-th subcarrier, CSI value H_i can be defined as:

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

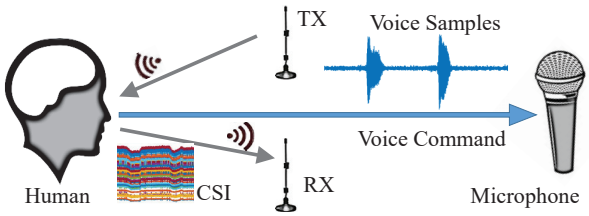
where α is the signal magnitude, f is the frequency and τ is the time-of-light.

Since the received signal reflects the constructive and destructive interference of several multi-path signals scattered from the surrounding objects, the movements of the lips and jaw while issuing a voice command can generate a unique pattern in the time-series of CSI values, which can be related to the voice wave of command. And CSI extraction is quite easy: we can deploy Universal Software Radio Peripheral (USRP) [2] and COTS device (*e.g.*, Intel 5300) to extract CSI with all subcarrier values and 30 subcarrier values respectively.

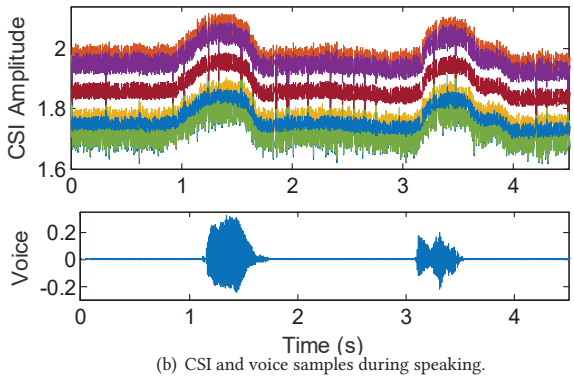
2.3 Articulatory Gesture

It is widely known that the articulation is related to human organs (*e.g.*, vocal cords, tongue, lips, jaw), as shown in Fig. 2. Voice differences depend on the motions of organs, which could affect the vibration frequency of the air, *i.e.*, the timbre. According to the position of the vibrating air, the procedure of voice generation is mainly divided into 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. 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. In this period, the user can produce different phonemes with different motions of tongue, lips and jaw.



(a) Human speaking scenario.



(b) CSI and voice samples during speaking.

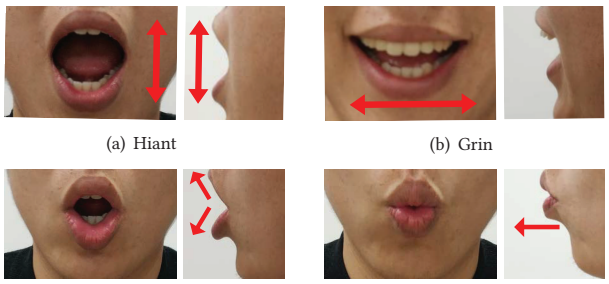
Figure 3: Illustrations of the two different attacks to VCS.

3 MOTIVATION

In this section, we elaborate the rationale behind WiVo by answering the following questions: firstly, do the mouth motions really have the correlation with the change of WiFi signals? Secondly, how can we capture this correlation between 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. 3(a) demonstrates the typical scenario of human speaking in VCS environment such as SmartThings or Amazon Alexa platform. When a user interacts with VCS, WiVo exploits a pair of antennas of the IoT devices in the proximity to collect the CSI data, and the microphone starts recording the voice samples simultaneously. Generally speaking, since CSI reflects the constructive and destructive interference of 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 liveness pattern of the user. As shown in Fig. 3(b), the dramatic fluctuations of CSI waveforms happen with the occurrence of human voice. If an adversary launches the various spoofing attacks described in Section 2.1, which means the fake voice command is injected without any corresponding mouth motions, 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 possible to leverage the consistency of fluctuations between voice samples and CSI streams to detect the spoofing attacks.



(a) Hiant (b) Grin

(c) Round (d) Pout

Figure 4: Mouth shapes of four basic syllables.

Table 1: Four categories of syllables.

Mouth motion	Syllables	Words
Hiant	/a:/ /æ/ /ai/	b**ar**, bicycle
Grin	/e/ /ei/	**A**, **base**
Round	/ɔ/ /ɔ:/	l**ot**, s**aw**
Pout	/u:/ /ʊ/	r**oot**, sh**oe**
Non-significant	/ə/ /iə/	s**ir**, h**ere**

3.2 Building the Correlation of CSI Vibration and Voice Syllables

The previous works have demonstrated the feasibility of sensing human movements via wireless signals. However, achieving very precise syllable recognition is less possible in IoT environment since it may be beyond the sensing capability of WiFi signal. As shown in Eqn. 1, the sensing capability of wireless signal depends on the wavelength of signals. In practice, the WiFi signal (*i.e.*, 12.5cm wavelength for 2.4GHz) based sensing mechanisms cannot accurately capture the tiny motion of human mouth. What is worse, WiFi can only recognize the motion of tongue, lips and jaw, and the impact of other vocal organs can not be recognized. According to the study of Dodd et al. [12], only 40% words in English can be recognized by only considering mouth motions.

To address the challenges above, in this paper, we classify the mouth motions into four categories, including hiant, grin, round and pout, which correspond to Fig 4.(a), (b), (c) and (d), respectively. Most voice syllables can be categorized into one of these types, while only a few syllables with non-significant mouth movements cannot be precisely captured by WiFi. More specifically, the hiant, the motion of opening the mouth largely, can pronounce the syllable that includes the phoneme like /*a:*/ and /æ/, which can be heard in words, such as "**bar**" and "**ha**". The grin, the motion of grinning human mouth like Fig 4.(b), can pronounce the syllable that is made of phoneme, like /e/ and /ei/, which can be heard in words, such as "**A**" and "**base**". The round, rounding lips at ease, can generate the syllable that made of phoneme, like /ɔ:/ , which can be heard in words, such as "**lot**" and "**saw**". Finally, the pout, the motion pouting the lips, can send out the syllable that is made of phoneme, like /u:/, which can be heard in words, such as "**root**" and "**shoe**". After such division, different types of syllables can be correlated with different CSI features as mentioned in the following sections.

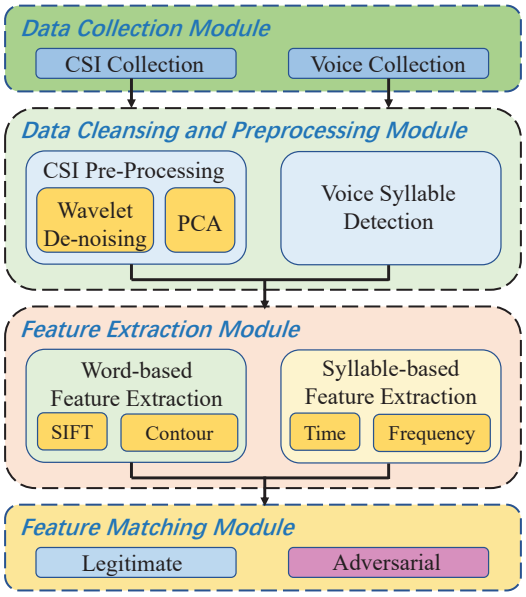


Figure 5: Workflow of WiVo.

4 SYSTEM DESIGN

4.1 System Overview

The basic strategy of WiVo is detecting if a voice command is an authentic one by 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 smart home, with the prevalent of IoT platforms such as Samsung SmartThings, which controls the smart devices with wireless signals, it is technically feasible to take advantage of these existing wireless infrastructures to collect the voice samples and their corresponding CSI data simultaneously.

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

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), it is required for the user to speak a predefined magic word, which can be utilized as a trigger. For instance, Apple iPhone needs "Hey, Siri" and Amazon Alexa needs "Alexa" to initialize the voice assistant. WiVo only starts when the voice trigger is recognized by the VCS. After WiVo is activated, WiVo utilizes two antennas to collect CSI data. These antennas

can be equipped by the different devices, or incorporated in the same device in IoT environment. WiVo allows a transmit antenna to continuously send wireless packets (such as broadcast packets) and another antenna to receive packets, and extracts CSI data from the preamble sequences of these packets.

4.3 Data Cleansing and Preprocessing

4.3.1 CSI Denoising. Before launching liveness detection, WiVo leverages wavelet denoising to eliminate the high frequency noises from the collected CSI data. Wavelet denoising includes three main steps as follows:

Discrete Wavelet Transform (DWT). Generally speaking, an original discrete signal $x[n]$ can be expressed in terms of the wavelet function by the following equation:

$$x[n] = \frac{1}{\sqrt{L}} \sum_{k} W_{\phi}[j_0, k]\phi_{j_0,k}[n] + \frac{1}{\sqrt{L}} \sum_{j=j_0}^{\infty} \sum_{k} W_{\psi}[j, k]\psi_{j,k}[n], \tag{2}$$

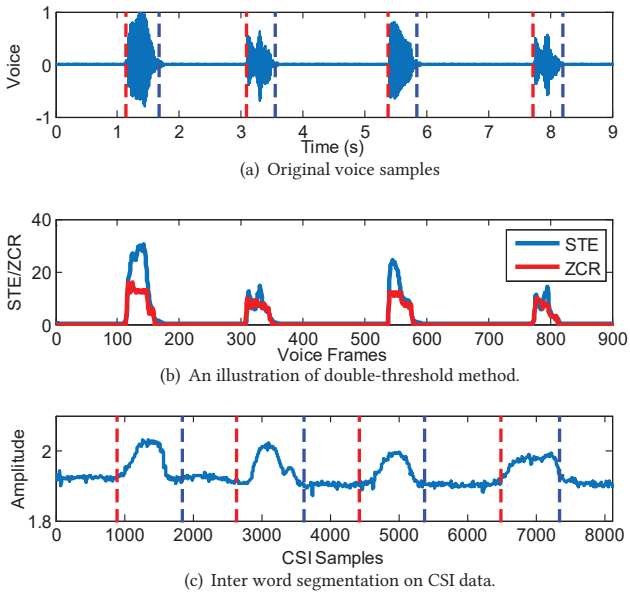
where L represents the length of $x[n]$. The functions $\phi_{j_0,k}[n]$ refer to scaling functions and the corresponding coefficients $W_{\phi}[j_0, k]$ refer to the approximation coefficients. Similarly, functions $\psi_{j,k}[n]$ refer to wavelet functions and coefficients $W_{\psi}[j, k]$ refer to detail coefficients. During the decomposition process, the origin signal is firstly divided into the approximation coefficients and detail coefficients. 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 (high-frequency) and approximation coefficients (low-frequency) at different frequency levels. Then, the threshold is applied to the detail coefficients to remove their noisy parts. In this study, we empirically choose an adaptive minimax threshold based on the experimental results.

Wavelet Reconstruction. After the two steps above, we reconstruct the signal to achieve noise removal by combining the coefficients of the last approximation level with all thresholded details. We choose Daubechies D4 wavelet [18] and perform 4-level DWT decomposition in wavelet denoising in this study. As shown in Fig. 6(c), after wavelet-based denoising, most of the burst noises can be removed.

4.3.2 Word and Syllable Extraction. After performing wavelet based denoising, it is observed that the CSI waveform shows a strong correlation with mouth motions. To compare the consistency between CSI and voice samples, it is critical to detect the start and end point of syllable on CSI data. However, directly applying burst detection method on CSI data does not work well in this case since the CSI waveform has many break points during the mouth motions. Therefore, we firstly perform word detection on voice samples, then divide the word sample into multiple syllables, and finally extract the corresponding CSI syllable data according to the timestamps.

Inter Word Segmentation. During speaking a command, there is a short interval (e.g., 200ms) between pronouncing two successive words. Therefore WiVo can leverage the interval between two voice samples to segment the words. WiVo exploits double-threshold detection method in this paper. Specifically, WiVo splits the voice samples $v[n]$ into frames of 512 points length, with shifting 256 points each time. For totally N frames, WiVo calculates the short



(a) Original voice samples
(b) An illustration of double-threshold method.
(c) Inter word segmentation on CSI data.

Figure 6: An example of inter word segmentation.

term energy $STE[n]$ and zero-crossing rate $ZCR[n]$, and chooses two adaptive thresholds for $STE[n]$ and $ZCR[n]$ to detect the start and end points of word. Fig. 6 illustrates the proceeding of dividing the CSI data into several word waveforms. Thus, we can divide the CSI data into several word waveforms. The i^{th} word's waveform W_i from the k^{th} subcarrier $H(:, k)$ can be represented as follows.

$$W_i = H(s_i : e_i, k), \tag{3}$$

where s_i and e_i are the start and end CSI indexes of the i^{th} words which are converted from the timestamps on voice samples. Note that, s_i and e_i are extended on both sides by 200 CSI samples, respectively, due to the fact that the CSI change introduced by the mouth motion can be observed a little bit earlier than the speech can be heard.

Inner Word Segmentation. The next step of WiVo is dividing the given CSI word waveform into multiple CSI syllable waveforms. Then WiVo calculates the similarity between the collected CSI syllable samples and pre-trained CSI syllables. Similar to inter word segmentation, WiVo processes the voice samples of a given word and obtains the corresponding syllables. To extract the syllables from a word, WiVo utilizes Munich Automatic Segmentation System (MAUS), a widely adopted phonetic segmentation system [14]. MAUS is based on the Hidden Markov Model method, and it can label the syllables of voice signals by analyzing the sound file and text description of the voice. Fig. 7 is an example of the operation result of MAUS, which segments a voice signal ("Open the door") into several phonemes. Then, WiVo combines these phonemes into multiple syllables based on phonetic knowledge. It is worth mentioning that there are still some subtle errors in the syllables segmentation results utilizing MAUS, so WiVo also utilizes the above inter word segmentation result to improve the segmentation performance. After these steps, we obtain the start and end points of each syllable

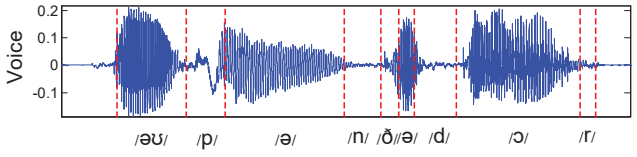


Figure 7: An example of the syllable detection.

in a voice signal, and then extract the corresponding CSI syllable samples.

4.4 Word Based Feature Extraction

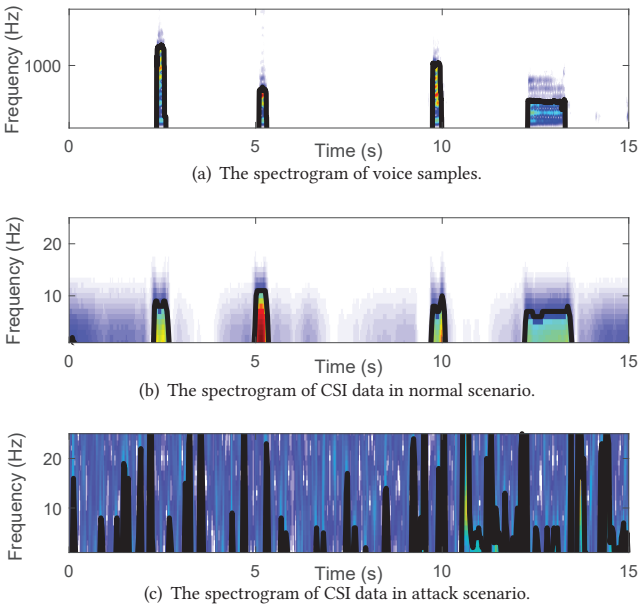
As mentioned in Section 3.1, it is observed that CSI variation occurs along with the human pronunciation. Thus, after performing the wavelet-based noise removing and inter-word segmentation, WiVo calculates the correlation between CSI and voice samples to determine if the voice command and the mouth motion are consistent.

In particular, after performing wavelet denoising, CSI data still composes of 52 subcarriers. To remove the DC components in all subcarriers and extract the strongest correlation component with mouth motions, WiVo adopts PCA to extract the first principle component of all CSI subcarriers. Then, we adopt Short Time Fourier Transform (STFT) to obtain the two-dimensional frequency spectrograms of the CSI data and voice samples. Fig. 8(a) shows the frequency shifts on the voice samples. Fig. 8(b) and Fig. 8(c) show the corresponding CSI data spectrograms in a non-attack scenario, and in a spoofing attack scenario, respectively. Note that, in a spoofing attack, the recorded voice is injected without any corresponding mouth motion. As shown in Fig. 8, the contours are marked by the black lines. It is observed that, in a non-attack scenario, the contours of CSI and voice samples have similar variation trends. However, in an attack scenario, the CSI variants are not in line with the corresponding voice due to lack of corresponding mouth motions.

Thus we can calculate the similarity between the contours of both CSI and voice spectrograms to measure its correlation. To extract the contour from the CSI frequency spectrogram, we firstly resize the CSI spectrogram with frequency from 0 to 30Hz into a m-by-n matrix $M_{CSI}(i,h)$ and normalize the $M_{CSI}(i,h)$ to a range between 0 and 1. Note that, in $M_{CSI}(i,h)$, each column represents the normalized frequency shifts during the i^{th} time slide. Then, we choose a pre-defined *threshold* and get the contour $C_{CSI}(i)$, where $i = 1...n$. $C_{CSI}(i)$ is the maximum value j which satisfies that $M_{CSI}(i,j) \geq threshold$. Calculating contours $C_{Voice}(i)$ for the voice spectrograms is similar to calculating $C_{CSI}(i)$. However, as mentioned in Section 4.3.2, we can set the value $C_{Voice}(i)$ to 0, if the i^{th} time slide is not within the word segments.

After obtaining $C_{CSI}(i)$ and $C_{Voice}(i)$, we measure the correlation between these two contours by adopting Pearson correlation coefficient [15], which is defined as *Corr*. *Corr* ranges from 0 to +1, where a higher value of *Corr* represents a higher level of similarity. To calculate *Corr*, we first re-sample $C_{CSI}(i)$ and $C_{Voice}(i)$ into the same length, and *Corr* can be calculated as:

$$Corr = \left| \frac{\sum_{i=1}^{n}(C_{CSI}(i) - \overline{C_{CSI}})(C_{Voice}(i) - \overline{C_{Voice}})}{(n-1)\delta_{CSI}\delta_{Voice}} \right|, \quad (4)$$



(a) The spectrogram of voice samples.

(b) The spectrogram of CSI data in normal scenario.

(c) The spectrogram of CSI data in attack scenario.

Figure 8: Illustration of the word based feature.

where n is the length of re-sampled sequences $C_{CSI}(i)$ and $C_{Voice}(i)$, δ_{CSI} and δ_{Voice} are the sample standard deviations of $C_{CSI}(i)$ and $C_{Voice}(i)$, respectively.

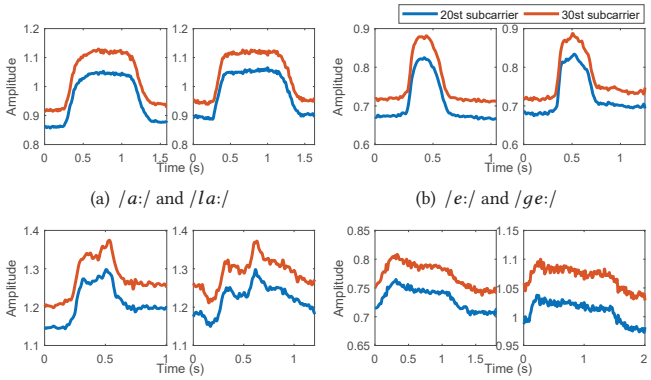
4.5 Syllable Based Feature Extraction

In the previous section, we have discussed how to obtain the word based feature *Corr* from CSI data and voice samples during the voice command pronunciation. However, it may be not enough to perform liveness detection only relying on *Corr*. For example, the dramatic change of environment will generate the drastic vibrations of CSI data, which lead to a deviated contour C_{CSI}. To further improve its performance, we will discuss how to extract time and frequency domain features of CSI syllable data in this section.

Time Domain Feature Extraction. Fig. 9 shows the amplitudes of CSI syllable data extracted from Section 4.3.2. It is observed that the CSI waveforms belonging to the same mouth motion category have the similar shapes. For instance, in Fig .9(a), the waveforms of syllable /*a*:/ and /*la*:/ have the similar waveform shapes and amplitude vibrations. And it is also discovered that the ranges of CSI amplitudes from different syllable categories are quite different. For instance, as shown in Fig. 9(a) and Fig. 9(d), 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 syllable data H_S, the CSI time domain feature $Range(H_S)$ can be calculated as:

$$Range(H_S) = \sum_{i=1}^{N} \frac{Max(H_{Si}) - Min(H_{Si})}{N \times Mean(H_{Si})}, \quad (5)$$

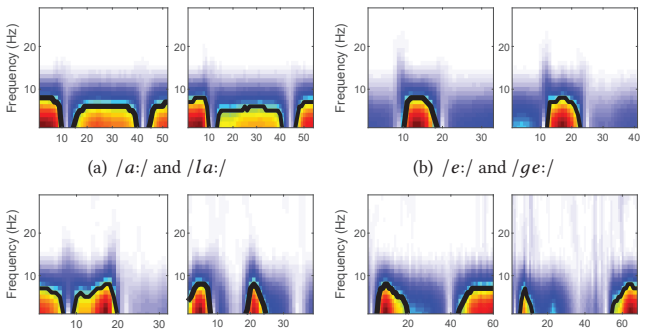
where N represents the number of CSI subcarriers and H_{Si} represents the i^{th} subcarrier of H_S.



(a) /a:/ and /la:/ (b) /e:/ and /ge:/

(c) /re:/ and /ra:/ (d) /u:/ and /gu:/

Figure 9: Time domain of four syllable types.



(a) /a:/ and /la:/ (b) /e:/ and /ge:/

(c) /re:/ and /ra:/ (d) /u:/ and /gu:/

Figure 10: Frequency domain of four syllable types.

Frequency Domain Feature Extraction. In time domain feature extraction part, we do not take the CSI waveform shapes as the time domain features, because the CSI waveform shape changes over time. However, the experimental results show that the frequency shifts of CSI data caused by mouth motion 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 are marked as black 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, we can utilize these contours as frequency features of the CSI syllable data. For a given CSI syllable data H_S, WiVo calculates the corresponding spectrogram contour C_{Hs} as described in Section 4.4.

4.6 Similarity Comparison and Liveness Identification

Before WiVo performs liveness detection, it is reasonable to assume that the user can provide totally M-by-N pre-collected CSI syllable data H_{Pre}, which contain M syllable categories (i.e., four mouth motion categories proposed in Section 3.2) and each category contains N syllables $H_{Pre}(i,j)$. The input data of WiVo are voice samples and CSI data H of a human voice command. After processing the input by using the above mentioned modules, WiVo calculates a word based similarity score $Corr$, and N_S syllable based features $Range(Hs(i))$ and $C_{Hs}(i)$ for N_S syllable $Hs(i)$ $(i = 1, ..., N_S)$ within the input voice samples.

Syllable Feature Combination. WiVo firstly calculates the range difference between the given CSI syllable data H_S and each CSI syllable category, which can be calculated as:

$$SMR_{Time}(i) = \sum_{j=1}^{N} \left| \frac{Range(Hs) - Range(H_{Pre}(i,j)}{N} \right|. \quad (6)$$

Since the corresponding syllable type of H_S can be calculated from the voice processing module as shown in Section 4.3.2, we can calculate the similarity score between H_S and its corresponding syllable type as follow:

$$S_{Time} = Min\{1 - \frac{SMR_{Time}(type)}{Max(SMR_{Time})} + \alpha, 1\}, \quad (7)$$

where $type$ represents the syllable type of H_S, which ranges from 1 to M. 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, WiVo compares the similarity between the spectrogram contour C_{Hs} and the M-by-N syllable contours $C_{Pre}(i,j)$ from pre-collected CSI syllables data H_{Pre}. WiVo utilizes Dynamic Time Wrapping (DTW) to calculate the similarity between C_{HS} and C_{Pre}. DTW is a dynamic programing method to calculate the similarity between two sequence with different length, and a smaller result represents a higher similarity. The similarity between C_{Hs} and four syllable categories (i.e. mouth motion categories) can be calculated as:

$$SMR_{Freq}(i) = \sum_{j=1}^{N} \frac{DTW(C_{Hs}, C_{Pre}(i,j))}{N}. \quad (8)$$

Similar to Eqn. 7, WiVo calculates the similarity score between C_{Hs} and its corresponding i^{th} syllable categories as:

$$S_{Freq} = Min\{1 - \frac{SMR_{Freq}(type)}{Max(SMR_{Freq})} + \alpha, 1\}, \quad (9)$$

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

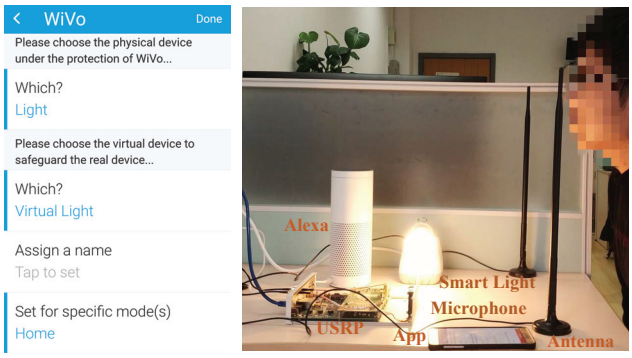
After obtaining the time domain similarity score S_{Time} and the frequency domain similarity score S_{Freq} of a given CSI syllable data H_S, we can calculate the combination syllable based similarity score S_{syll} as:

$$S_{syll} = S_{Time} \times S_{Freq}. \quad (10)$$

Note that, if the syllable is not within the M syllable categories, we will discard its similarity score.

Liveness Detection. After performing syllable feature combination for a given voice command, we obtain its word based feature $Corr$ and the $S_{syll}(i)$ of each syllable, where $i = 1, 2, ..., N_S$. Then, we can calculate the finial decision score of the input, which is calculated as:

$$Score = Corr \times \prod_{i=1}^{N_S} S_{syll}(i). \quad (11)$$



(a) SmartApp user interface (b) Real case scenario

Figure 11: SmartThings App UI and system testbed.

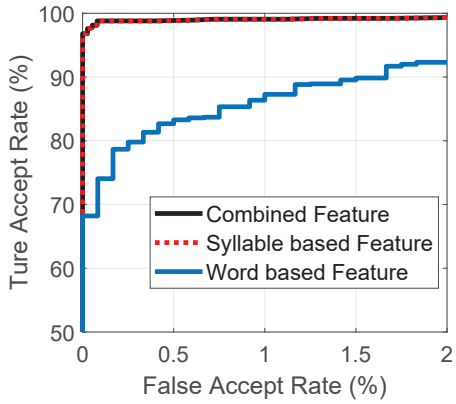


Figure 12: Performance on thwarting spoofing attacks.

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 threshold, WiVo regards it as an authentic voice command. Otherwise, WiVo 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

5.1 System Setup

Hardwares. WiVo consists of two hardwares: i) an Universal Software Radio Peripheral (USRP) N210 device which connects two commercial WiFi antennas, and ii) a microphone, responsible for collecting voice samples. In the experiment, the distance between antennas and human is 20cm. The USRP N210 collects CSI data at the rate of 1000 packets/second in 2.4GHz WiFi frequency with the 1/2 BPSK modulation mechanism. We exploit USRP rather than COTS device (e.g., Intel 5300 NIC) to collect CSI data, since some commercial devices change its power adaptively and result in instable CSI measurements. Choosing USRP can achieve more stable CSI data. However, USRP and COTS devices have the same functions in essential.

Feasibility of WiVo. In the experiment, WiVo is incorporated with Samsung SmartThings platform, which is compatible with Amazon Alexa, a popular VCS around the world. We develop a SmartApp in SmartThings platform to implement the function of WiVo. As shown in Fig. 11, the SmartThings hub interacts with the Amazon Alexa, WiVo and a smart light with wireless connections. Note that, the WiVo interacts with SmartApp by generating a virtual device in the SmartApp. In the experiments, when the Alexa receives the human voice command such as "let there be light", it will send the corresponding command to the hub, and at the same time, WiVo performs liveness detection by analyzing the collected CSI and voice samples. SmartApp will execute the Alexa's command if and only if the liveness detection of WiVo is successful, and then open the smart light. Otherwise, SmartApp regards the voice command as an unauthentic one without executing.

Data Collection. We totally recruit 6 volunteers in the experiment. 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 WiVo's pre-collected syllable profiles. Then, each volunteer performs voice commands and the adversary performs spoofing attacks for this volunteer's profiles. Then WiVo performs liveness detection by analyzing the collected CSI data and voice samples with the volunteer's syllable profiles.

Metrics. To assess the performance of WiVo, we choose the False Accept Rate (FAR) and the True Accept rate (TAR) as metrics. TAR is the rate which WiVo detects the authentic user correctly. 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 adjusting the verification threshold, and we show their relationship using Receiver Operating Characteristic (ROC) curve. In our experiment, we adjust the threshold value of WiVo to study more comprehensive results.

5.2 Thwarting Spoofing Attacks

In this subsection, we evaluate the effectiveness of WiVo to defend against the spoofing attacks. To perform legitimate voice commands, each volunteer is required to speak 150 voice commands. After that, we perform spoofing attacks for each user's syllable profiles for 750 times. For totally 5400 voice commands, the lengths of those four types of syllables proposed in Section 3.2 range from 4 to 8. Fig. 12 depicts ROC curve of WiVo in detecting live users in non-attack scenario and in spoofing attack scenario. We observe that with 1% FAR, the detection rate is as high as 99.1%, using combined word and syllable based features. More specifically, we find that the syllable based feature is much better than word based features. For instance, with 1% FAR, the syllable based detection rate still keeps 99%. However, the word based detection rate is reduced to 87.3%. The reason is that the word based features are more susceptible to the environment noise. After WiVo collecting voice and CSI data, the average time delay of performing per liveness detection is 0.32 second, which is acceptable in practice. In summary, our experimental results well validate the effectiveness of WiVo on defending spoofing attack.

5.3 Scale up to Multiple User's Scenario

In Section 5.2, for each user, WiVo performs liveness detection based on his/her CSI syllable profiles. However, in some smart

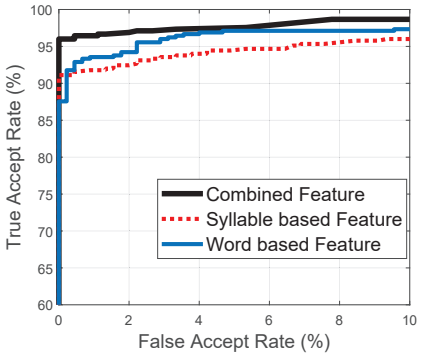


Figure 13: Scaling up to multiple users.

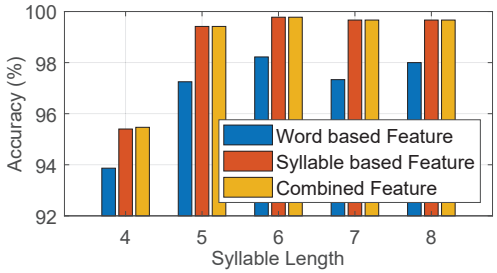
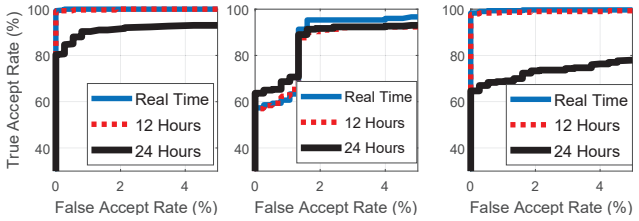


Figure 14: The impact of syllable length.

home environments with multiple users, it is less likely to collect each user's syllable profiles. A more desirable design is to collect once but work for multiple users. In this section, we perform experiments to evaluate the scalability of WiVo. In experiment, we firstly recruit a volunteer to provide WiVo with his/her syllable profiles and record his/her articulatory gesture. Then we recruit another volunteer to study the articulatory gesture and perform voice commands for 450 times. After that, we implement spoofing attacks on Amazon Alexa for 450 times too. Fig. 13 shows the evaluation result of WiVo, where WiVo achieves 96.4% TAR with 1% FAR, and 96.8% TAR with 2% FAR. Note that, the detection rate of syllable based feature is smaller than that in Section 5.2. The reason is that the mouth motion of another volunteer is not the same as the user which provides the pre-collected syllable profiles. However, compared with spoofing attacks, WiVo can still achieve a high detection accuracy, which demonstrates that it is also highly effective in multiple users scenario.

5.4 Impact of Syllable Lengths

In this subsection, we investigate the impact of different syllable lengths of the voice command. Fig. 14 shows the accuracy of different syllable lengths at 2% FAR. The accuracy is the rate of successfully detecting authentic and spoofing commands among all commands. We can find that with the increase of syllable lengths, the accuracy raises from 99.5% to 99.7% slightly. This result indicates that a longer syllable lengths can reduce the impact of misjudgment of a single syllable. The accuracy of every syllable length is over



(a) Combined feature (b) Word based feature (c) Syllable based feature

Figure 15: The impact of time.

90%, which means the features extracted from syllables by WiVo are accurate enough for liveness detection.

5.5 Timeliness of CSI Syllable Profiles

It is well known that the mouth motion can be affected by the emotion or vigor of the user. Further, wireless signals are quite dynamic. Therefore, the timeliness of user's syllable profiles may affect the WiVo's effectiveness. In the experiment, we recruit a volunteer to provide syllable profiles. After that, we require the volunteer to perform voice commands, and the adversary launches spoofing attacks every 12 hours. In each time, the volunteer speaks 75 voice commands and the adversary performs spoofing attack for 75 times too. Fig. 15 shows the performance of WiVo under real-time, 12 hours and 24 hours. It is observed that after 12 hours, WiVo achieves 99% TAR with 1% FAR, which is similar to real time performance. After 24 hours, WiVo can still achieve 90.3% TAR with 1% FAR by utilizing the combined feature. Note that, after 24 hours, the performance of syllable based feature is decreased to 73.3% TAR with 2% FAR. The performance degradation may be caused by emotion changes of user or the background environment changes. However, by utilizing the combined feature, WiVo can still achieve at least 90% TAR with 2% FAR, which is acceptable in practice. And we can furthermore enable the WiVo to adaptively update the user's profiles to improve its performance.

6 DISCUSSIONS

The performance evaluation part demonstrates the effectiveness of WiVo on thwarting spoofing attacks. However, there are some limitations that may degrade the detection accuracy of WiVo and leave possibilities for adversary to attack the VCS successfully. In this study, the distance between the user and the antennas of WiVo affects the performance of WiVo. When the distance is too long, the collected CSI cannot reflect the mouth motion components and result in inaccurate judgment of WiVo. To solve this limitation, a practical solution is increasing the density of IoT devices to make sure that the user locates in the effective range of WiVo in most usage scenarios of VCS. Therefore, when user interacts with VCS, WiVo could dynamically choose the antennas which are closest to the user to collect CSI data. Besides, the adversary can launch insider attack, which is not considered in this study. The insider 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. To thwart the insider attack, more

sophisticated wireless sensing techniques need to be proposed, and we will leave it for the future work.

7 RELATED WORK

Attacks Towards VCS. With the prevalent of VCS, the security issues have been proposed in recent researches [7, 17, 25]. Besides traditional replay attacks, Carlini et al. [7] 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. [17] demonstrated that its practical to exploits two high-frequency waves to inject voice commands into VCS. Zhang et al. [25] exploited inaudible ultrasonic waves to inject state-of-the-art VCS (e.g., Siri and Amazon Alexa) and this attack could achieve almost 100% 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 [9, 10, 13, 26, 27]. Feng et al. [13] 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. [26, 27] utilized the Doppler effect of ultrasonic generated from the loudspeaker of smartphone to perform liveness detection. However, these schemes required the user either to wear specialized devices or hold the phone with a fixed manner.

Wireless Sensing Technologies. Using wireless signals to sense human motion has the advantages of device-free and non-invasion, and recent studies[16, 19, 21, 23] demonstrate its feasibility. Shi et al. [19] showed that existing WiFi signals generated by indoor IoT devices can be utilized to achieve user authentication based on the daily activities. Tan et al. [21] developed WiFinger to capture subtle changes of finger movements for fine-grained gesture recognition. Qian et al. [16] and Wang et al. [23] demonstrated using WiFi signals could achieve human localization and tracking with centimeter-level precisions.

8 CONCLUSION

In this paper, we propose WiVo, a device-free liveness detection system to thwart the spoofing attacks toward VCS. WiVo 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. WiVo does not require the user to carry any device or demand a large number of training data. We implement WiVo on SmartThings platform to demonstrate its feasibility and the results show that WiVo can achieve 99% detection accuracy with 1% false accept rate.

ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China (No. 61672350, U1401253) and National Science Foundation (No. 1618893, No. 1553304, No. 1527144).

REFERENCES

[1] 2009. IEEE Std. 802.11n-2009: Enhancements for higher throughput. http://www.ieee802.org.
[2] 2017. Ettus Research. https://www.ettus.com/
[3] 2017. Places of articulation. https://en.wikipedia.org/wiki/File:Places_of_articulation.svg
[4] 2017. Speech and Voice Recognition Market by Technology. https://www.marketsandmarkets.com/Market-Reports/speech-voice-recognition-market-202401714.html
[5] 2017. Top 10 Consumer IoT Trends in 2017. http://www.parksassociates.com/whitepapers/top10-2017
[6] Almog Aley-Raz, Nir Moshe Krause, Michael Itzhak Salmon, and Ran Yehoshua Gazit. 2013. Device, system, and method of liveness detection utilizing voice biometrics.
[7] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. 2016. Hidden Voice Commands.. In *USENIX Security Symposium*. 513–530.
[8] S. Chen, K. Ren, S. Piao, C. Wang, Q. Wang, J. Weng, L. Su, and A. 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)*. 183–195.
[9] Y. Chen, J. Sun, X. Jin, T. Li, R. Zhang, and Y. Zhang. 2017. Your face your heart: Secure mobile face authentication with photoplethysmograms. In *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*. 1–9.
[10] Y. Chen, J. Sun, R. Zhang, and Y. Zhang. 2015. Your song your way: Rhythm-based two-factor authentication for multi-touch mobile devices. In *2015 IEEE Conference on Computer Communications (INFOCOM)*. 2686–2694.
[11] Wenrui Diao, Xiangyu Liu, Zhe Zhou, and Kehuan Zhang. 2014. Your Voice Assistant is Mine: How to Abuse Speakers to Steal Information and Control Your Phone. In *Proceedings of the 4th ACM Workshop on Security and Privacy in Smartphones & Mobile Devices (SPSM '14)*. 63–74.
[12] Barbara Dodd and Ruth Campbell. 1987. Hearing by eye: The psychology of lip-reading. *American Journal of Psychology* 72, 6 (1987).
[13] 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 (MobiCom '17)*. 343–355.
[14] T. Kisler, F. Schiel, and H. Sloetjes. 2012. Signal processing via web services: the use case WebMAUS. In *Digital Humanities 2012, Hamburg, Germany*. 30–34.
[15] Lawrence I-Kuei Lin. 1989. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* 45, 1 (1989), 255–268. http://www.jstor.org/stable/2532051
[16] Kun Qian, Chenshu Wu, Zheng Yang, Yunhao Liu, and Kyle Jamieson. 2017. Widar: Decimeter-Level Passive Tracking via Velocity Monitoring with Commodity Wi-Fi. In *Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing (Mobihoc '17)*. Article 6, 10 pages.
[17] 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 (MobiSys '17)*. 2–14.
[18] S. Sardy, P. Tseng, and A. Bruce. 2001. Robust wavelet denoising. *IEEE Transactions on Signal Processing* 49, 6 (2001), 1146–1152.
[19] Cong Shi, Jian Liu, Hongbo Liu, and Yingying Chen. 2017. Smart User Authentication Through Actuation of Daily Activities Leveraging WiFi-enabled IoT. In *Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing (Mobihoc '17)*. Article 5, 10 pages.
[20] SmartThings. 2017. SmartThings Public. https://github.com/SmartThingsCommunity/SmartThingsPublic
[21] Sheng Tan and Jie Yang. 2016. WiFinger: Leveraging Commodity WiFi for Fine-grained Finger Gesture Recognition. In *Proceedings of the 17th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc '16)*. 201–210.
[22] Guanhua Wang, Yongpan Zou, Zimu Zhou, Kaishun Wu, and Lionel M. Ni. 2014. We Can Hear You with Wi-Fi!. In *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking (MobiCom '14)*. 593–604.
[23] Ju Wang, Hongbo Jiang, Jie Xiong, Kyle Jamieson, Xiaojiang Chen, Dingyi Fang, and Binbin Xie. 2016. LiFS: Low Human-effort, Device-free Localization with Fine-grained Subcarrier Information. In *Proceedings of the 22Nd Annual International Conference on Mobile Computing and Networking (MobiCom '16)*. 243–256.
[24] Teng Wei, Shu Wang, Anfu Zhou, and Xinyu Zhang. 2015. Acoustic Eavesdropping Through Wireless Vibrometry. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking (MobiCom '15)*. 130–141.
[25] 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 (CCS '17)*. 103–117.
[26] 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 (CCS '17)*. 57–71.
[27] 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 (CCS '16)*. 1080–1091.