Anti-Spoofing Voice Commands: A Generic Wireless Assisted Design

CUI ZHAO, Xi'an Jiaotong University, China
ZHENJIANG LI*, City University of Hong Kong, China and CityU Shenzhen Research Institute, China
HAN DING, Xi'an Jiaotong University, China
WEI XI, Xi'an Jiaotong University, China
GE WANG, Xi'an Jiaotong University, China
JIZHONG ZHAO, Xi'an Jiaotong University, China

This paper presents an anti-spoofing design to verify whether a voice command is spoken by one live legal user, which supplements existing speech recognition systems and could enable new application potentials when many crucial voice commands need a higher-standard verification in applications. In the literature, verifying the liveness and legality of the command's speaker has been studied separately. However, to accept a voice command from a live legal user, prior solutions cannot be combined directly due to two reasons. First, previous methods have introduced various sensing channels for the liveness detection, while the safety of a sensing channel itself cannot be guaranteed. Second, a direct combination is also vulnerable when an attacker plays a recorded voice command from the legal user and mimics this user to speak the command simultaneously. In this paper, we introduce an anti-spoofing sensing channel to fulfill the design. More importantly, our design provides a generic interface to form the sensing channel, which is compatible to a variety of widely-used signals, including RFID, Wi-Fi and acoustic signals. This offers a flexibility to balance the system cost and verification requirement. We develop a prototype system with three versions by using these sensing signals. We conduct extensive experiments in six different real-world environments under a variety of settings to examine the effectiveness of our design.

CCS Concepts: • **Security and Privacy** → **Network security**; • **Networks** → Network services.

Additional Key Words and Phrases: Voice commands, wireless sensing, speaker verification

ACM Reference Format:
Cui Zhao, Zhenjiang Li, Han Ding, Wei Xi, Ge Wang, and Jizhong Zhao. 2021. Anti-Spoofing Voice Commands: A Generic Wireless Assisted Design. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 3, Article 139 (September 2021), 22 pages. https://doi.org/10.1145/3478116

1 INTRODUCTION

Voice-user interfaces (VUIs, *e.g.*, Apple Siri, Google Assistant, Amazon Alexa, *etc.*) is becoming increasingly popular in recent years, and more and more devices or objects could be controlled through VUIs in our current

*Corresponding author.

Authors' addresses: Cui Zhao, School of Cyber Science and Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, 710048, China, zhaocui@stu.xjtu.edu.cn; Zhenjiang Li, City University of Hong Kong, Hong Kong, China, CityU Shenzhen Research Institute, Shen Zhen, China, zhenjiang.li@cityu.edu.hk; Han Ding, School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China, dinghan@xjtu.edu.cn; Wei Xi, School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China, xiwei@xjtu.edu.cn; Ge Wang, School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China, gewang@xjtu.edu.cn; Jizhong Zhao, School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China, zjz@xjtu.edu.cn.

2474-9567/2021/9-ART139 $15.00
https://doi.org/10.1145/3478116

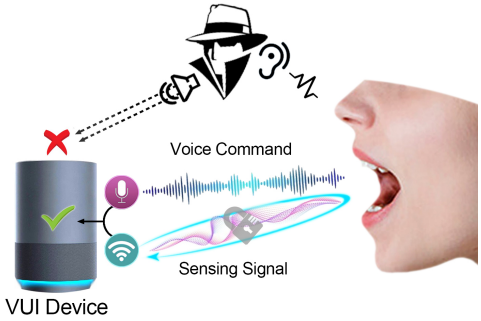


Fig. 1. Our design WISE can accept the voice commands from live legal users only. When the device is receiving the voice command, it also transmits sensing signals actively to recognize the speaker's mouth movement. With our sensing channel design, WISE is against replay attacks and compatible to a variety of signals to be used for the sensing.

daily life or the near future smart homes. However, when the number and importance of VUI-controlled devices are increasing, there is an urgent need of the ability to accept certain (crucial) voice commands from *live legal users* merely in many scenarios, especially for those that need a higher-standard verification. In other words, for some control targets that are crucial (*e.g.*, security camera, alarm system, door) or under a special control model, only the authorized (legal) users can execute commands to operate them in person (live). This verification capability, as a strong supplement to VUI's existing ability to recognize common voice commands, could benefit many useful applications in practice, including smart homes, entrance guard systems, automobiles, *etc.*

Verifying *live* and *legal* users these two aspects has been studied separately in the literature. The user's legality can be verified by extracting the user's representative voice features, *e.g.*, Mel Frequency Cepstrum Coefficient (MFCC) [38]. To verify a live user, also known as the *liveness detection* [33], some of the existing methods [20, 23, 48] propose to ask the VUI device to transmit certain signals to form a sensing channel, which can capture the user's body movement to defend the attack of replaying the legal user's voice. However, to accept a voice command from a live legal user, we cannot combine the prior solutions directly, because there are some overlooked and unsolved challenges when these two aspects are considered jointly in one system.

1) First, the safety of the sensing channel is ignored in previous designs. However, the recent advances in the wireless domain have revealed that wireless and acoustic channels are not safe completely, which may suffer the replay attack as well [34]. Hence, once the attacker replays both the voice command and sensing-channel signal simultaneously, the previous methods will be compromised and the overall verification becomes untrustworthy.

2) Even the safety of the speaker's liveness could be guaranteed, a direct combination of the user's legality verification and liveness detection is still inadequate — the attacker can play a recorded voice command from the legal user and mimic this user speaking the command (without sound) simultaneously to fool the system, because the existing sensing based liveness detection methods only ensure that a person is indeed speaking (instead of a recorded audio), but they do not consider the identity of the current speaker. Therefore, we need to further consider the user's legality and liveness jointly in the system design, which has been rarely studied yet.

This paper introduces WISE to address above two challenges. WISE contains an anti-spoofing sensing channel. Users can pre-define a set of crucial keywords (*e.g.*, involving important control devices or under special security configurations) and register them with the WISE service. After receiving a voice command, VUI conducts the ordinary speech recognition [45] first and temporarily records both the voice command and sensing-channel signals. Only when the recognized command contains the pre-defined keyword(s), VUI further launches the WISE service for the live legal user verification before the command could be executed. The verification can be

finished in the mill-second level (§5.1), which does not cause any perceptual latency to the user. WISE could defend many attacks to the speech recognition systems, including impersonation attacks, voice replay attacks, dual-channel replay attacks, *etc.* To enable such a design, we propose the following techniques.

1) We introduce a transceiver sheltering technique to ensure the safety of the sensing channel. The key idea is to add some random noise to "pollute" the sensing signals to sense the user's mouth movement. The random noise is known by the VUI device only, and it can remove the noise from the received signal reflected from the user. The polluted signals can still be used for the liveness detection, while it can hardly be replayed by an attacker because the noise is random each time. In addition, we also provide a *generic* interface to form this sensing channel, which is compatible to a variety of widely-used wireless signals including RFID, Wi-Fi and acoustic signals. We analyze their advantages and disadvantages, offering a flexibility for manufactures to enable the WISE service by balancing the system cost and the verification requirement.

2) To further verify the user's legality and liveness jointly, we introduce a dual-channel signal processing pipeline and address the following technical issues. As the mouth movement is tiny when the user is speaking, we process the voice command and sensing signal simultaneously, and identify the most representative parts from both signals for the verification. After the signal processing, we further introduce a neural network based design to fulfill the verification. The network is designed carefully to ensure a high-accuracy verification.

To validate the efficacy of our design, we develop a WISE prototype using three types of sensing-channel signals, including RFID, Wi-Fi and inaudible sound (near 20 kHz). All these sensing signals are transmitted and received by commercial devices in our experiments.

- Overall, WISE achieves over 94% average verification accuracy and less than 4% false acceptance rate to defend various attacks.
- WISE remains a similar working distance as the state-of-the-art sensing-based liveness detection method, *e.g.*, 20 cm, while WISE can ensure the safety of the sensing channel as well.

In summary, we make the following contributions in this paper. First, we propose a generic wireless-assisted framework to accept voice commands from live legal users, which could enable many useful application potentials in practice. Second, we propose a series of techniques in the WISE, which can achieve an anti-spoofing sensing-channel design and verify the user's legality and liveness jointly. Finally, we develop a WISE prototype with three sensing-channel signals. We conduct extensive real-world experiments in six different environments to evaluate its performance.

2 BACKGROUND AND OVERVIEW

WISE is positioned as a supplement to the recognition of voice commands in existing VUI devices, instead of replacing this function.

2.1 Applications

With WISE, we could enable (at least) the following application potentials in practice:

1) Access control of devices at home. As a convenient interface, more and more devices and objects could be controlled by the VUI device at home in the future. To achieve an intelligent smart-home management, different people should have different permissions to access in-home devices. For instance, some devices cannot be accessed by children when their parents are not at home, such as TV, play station, microwave oven, *etc.* Security cameras and alarm systems should not be accessed by the elder people for the safety concern. Most of devices cannot be accessed by the visitors. WISE can enable such an access control design in smart homes.

2) Entrance guard system. Many entrance guard systems nowadays mainly use passwords or membership cards to verify a user's identity, such as in residential buildings, clubs, gyms, *etc.* It could cause non-negligible

management efforts, *e.g.*, to periodically update passwords and notify users, to handle the lost or broken membership cards, *etc.* On the contrary, WISE provides an efficient alternative to avoid such issues. Moreover, the design with WISE is also *contactless*, which could be more convenient (*e.g.*, when user's hands are occupied by bags) and healthy (*e.g.*, users do not need to touch the panel to input password for avoiding the virus infection).

3) Automobiles. VUI also becomes increasingly popular in automobiles due to the hand-free control, and many functions could be triggered through such a convenient interface. However, for the safety concern, most commands should be launched by the driver directly. WISE can help avoid that other people in the car may mis-trigger some inappropriate commands. Moreover, recent studies [4, 52] unveil the possibility of a potential security risk for the in-car VUI suffering the audio adversarial attacks. WISE can help avoid such attacks as well.

Compared with the traditional biometric methods, *e.g.*, fingerprints and face recognition, WISE is a contactless and privacy-preserved alternative which requires no extra user cooperation, *i.e.*, just speaking the voice command towards the VUI device.

2.2 System Overview

WISE provides a software-based pipeline for the voice command and sensing signal, as depicted in Figure 2 with three main modules.

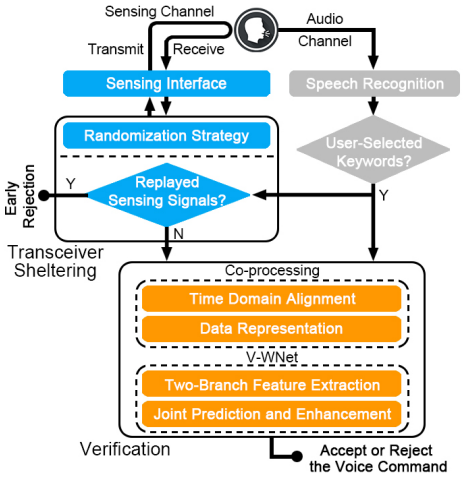


Fig. 2. Overview of the WISE design.

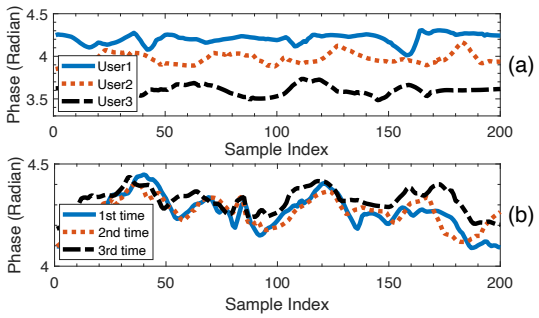


Fig. 3. The impact of the user's mouth movement on the RFID signals, when (a) three users speak the same voice command, and when (b) one user speaks a command three times.

1) Sensing interface. The sensing channel aims to defend the replay attack on the voice command (*e.g.*, using a recorded voice command to fool the system), and WISE has a generic interface to cooperate with different wireless modules for constructing the sensing channel. It can support the inaudible sound (*e.g.*, near 20 kHz) using the default speaker and microphone from the VUI device directly. It can also leverage an extra wireless module (*e.g.*, RFID or Wi-Fi) to augment the verification performance.

2) Transceiver sheltering. This module further prevents the sensing channel from the replay attack, which utilizes the one-time random noise to "pollute" the sensing signal for the anti-spoofing purpose. According to the communication characteristics, we design tailored randomization strategies for each type of sensing signals. Once a replay attack is detected, an "Early Rejection" will be given to the current voice command; Otherwise, both the voice command and sensing signal will be further processed next.

3) Verification. This module checks both whether the received voice command is spoken by a person and whether this live user is a legal user. To this end, this module first co-processes the voice command and sensing signal, and identify the most representative features from both signals. We then introduce a neural network based design to fulfill the verification.

2.3 Sensing Feasibility Study

In WISE, we use the sensing signal to capture the user's mouth movement for the liveness detection. According to the verification module introduced above, the sensing signal also needs to distinguish that this live user is a legal user. Prior studies [19] find that the vocal tracts, lips and tongues vary among people, and different people have distinguishable yet consistent ways of the mouth movement when they are speaking [35]. This offers a possibility to use wireless signals to fulfill our verification design. To unveil this potential, we conduct a proof-of-concept experiment in Figure 3, which shows the result using the RFID signal and similar results can be also obtained using Wi-Fi and acoustic signals.

1) Distinguishability. We use the *phase* information from the RFID signal to capture the user's mouth movement during the speaking. Fig. 3(a) shows the phase value changes when three volunteers say the same voice command towards an RFID tag (the experimental setting is detailed in §4). We can see that the impacts from different users on the received RFID signals are significantly different and highly distinguishable.

2) Consistency. Fig. 3(b) further plots the phase values when one volunteer speaks the same voice command three times. The result suggests that the impact from the same person on the received sensing signal keeps relatively consistent. As they are not exactly the same each time, we still need to propose effective techniques to achieve a consistent and reliable verification.

In summary, Fig. 3 reveals that the RFID signal has a great potential to fulfill the sensing channel design in WISE. We find that Wi-Fi and acoustic signals have this potential as well, and we will leverage such sensing abilities to design WISE in §3.

2.4 Threat Model

WISE can defend the attack that attempts to conduct unauthorized operations on important control targets through the legal user's VUI device. To this end, we consider the following threat models.

1) Impersonation attack. When a legal user is speaking a voice command, it could be overheard by an attacker nearby, who may conduct an impersonation attack by mimicking the voice and the style of the user's speaking.

2) Voice replay attack. Attacker can record a crucial voice command spoken by the legal user in advance. The attacker then replays the recorded voice command and mimics the legal user's mouth movement (without sound) at the same time to fool the system.

3) Dual-channel replay attack. We further consider a more advanced attacker, who can replay both the recorded voice command and the eavesdropped corresponding sensing signal simultaneously.

Attacker can launch the attack at a common working distance to the VUI device (as the legal user), so that the system rejects the attacker's request because of our effective system design, instead of the relatively weak signals received when the attacker is too far away from the system.

3 SYSTEM DESIGN

We elaborate the design of the *transceiver sheltering* and *verification* two main technical modules in this section.

3.1 Transceiver Sheltering

As introduced in §2.2, the sensing signal can be used for the live user detection, while the sensing signal itself can be replayed as well. An advanced attacker can thus launch a dual-channel replay attack (for both the voice

command and the sensing signal) to attack the system. To cope with this issue, we introduce a transceiver sheltering module to form an anti-spoofing sensing channel.

3.1.1 Principle. The sensing signal contains some "credential" information (the content of this information is stated later), and the transceiver sheltering module adds the *one-time random* noise to the sensing signal to *hide* its credential information. The added noise is known by the sensing interface only. Therefore, the device can add the noise before the transmission and later cancel the noise from the received signal to recover the credential information. On the contrary, the replayed sensing signal (eavesdropped in advance) is invalid, since the one-time random noise is added differently each time. Only the sensing signal with the credential information recovered correctly is viewed to be safe, which will be used for the following liveness detection. *We note that our design mainly relies on the correct decoding of the credential information, instead of its content itself.*

Therefore, when the recognized voice command contains crucial keywords predefined by the user, the *transceiver sheltering* module will be triggered and accesses the temporarily stored sensing signal received during the period of this voice command. The credential information can be obtained multiple times during this period. If the error rate (to recover the credential information) is large, *e.g.*, much higher than the ordinary communication error rate, WISE will output an "Early Rejection" to prevent the execution of the current voice command for the security and system overhead considerations; Otherwise, the *verification module* will be functioned to further check whether the command is from a live legal user (§3.2).

Next, we elaborate how to extend this principle in a practical system by using different wireless signals.

3.1.2 Transceiver Sheltering with RFID. The RFID protocol involves two standard parties: the RFID tag and reader. The RFID-enabled WISE can be suitable for the entrance guard system, wherein the tag is installed close to the microphone and the reader is deployed nearby on the wall. We adopt two passive tags which are battery-free. The reader queries two tags periodically following the EPC Gen 2 protocol [8]. Every period is named as an *inventory round*, containing several *inventories*. In each inventory, one tags will reply and each inventory lasts milliseconds.

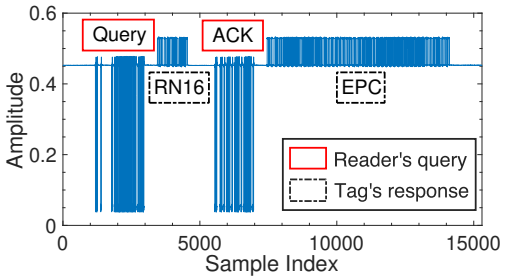


Fig. 4. Illustration of the reader's query and the tag's response in one inventory.

EPC as credential information. Fig. 4 shows the baseband process of the reader's query and the tag's response in one inventory. Throughout the entire inventory, the reader keeps transmitting the Continuous Wave (CW).[1] In the first phase of the inventory, the reader modulates the "Query" bits into CW to send a "Query" command. By receiving the energy from the reader's signal, the tag is activated and then replies (*i.e.*, backscatters) an 16-bit binary sequence (RN16) to notify the reader about its existence. Next, the reader transmits an ACK command appended with the RN16 to request the tag to reply its identity which is denoted as EPC (Electronic Product Code). The tag finally replies its EPC to terminate this inventory. After the inventory, the reader can decode

[1]In RFID communications, the carrier wave is a periodical signal, while the continuous wave is the wave with a constant amplitude value to be modulated to the carrier wave for providing sufficient energy to activate tags only.

the tag's EPC from the demodulated signal. Because the EPC of each tag can be configured in the system, we can thus use EPC as the credential information for each tag in WISE. Moreover, if a user is speaking during this communication process, the impact of the user's mouth movement can be unveiled from the received signal's amplitude and phase values. Within the duration of the user's speaking, we can collect adequate signals from a series of consecutive inventories. For sensing, the varying of signal's amplitude and phase matters, based on which we can complete the verification design (§3.2).

Adding one-time random noise. In the EPC Gen 2 protocol, the EPC is transmitted in plain text. The attacker can thus eavesdrop the communication between the reader and tag in advance to obtain the tag's EPC information. On the other hand, as the tag harvests energy from the reader's transmitted signal to power its reply, the tag cannot conduct complicated computations [5], *e.g.*, encryption. We thus propose to add the one-time Gaussian random noise (both its average and variance equal to one) to the CW at the reader side, which is generated in the baseband from the GNU Radio Reader Block [24] to replace the traditional continuous wave in the transmission. The duration of the random noise will fully cover the period of the EPC reply from the tag, *i.e.*, EPC will be hidden in the random noise. After tag is activated by the reader's transmitted signal, it will modulate its EPC into the noisy CW. Fig. 5(a) shows the effect of utilizing the noisy CW to hide the EPC credential in the time domain. For the illustration purpose, we only generate the noisy CW to cover the first half of the EPC. We find that the EPC bits hidden inside the noise cannot be decoded by the reader unless the noise can be cancelled precisely. In the frequency domain, the EPC is also hidden well, *i.e.*, we cannot specify the spectrum of the clean EPC (Fig. 5(c)) from the spectrum of the noisy EPC Fig. 5(d).

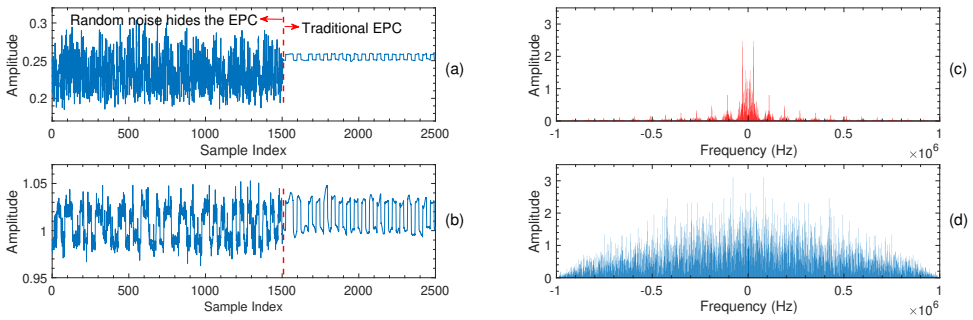


Fig. 5. (a) One-time random noise is added to hide the EPC of one tag's reply; (b) The recovered signal with our proposed channel estimation design; The spectrum of the clean (c) and noisy (d) EPC replies.

Recovering the EPC. Although the reader knows the one-time random noise (noisy CW) added each time, the challenge is that we cannot deduct it from the received signal directly because the noise will be distorted by the wireless channel during the communication. To tackle this issue, we propose to conduct a lightweight channel estimation before the noise removal. To this end, we add one more inventory at the beginning of each inventory round, and tags will reply nothing in this inventory. Therefore, the EPC field contains the random noise only, based on which the reader can estimate the channel. The reader performs the channel estimation once in each inventory round. Because the duration of the entire inventory round is short (*e.g.*, around tens of milliseconds, which depends on the number of tags within the reader's communication range), we thus utilize this channel estimation result for other inventories in the same inventory round for the noise removal. In the inventory for the channel estimation, we denote the original and the received noise samples as N_i^t and N_i^r respectively, where i is the sample index. We can obtain a series of channel distortion ratios $H_i = N_i^r / N_i^t$. Then, for each inventory j in the same inventory round, after the reader receives the EPC field samples $\hat{N}_i^r(j)$ (we add a hat for N here

to distinguish from the pure noise since this noise is hiding the EPC information), it can recover the EPC-field signal by:

$$EPC_i(j) \quad = \quad \hat{N}_i^r(j)/(H_i \cdot N_i^t(j)). \tag{1}$$

Fig. 5(b) shows the recovered EPC signal from Fig. 5(a) with the user's mouth movement. From the result, we can see that the recovered EPC-field signal after the noise removal (the left-hand part) is a square-like wave, which has a similar wave format as the original EPC-field signal (the right-hand part). When we feed such a recovered signal to the reader, we can obtain the correct EPC. Through our experiment, we find that the error rate (due to communication) to recover EPC by the legal device is very small, *e.g.*, less than 0.3%. Therefore, we adopt a relatively conservative error rate threshold 3% in our implementation. If the percentage of the EPCs' recovering error rate is higher than such a threshold, this voice command will be early rejected.

Even with above design, RFID tags can still be sniffed by other readers to get their EPCs (because they are transmitted in plain text), which however will not compromise our system. The major reason is that the transceiver sheltering proposed above replies on whether the EPC can be decoded correctly by the reader each time, instead of any pre-knowledge of the EPC's content. Hence, whether tags' EPCs are known by the attacker or not does not impact the verification in WISE. This is because the reader will add the *one-time* noises to the EPC field in each inventory, and only the reader knows what to be subtracted from the received signal. As a result, replaying the EPC alone will not work (since the noises are missing), and any previously recorded EPCs together with the noises will hardly cause a successful decoding neither for the future inventory at the reader side (since the noises are one-time noises). Another possible attack is that the attacker replays the EPC information only and also masks the system's RFID tags, this in principle has a chance to pass the transceiver sheltering, but it requires to tightly synchronize the reader's signal and the attacker's signal in the air, which is non-trivial to be achieved in practice. Moreover, even the transceiver sheltering could be passed, the attacker still cannot pass the network verification (Section 3.2), as this synthetic sensing signal does not contain the feature of any legal user.

3.1.3 Transceiver Sheltering with Wi-Fi. Wi-Fi is another popular wireless standard in our daily life. WISE enabled by Wi-Fi can be suitable for the voice assistant devices at home. Nowadays, many VUI devices are equipped with the Wi-Fi NICs already. We envision that in the near future, if a pair of NICs are added to the VUI device or the full-duplex NIC is equipped, we can leverage Wi-Fi to form the sensing channel for WISE.[2]

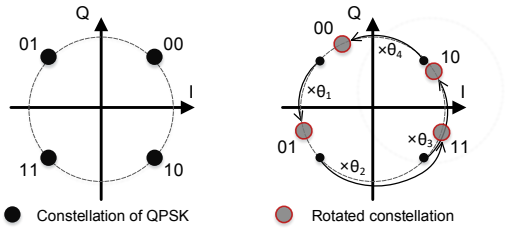
Sensing packet as credential information. Wi-Fi follows the 802.11 standard, wherein bits are organized as packets to deliver over the Wi-Fi signal. Each packet includes packet preamble, data bits and cyclic redundancy check (CRC) three major fields. With the random noise-adding mechanism stated next, only the legal device can remove the added noise from the received packet to recover the transmitted bits. Therefore, we leverage the packet's data field (regardless its content) as the credential information — Only when the decoded data bits match the decoded CRC field, this packet can be viewed as a valid sensing packet.

Adding one-time random noise. Due to the physical layer and communication protocol differences, the noise-adding design proposed for RFID cannot be adopted here. Therefore, we propose a constellation masking (CM) mechanism for the Wi-Fi signals, *i.e.*, the noises are generated by modifying the constellation diagram. Fig. 6(a) shows four points on the I-Q plane (with the QPSK modulation) to represent "00"∼"11", respectively. The position of each point on the I-Q plane is pre-defined. With OFDM (Orthogonal Frequency Division Multiplexing) [27] technology, packet bits are converted to the OFDM symbols before the transmission. The main idea of the CM mechanism is to rotate a random degree on constellation for each packet before the transmission, and the added noise is a random re-mapping from the constellation points to these pre-defined positions on the I-Q plane. However, to ensure the detection of the incoming packet at the receiver, we skip the preamble field and rotate

[2]In principle, by receiving the Wi-Fi signal from another Wi-Fi device nearby, we can also fulfill the sensing design. However, when the sender and receiver are located on one device, we can avoid setting up an extra secure channel to exchange the constellation rotation factors.

symbols for the rest data bits and CRC two fields only. In particular, we denote the original symbols from these two fields in one packet as S. The rotated version of S used in the transmission is $\hat{S} = S \times \theta$, where θ denotes the one-time random re-mapping from the constellation points to these pre-defined positions on the I-Q plane.

Decoding the packet bits. With the CM design, only the legal device can rotate them back on the I-Q plane before the packet decoding. Fig. 6(b) shows a rotation example. All the points are rotated to other quadrants, which will cause the decoding error if such a rotation is not compensated before the decoding. Through our study, we find that the error rate to decode packet bits by the legal device is less than 0.2%. Therefore, we adopt 2% as the error rate threshold in our current Wi-Fi based design. We note that the CM mechanism can be regarded as a simple yet efficient symmetrical encryption (yet without introducing any extra bits to each packet) and the rotation factors θs serve as the key. Therefore, we do not require the content of the packet. As long as the decoded bits can match the CRC, we view such a sensing packet to be valid. Such a mechanism is transparent to users as it does not impact the Wi-Fi communication.



(a) Original constellation (b) Rotation example

Fig. 6. Illustration of the constellation masking design.

Table 1. Key parameters for acoustic-based transmissions.

Baseband/MAC Parameters	Settings/Values
Bandwidth	4 KHz
Carrier frequency	19 KHz
Number of Subcarriers	128
OFDM symbol length	160
Cyclic prefix	32
Subcarrier modulation	QPSK
Coding scheme	RS code 17/32
Decoding scheme	Viterbi-decoding

In WISE, the symbols in one packet are rotated following the same random mapping. Therefore, the probability to crack the rotation masking for one packet is not very small usually, *e.g.*, $\frac{1}{4!} = 4.17\%$ for QPSK. However, when a user speaks one voice command, there will be 1000 packets received per second and different packets employ different mappings. For example, even for a short voice command with one second merely, the cracking probability, *e.g.*, for QPSK, $\sum_{i=980}^{1000}(\frac{1}{4!})^i < 20 \times (\frac{1}{4!})^{980} < 10^2 \times (\frac{1}{10})^{980} < 10^{-978}$, becomes very small.

3.1.4 Transceiver Sheltering with Acoustic Signal. Recent studies have revealed the sensing ability by using acoustic signals [3, 50, 51], and many smart devices have the build-in speaker and microphone, which can form the Tx-Rx link to fulfill the sensing design. Hence, the acoustic signal enabled WISE can be suitable for many personal devices, *e.g.*, smart phone, laptop, voice assistant, *etc.* For the acoustic signal enabled transceiver sheltering module, the speaker of the device will transmit dedicated acoustic signals and its microphone receives the transmitted signals. When a user is speaking, the received acoustic signals (reflected from the user's mouth) contain the unique features that can be used for sensing.

To enable the credential information delivery over acoustic signals, we employ the OFDM physical layer design at the upper edge of the frequency range of the acoustic signal that most personal devices can support, *e.g.*, 19 KHz. With this configuration, the transmission is robust to multi-path for reliably delivering credential information.The key parameters are listed in Table 1 and more development details are introduced in §4. Because the constellation masking design for the acoustic signal still follows the 802.11 standard, it can be transplanted from the Wi-Fi based WISE directly, *i.e.*, only the legal device can recover the rotated constellation symbols and

decode the packets successfully. The error rate threshold is set to 5% (vs. less than 0.5% error rate due to the communication) in our current implementation.

3.2 Verification

If the sensing signals are not early rejected by the transceiver sheltering module, the verification module then can verify the live legal user. However, if the attacker plays a recorded voice command from the legal user and mimics this user's mouth movement (without sound) simultaneously, the system can still be compromised. To verify the current user is indeed a legal user, we should leverage the unique features of this user from both the voice command and sensing signal. Hence, we introduce a dual-channel pipeline to pre-process both signals first.

3.2.1 Dual-channel Signal Pre-processing. When a legal user is speaking, the WISE device will receive the voice command through microphone, and it can also receive the sensing signal reflected from the user's mouth. The human voice is a well-studied signal and we can easily extract the unique features from the user's voice. Moreover, the recent advances in the wireless sensing studies also reveal the possibilities to obtain the representative features from the user's movement [20, 23, 48]. Hence, the first task is to identify the part in the sensing signal that corresponds to the user's speaking.

To this end, the opportunity in WISE is that the voice command and the sensing signal are received simultaneously. Because microphone receives a voice command only when the user is speaking, we can thus use the boundaries of the voice command's content to segment the sensing signal. After the speaking parts from two signals are identified, they can serve as the input of the neural network for verification in §3.2.2.

1) Voice command segmentation. We first detect the boundaries of the voice command's content. To process audio signals, the signal is usually divided into frames, where the frame duration is usually 5 to 40 ms [22]. For each frame i, its power $P_i = 10 \times \lg V_i$, where V_i indicates the amplitude variance of frame i. Next, the Zero-Crossing Rate (ZCR) [2] of this frame can be calculated as:

$$\mathrm{ZCR}_i = \frac{1}{2N} \sum_{n=1}^{N} |sgn[x_i(n)] - sgn[x_i(n-1)]|, \tag{2}$$

where N is the number of the samples in the frame and $sgn(\cdot)$ is the sign function, *i.e.*,

$$sgn[x_i(n)] = \begin{cases} 1, & x_i(n) \geq 0, \\ -1, & x_i(n) < 0. \end{cases} \tag{3}$$

The ZCR value of one voice frame represents the rate of the sign's changes (*i.e.*, 1 or -1) cross all the samples of this frame. In particular, when the voice signal contains only the background noise, the signs of the samples of each frame will cross zero frequently, which lead to a larger ZCR value. On the contrary, the samples from one frame tend to have much more consistent signs for the human voice, which lead to a much lower ZCR value. Fig. 7(a) illustrates one example, where the start-end boundaries of the segmented voice command content are plotted in red dotted lines. On the other hand, the voice content part usually has much higher energy compared with the background noise part. Hence, based on these two observations, we can reliably segment the content part from the voice command with the proper ZCR and energy thresholds.

2) Time-aligned sensing signal. With the boundaries for the user's speaking detected from the voice command, we can use them to segment the concurrent sensing signal as well. Fig. 7(b) shows an example to segment the received RFID sensing signal.[3] Afterwards, the segmented voice command and sensing signal will serve as the input of the verification design.

[3]The RFID signal in Fig. 7(b) is filtered by WTD (Wavelet Thresholding Denoising) [9] for removing the high-frequency noise.

Before we elaborate the verification design, one natural question is why not use the entire sensing signal in Fig. 7(b) for verification? Through our investigation, we find that before and after the part aligned with the voice command's content, the sensing signal could be impacted by the user's other activities, such as swallowing or licking lips before/after speaking the voice command. However, these activities are not necessarily occur and follow the same order each time, which may impact the system performance.

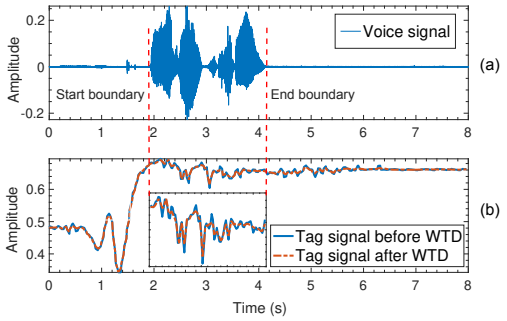


Fig. 7. Time-aligned segmentation for the sensing signal.

Algorithm 1: Sensing signal warping

Require: Current sequence: $S_n, n \in [1, N]$;
Target length: M
Ensure: Warped sequence: $R_m, m \in [1, M]$
$\alpha \leftarrow N/M$
$d \leftarrow [1, 2, ..., M]$
$d \leftarrow d * \alpha$
$d_{int} \leftarrow floor(d)$
if $N < M$ **then**
  $d_{dec} \leftarrow d - d_{int}$
  $S_{diff} \leftarrow S_j - S_i, i \in [1, N-1], j = i + 1$
  $R_{int} = S[d_{int}]$
  $R_{dec} = S_{diff}[d_{int}] . * d_{dec}$
  $R = R_{int} + R_{dec}$
else if $N >= M$ **then**
  $R = mean(S[d_{int}], S[d_{int} + 1])$
end if

3.2.2 Design of V-WNet. Fig. 8 illustrates our neural network design for verification, which is named as V-WNet. For both the voice and sensing inputs, we first convert them to another representation that facilitates the neural network's processing. Then, for each input branch, we utilize neural networks to extract their user-specific features, and the neural networks are designed differently to be tailored for the voice and sensing signals. In addition, we consider a practical constraint with an additional loss function design to enhance the V-WNet's overall performance. Finally, we employ the fully-connected layer to output the final decision about whether the current user is a live legal one.

1) Input representation. To better unveil the signal features and facilitate the neural network's processing, we convert the pre-processed voice and sensing signals to another representation first:

Voice spectrogram. We convert the segmented voice command to its spectrogram prior to apply the neural network. In the literature, there are four popular hand-crafted features to describe the sound/voice characteristics [26, 36], including linear predictive coding (LPC), linear prediction cepstrum coefficient (LPCC), perceptual linear prediction (PLP), and Mel frequency cepstral coefficient (MFCC). Recently, neural networks show another possibility to automatically extract effective features from the audio sound's spectrogram according to the training data collected from applications [40, 42, 47], because the user's voice features are mainly distinguished in the frequency domain [25] and also exhibit certain temporal dependence. Spectrogram can take care of both aspects. Our design leverages this recent trend and we introduce a neural network to enable an anti-spoofing approach. In particular, we employ a sliding window (20 ms) to divide the voice command into frames and each frame has 50% content overlap with the previous one. Next, a Hamming window is applied to each frame, based on which we can convert the time-domain frames to the spectrogram by Fast Fourier Transform (FFT). With our current implementation, every 200×200 spectrogram serves as the input of the following neural network. The most-left

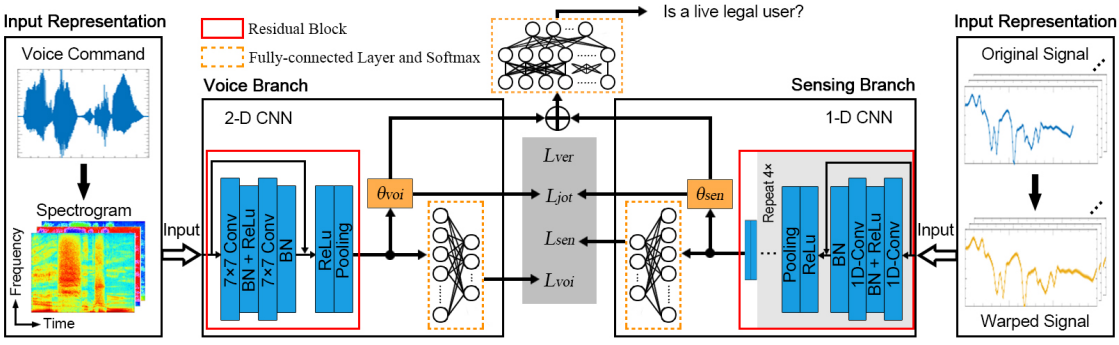


Fig. 8. Overall architecture of our neural network design, which is named as V-WNet in WISE.

block in Fig. 8 shows the input of spectrograms (including RGB channels) from the current voice command, wherein the x-axis and y-axis of one spectrogram present the time and frequency dimension respectively, and the color indicates the power at a particular frequency and time.

Sensing signal warping. On the other hand, the user's mouth motion features are captured by the sensing signal. However, the user cannot ensure that the speaking speed of a voice command is exactly the same, leading to a fact that the length of the received sensing signal can be different each time. We thus apply a discrete sequence warping approach, as shown in Algorithm 1, to unify the signal length to match the input size of the consequent neural network. Given the target length of the sensing signal defined in the network, we first calculate a warping factor α (line 1). Then (in lines 2-4), we align each index of the target (warped) sequence to that of the current sequence using α and obtain the aligned index d_{int}. If the target length is larger than the actual length (lines 5-10), we re-scale the differences between adjacent samples with a step d_{dec} (line 6) to fulfill the warping operation. The warped sequence can be also obtained by calculating the mean values of the adjacent (in terms of d_{int}) samples if the target length is shorter than the actual length (lines 11-13). The most-right block in Fig. 8 shows one warping example. The length of the current signal is 123, which is warped to the target length of 150. The target length is configured as 400 (for RFID), 900 (for Wi-Fi) and 60 (for acoustic signal) in the current WISE.

2) Neural network design. There are two neural network branches in V-WNet, including the voice branch and sensing branch. Moreover, we adopt the residual structure [14] to form two branches (voice branch and sensing branch) in our network to avoid over-fitting.

Voice command branch takes 2D spectrograms (200×200) as input, and processes it by the 2D-CNN layers. In particular, the voice branch contains one residual block with two 2D convolutional layers (64 kernels of size 7×7 and 3 input channels for RGB), followed by the Batch Normalization (normalizing the mean and variance of the input data), ReLu (introducing non-linearity) and Pooling (reducing the size of features) to extract the user's identity feature. We also append one Fully-connected and Soft-Max layer for training only. With this branch, we can obtain one loss function as:

$$L_{voi} = -\frac{1}{n}\sum_{i=1}^{n} y_{voi}^{i}\log(\hat{y}_{voi}^{i}), \tag{4}$$

where n is the number of input spectrogram i, $\hat{y}_{voi}^{i}$ indicates the predicted user's identity in the voice branch and y_{voi}^{i} is the user's identity label. We denote the feature extracted from this branch as θ_{voi}.

Sensing signal branch adopts the 1D-CNN [18] layers to process the sensing signal input, because the warped signal segment is the 1D temporal sequence. The sensing branch contains four residual blocks with different numbers and sizes of the kernels for different sensing signals, which are also followed by Batch Normalization, ReLu and Pooling. Specifically, 64, 128, 256, and 512 kernels with size of 7 for RFID (2 channels), 9 for Wi-Fi (10

channels) and 3 for acoustic inputs (10 channels), where the number of channels equals to the number of tags or subcarriers of the sensing signal. Similarly as the voice branch, this branch also leads to the loss function:

$$L_{sen} = -\frac{1}{n}\sum_{i=1}^{n} y_{sen}^{i}\log(\hat{y}_{sen}^{i}), \tag{5}$$

where $\hat{y}_{sen}^{i}$ is the predicted user's identity by the Fully-connected layers of this branch and y_{sen}^{i} is the user's identity label. We denote the feature extracted from this branch as θ_{sen}.

For θ_{voi} and θ_{sen}, we can further concatenate them, and employ two Fully-connected layers to fulfill the user verification, which leads to the following loss function:

$$L_{ver} = -\frac{1}{n}\sum_{i=1}^{n} y_{ver}^{i}\log(\hat{y}_{ver}^{i}), \tag{6}$$

where $\hat{y}_{ver}^{i}$ is the predicted user's identity of the V-WNet's final output and y_{ver}^{i} is the user's identity label.

With the two basic neural network branches, we further introduce an enhancement to improve the final performance of V-WNet, which is to jointly consider the features extracted from both branches (*i.e.*, θ_{voi} and θ_{sen}) in the final verification. Since these features refer to the same user, we expect them should preserve certain "similarity" in some dimension and the norm-2 operation is a feasible option [46]. In V-WNet, we thus further introduce the following loss function:

$$L_{jot} = \|\theta_{sen} - \theta_{voi}\|_2^2. \tag{7}$$

Combining all the loss functions above, we can obtain the final loss function used to train V-WNet.

In WISE, we design such a neural network mainly because our system takes the voice-command and sensing-channel two different types of signals as input. We find that there is no existing network that is dedicated to match our input formats and consider their characteristics jointly.

4 IMPLEMENTATION

In this section, we introduce the implementation by using RFID, Wi-Fi and inaudible sound three types of sensing signals. The usage diagram of WISE is illustrated through Fig. 9(a).

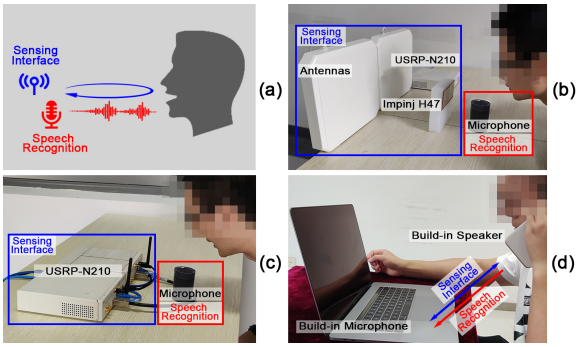
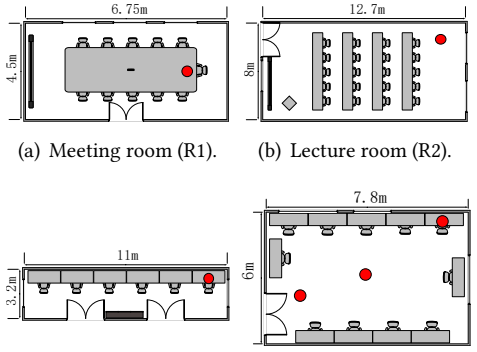


Fig. 9. Experimental setups: (a) Diagram, (b) RFID, (c) Wi-Fi and (d) Inaudible sound.



(a) Meeting room (R1). (b) Lecture room (R2).

(c) Office-one (R3). (d) Office-two (R41 to R43).

Fig. 10. Floor plans of the experimental rooms, wherein the red dots denote the device positions.

1) Implementation with RFID. The RFID reader is developed using a USRP N210 software-defined radio with the SBX daughter-board. The transmission frequency and the baseband sample rate are set to be 922.375 MHz

(25dBm) and 2 MHz respectively, making the RFID communication within the ISM band. Two directional antennas (Laird S9028PCR with the gain of 8 dBi) are installed to transmit and receive the RFID signals. The reader runs the EPC-Gen2 protocol to identify RFID tags with the standard FM0 coding and the 40 kHz BLF (backscatter link frequency). As Fig. 9(b) shows, we use one commercial speaker-microphone device (Xiaomi Cannon 2) to receive the voice commands, and one tag array to sense the user's mouth movement by computing the phase values of each replied EPC. The number of tags in the array is selected with the following two considerations:

To capture a comprehensive mouth movement, the size of the tag array is preferred to be comparable to or greater than that of the user's opened mouth. According to the surveys conducted in different countries, the average size of an opening mouth is about 50 mm (diameter of a circle). In WISE, we adopt the Impinj H47 tags (about 44 mm × 44 mm for each). Therefore, we need at least two tags (outperforming one tag as shown in Fig. 15) to form an array in the implementation.

The reader can query one tag around 300∼400 times per second in practice. In an array, each tag will be read less than $\frac{400}{n}$ times (n is the number of tags) because of the collision. Therefore, to ensure a sufficient reading rate, we adopt $n = 2$ without further increasing the tag number in our current implementation.

2) Implementation with Wi-Fi. We have utilized two N210 USRP software-defined radios with the UBX daughter-boards to setup the Wi-Fi sensing channel (following the 802.11n standard and adopting 3dBi omni-directional antennas) in the experiments. The sensing is conducted based on the amplitude of channel state information (CSI). The Wi-Fi transmissions are configured at the 5 GHz central frequency with the 20 MHz bandwidth divided into 64 subcarriers. The packet transmission rate is about 1000 packets per second with a 1000 ns interval between two packets. The voice commands are still received by the commercial Xiaomi Cannon 2 device, as shown in Fig. 9(c).

3) Implementation with inaudible sounds. We adopt the build-in loudspeaker of a commercial smart phone (iPhone X) to transmit the inaudible sound (as sensing signal), and receive both the voice command and the sensing signal by using the build-in microphone of a laptop (MacBook Pro), as Fig. 9(d) depicts. The sensing signal is transmitted at 19 kHz and the audio sampling rate is set to be 44.1 kHz. At the receiver side, the voice command can be separated from the sensing signal by a low-pass filter because the frequency of the human's voice is usually less than 3 kHz [43]. On this test bed, we further implement an end-to-end acoustic-based transmission protocol. The transmitted acoustic sensing signals from the smart phone employ the QAM-based modulation with OFDM (with 128 sub-carriers) and the constellation masking based transceiver sheltering design.

All the three test beds share a common back-end design. The back-end conducts the early-rejection decision based on the received sensing signal. Moreover, we also implement V-WNet on the back-end using PyTorch and employ the adaptive moment estimation (Adam) algorithm to train the network. The training of each V-WNet network takes no more than 10 minutes using a desktop with i7-9700 CPU and Nvidia RTX 2080 GPU. Different test-beds adopt the same network architecture, while we train three versions of V-WNet with different parameters to match the input differences for each type of the sensing signals. We note that in practice, we only need to train the network for the sensing signal adopted in the system, instead of always three versions.

5 EVALUATION

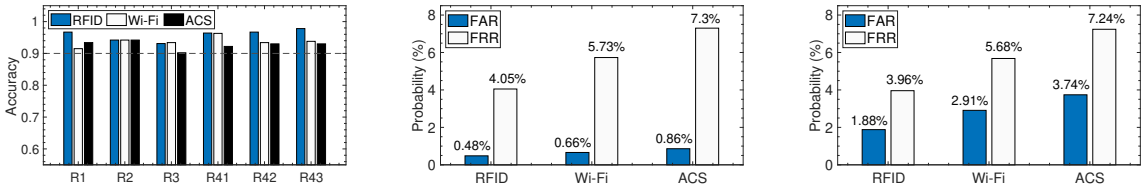
We invite 12 volunteers, including 5 females and 7 males, to participate in our experiments. A pool of voice commands is provided, which includes 20 commands (contains three to five words) in total, *e.g.*, "open the door", "play paid-TV shows", "switch off the alarm system", *etc.* All the volunteers are asked to select the same command "yes, I'm sure" as well as another randomly selected command from the pool, and then register them in WISE. We evaluate the system performance in six experimental environments as shown in Fig. 10, including four different rooms (denoted as R1, R2, R3 and R4) and three different places in room 4 (denoted as R41, R42, R43). In each environment, all the volunteers say their registered voice commands for 75 times on each test-bed, where 50 times

of the data is used to train V-WNet and 25 times of data for evaluation. To investigate the system performance under various attacks, each volunteer serves as the victim in turn. All other volunteers then play as attackers and say the same command as the victim's for compromising the system. The working distance measures the distance between the user's mouth and the sensors, *e.g.*, the RFID tag, Wi-Fi module or microphone. In the experiments, the training data are collected at the working distance of 5 cm only, while we further vary the working distance, as well as some other factors, *e.g.*, the speaking speed, the mouth's angle towards the microphone, *etc.*, to thoroughly examine the system performance.

We evaluate WISE using three main metrics — verification accuracy, false acceptance rate (FAR) and false rejection rate (FRR). The verification accuracy refers to the probability that the system can detect live legal users correctly. FAR and FRR are probabilities that the system accepts attackers' commands or rejects legal users' by mistake, respectively, due to the impersonators, voice replay or dual-channel replay attacks.

5.1 Overall Performance

We first examine the overall system performance in this subsection, and further investigate how different practical factors may impact the system performance in the next subsection.



(a) Verifying live legal users. (b) Resisting the impersonation attack. (c) Resisting the voice replay attack.

Fig. 11. Overall performance, including the verification accuracy and the defense to impersonation and voice replay attacks.

5.1.1 Distinguishing live legal users. Fig. 11(a) reports the accuracy of verifying 12 live legal users in these six environments, which is above 90% in all the six environments. In particular, the average accuracy is 95.8% for RFID, 93.8% for Wi-Fi, and 92.6% for acoustic (denoted as ACS in this section) test-beds respectively, which demonstrate that WISE can perform well in different environments. The reason why the ACS test-bed performs worse than other two is due to the relatively low packet transmission rate (*e.g.*, about 16.5 packets per second), which limits the granularity of the acoustic sensing and lower its sensing ability. We envision that the advance of the hardware and the optimization of the acoustic transmission protocol can further improve its performance.

5.1.2 Resisting impersonation attack. Voice command is vulnerable to the eavesdropping, an attacker can launch an impersonation attack by mimicking the voice and the style of the victim's speaking (*e.g.*, tone, volume and speed). In this experiment, each volunteer acts as the victim (also conducts the verification) in turn and all other volunteers play as attackers to repeat (25 times) the victim's voice command and mimic the victim's speaking style. Fig. 11(b) shows the system's average FAR and FRR. We can see that all the FARs and FRRs are less than 1% and 7.5% for three types of sensing signals. The results suggest that WISE can effectively reject an attacker who conducts the impersonation attack to compromise the system, and ensure a reliable verification for legal users. The FRR can be degraded by adjusting the network (*e.g.*, decreasing the number of the residual blocks and kernels) to tolerate more sensing "diversity or uncertainty" from the users. However, this tolerance will inevitably increase the chance (FAR) that an attacker can pass the verification. It is essentially a trade-off between FRR and

FAR. Our current network gives FAR a higher priority, because falsely rejecting a legal user only takes less than one second to verify again, but it is highly undesired to let an attacker pass the verification.

5.1.3 Resisting voice replay attack. Another crucial attack that needs to be considered is the voice replay attack. We play the recorded voice command from one of the volunteers in turn, and ask all the other volunteers to mimic the victim's mouth movement (without sound) 25 times. In this experiment, the victim also conducts the verification. Fig. 11(c) shows the system performance under this attack. We can see WISE is robust to defend this attack, wherein the FARs are only 1.88%, 2.91% and 3.74% for these three types of sensing signals, while the FRRs are all less than 7.5% (3.96% for RFID, 5.68% for Wi-Fi and 7.24% for ACS, respetively).

5.1.4 Resisting dual-channel replay attack. Finally, we examine how the system performs under the dual-channel replay attack. We replays both the legal user's voice command and the concurrent sensing signal (recorded when collecting the training/testing dataset) within the system working distance. Fig. 12(a) shows that the FARs from all the three test-beds are quite low, *e.g.*, 0%, 0.28% and 0.46% for RFID, Wi-Fi and ACS, respectively. In particular, Figure 12(b) shows that most of the dual-channel replay attacks are early rejected (*e.g.*, 100% for RFID, 99.7% for Wi-Fi and 99.5% for ACS), which can demonstrate that WISE can defend the dual-channel replay attack is indeed contributed by our transceiver sheltering module design.

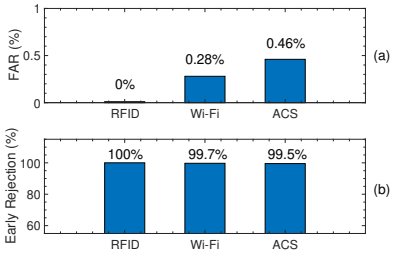


Fig. 12. Resisting the dual-channel replay attack.

Table 2. Execution latency on different test-beds.

Test Bed	Signal Processing	Network Inference
RFID	4 ms / 500 ms	12 ms
Wi-Fi	75 ms / 500 ms	16 ms
ACS	6 ms / 500 ms	9 ms

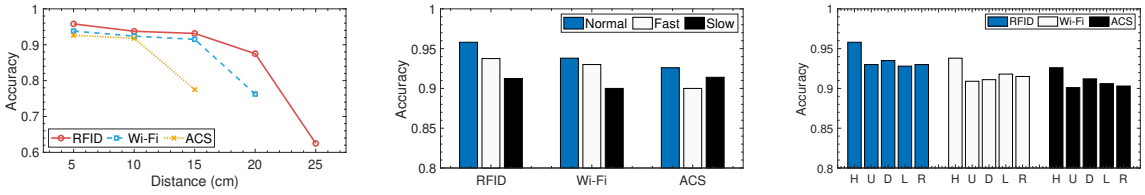
5.1.5 Execution latency. We have studied the execution latency of these three test-beds. The latency comes mainly from two parts: signal processing and the neural network inference. The signal processing further contains the processing of the sensing signal and the voice command. Because the sampling rates are different for different sensing signals, the latency of the signal processing varies. As Table 2 shows (the first number in the "Signal Processing" column), for a voice command of about two seconds, the latency varies from 4 ms to 75 ms to complete its sensing signal processing. Moreover, the processing of the voice command itself is much more time consuming — taking about 500 ms (the second number in the "Signal Processing" column), which dominates the signal processing latency. For the neural network inference part, the latency is similar to each other because the three versions of V-WNet for different sensing signals have the same neural network structure. Therefore, WISE takes less than one second (516 ms for RFID, 591 ms for Wi-Fi and 515 ms for inaudible sounds) to verify a user, which does not cause a perceptual latency to the user.

5.2 Performance Impacted by Different Factors

In this subsection, we conduct a series of additional experiments to further investigate how the different practical factors will impact the user's own usage of WISE.

5.2.1 Impact of working distance. In our experiment, the training data are collected at the working distance of 5 cm only, while we evaluate the performance of WISE by varying the working distance from 5~25 cm for RFID,

5~20 cm for Wi-Fi and 5~15 cm for acoustic three test-beds. We do not need to collect the training data from all different working distances. Through our experiments in Fig. 13(a), we find that three test-beds can achieve a good verification accuracy (over 90% accuracy) within 5~15 cm (for RFID), 5~15 cm (for Wi-Fi), and 5~10 cm (for Acoustic), respectively. When the distance is further increased, the sensing ability from each sensing channel is degraded due to the relatively low signal-to-noise ratio, which in turn lowers the accuracy. Although WISE remains a similar working distance as the prior liveness detection methods within 20 cm (yet they do not ensure the sensing-channel's safety), the working distance of all these methods, including WISE, is limited in general. Considering that that the WISE's service is triggered for a set of crucial voice comments only, it will be acceptable for the user to be relatively close to the system with our current design, while we will explore how to further prolong its working distance as an important future work of this paper.



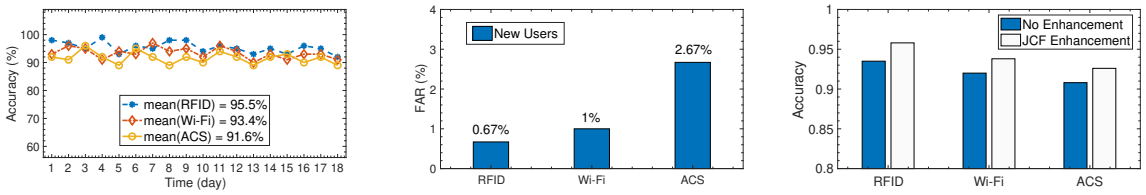
(a) Accuracy vs. working distance. (b) Accuracy vs. speaking speed. (c) Accuracy under different angles.

Fig. 13. Performance under different impacts, including the working distance, user's speaking speed and the angle of user's mouth towards the microphone.

5.2.2 Impact of speaking speed. We compare the performance under a normal speaking speed, *e.g.*, four words within two seconds, with the performance when the volunteers intentionally speak a little faster and slower. From Fig. 13(b), we have three observations: 1) Abnormal speaking speed indeed reduces the verification accuracy but all the results are still over 90%. This is mainly due to the sensing signal warping (§3.2.2) that can warp the sensing signals to a constant sequence. 2) For RFID and Wi-Fi, the faster speaking speed has a less impact than the lower speed, because intentionally slowing down the speaking speed likely changes a user's speaking habit, which further impacts the system performance. 3) On the contrary, for ACS, the faster speaking speed has a more impact than a lower speed. We believe this is because the limited acoustic sensing packet transmission rate, which dominates the performance. Overall, WISE can perform well when the user speaks under different speeds, while the normal speed is preferred and suggested to ensure the better system performance.

5.2.3 Impact of angle. In this experiment, we evaluate the impact of the angle with respect to how the user's mouth towards the device (*e.g.*, the RFID tag, Wi-Fi module or microphone), which includes keeping the user's head horizontal (H, considered as the baseline setting), rising head up (U), lowering head down (D), turning head left (L) and turning head right (R). The rotation of U, D, L and R is about 10 degrees with respect to the baseline setting (H). Fig. 13(c) shows that WISE is robust to such angle differences, and there is only a slight impact (about 1%~3% accuracy decrease) in each scenario compared to the baseline setting.

5.2.4 Impact of time span. Next, we examine the performance of WISE cross a relatively long time span of 18 consecutive days as shown in Fig. 14(a), wherein the system is trained using the data collected on the first day only. The average accuracies are 95.5% for RFID, 93.4% for Wi-Fi and 91.6% for ACS, respectively. Such a result is understandable because the user's voice features and speaking habits are relatively stable, which do not exhibit dramatic changes within a short period of time. Therefore, the system can be re-fined (by using the identifier extracted from the user's latest successful verification as the new one) with a period of several weeks in practice.



(a) Performance over a long time span. (b) Performance for the new users. (c) Impact of network enhancement.

Fig. 14. Performance under different impacts, including the accuracy over 18 consecutive days, FAR for new users and the impact of network enhancement.

5.2.5 Performance for the new users. We have also examined the system performance for the new users. The network is trained using both the legal users' data and some additional users'. All the additional users' data shares the same label, *e.g.*, "others", in the network training. We note when we test the new users' performance, the new users are not from these additional users involved in the training. In Fig. 14(b), we evaluate the FAR for the new users (recognized as one of the legal users). The result shows that WISE performs well for the new users and FAR is less than 1%~3% on three test-beds.

5.2.6 Impact of network module. In the design of our neural network, we introduce an enhancement to jointly consider the features (JCF) from both the voice command and sensing signal to improve the performance. To investigate the utility of this enhancement design, we show its impact in Fig. 14(c). We find that if we disable this JCF enhancement, the average verification accuracy is 93.5%, 92% and 90.8% for RFID, Wi-Fi and ASC, respectively. After JCF is enabled, it can improve the accuracy to 95.8% for RFID, 93.8% for Wi-Fi and 92.6% for ACS.

5.2.7 Impact of the number of tags. The size of one tag is less than the size of the user's mouth usually. In the current WISE, we adopt two tags in the RFID-based test-bed (Section 4) because their size is comparable to our mouth size. To investigate this setting, we examine the WISE's performance when only one tag is used. As shown in Fig. 15, the accuracy to verify legal users could degrade by more than 2%. For both the impersonation attack and voice-replay attacks, both the FAR and FRR with one tag are higher than them with our current setting of two tags, *e.g.*, about 1% and 2% rises for FAR and FRR, respectively.

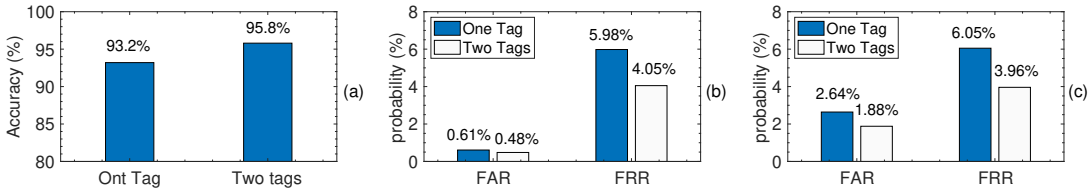


Fig. 15. Performance comparison between one tag and two tags, (a) verifying live legal users, (b) resisting the impersonation attack and (c) resisting the voice-replay attack.

5.3 Summary of the Sensing Interface Selection

According to the evaluation, we finally summarize the sensing interface selection, which can offer a flexibility to balance the system performance and cost according to application scenarios.

RFID achieves the best performance and is more robust to different factors. The pay is the hardware cost and makes it more suitable for the entrance guard systems, wherein the RFID device has already deployed.

Wi-Fi performs generally between RFID and ACS. There is still bar for the hardware with a moderate cost, *e.g.*, two NICs or one duplex NIC, which is more suitable for the smart speaker or voice assistant devices at home.

Inaudible sound performs slightly worse than Wi-Fi in our current implementation, mainly due to the limited packet transmission rate. It could have the minimal hardware requirement, which will be suitable for the personal devices. We envision that optimizing the transmission protocol can further improve the performance.

6 POINTS OF DISCUSSION

In the experiments above, we have evaluated WISE in six different environments with the training data collected from these environments. In this study, we have also investigated WISE in a more challenging cross-environmental setting. In particular, we 1) use the data from only two to five environments to form the training data set, 2) utilize the adversarial framework [15] to train our network, and 3) conduct the experiment to examine the system performance on the rest (new) environments. We find that with the training data collected from more environments, the accuracy to recognize legal users is increasing, while the negative aspect is that FAR in both the impersonation and voice replay attacks is increased as well. This is understandable, since the environment-independent ability enforces the network to discard certain user-specific features, which will of course increase the chance to falsely accept attackers. Due to this reason (as well as the fact that the voice-user interface devices have a static deployment usually), we suggest to train the system using the data collected from the targeted environment, which is more effective to avoid various attacks.

If an attacker keeps speaking some crucial commands intentionally, the WISE service will be triggered frequently and ended with a verification failure in the network. Fortunately, such a system reaction is unlikely to cause a DoS. This is because the overall execution time to process one crucial voice command in WISE is small, *e.g.*, about 500 ms (wherein the network execution itself takes about 20 ms only). This latency is much less than the time duration for the attacker speaking one voice command. As a result, it is hardly for the attacker to jam WISE due to human's limited speaking speed. On the other hand, the recourse consumption caused by this malicious behavior is similar as that when legal users use WISE one after another, which will not incur dramatic extra system overhead.

For a sensing system, it is difficult to work if its sensors cannot function properly, *e.g.*, the RFID tags are masked in our system. Therefore, we can consider to monitor and alert the consistent high-power signals that may jam the RFID tag's communication. On the other hand, when the masking indeed happens, WISE cannot decode useful information and it will keep rejecting the user's verification. Therefore, another potential countermeasure is when the WISE service is frequently triggered but without receiving the tag's responses for a long time and/or WISE keeps outputting rejections, the system can inform the administrator to double check whether the tags get masked.

7 RELATED WORK

Since sound is transmitted through an open channel, traditional speaker verification systems are vulnerable to the replay attacks [6]. As a result, prior research studies have proposed various types of countermeasures: 1) the challenge-response based approaches, *e.g.*, [17], that ask the user to speak different system randomly generated texts each time; 2) the approaches, *e.g.*, VOID in [1] and [29], will reject a verification request if they detect the sound's special characteristics caused by the hardware; and 3) the solutions like [32, 44], that will reject a verification request if the current one is exactly the same as the previous one. However, these approaches may become less effective if an attacker uses the speech conversion and synthesis techniques, or the legal user's

Table 3. Comparison among the sensing channel based system designs.

Features	TM [30]	VAuth [7]	LipPass [20]	WiVo [23]	VoiceGesture [48]	WISE
Sensor free	×	×	✓	✓	✓	✓
Resist impersonation	✓	✓	✓	✓	✓	✓
Resist voice replay	✓	✓	×	×	×	✓
Resist dual-channel replay	×	×	×	×	×	✓

voice command is recorded and replayed by a high-end device that can alleviate the characteristics or differences caused by the replay device's hardware.

In the literature, some recent works find that introducing a sensing channel could potentially avoid above limitations, *e.g.*, VoiceGesture [48], WiVo [23], LipPass [20], VAuth [7] and TM (Throat Microphone) based system [30]. Our WISE design also belongs to this category. Among these six systems, VAuth and TM-based design require users to wear the wearable devices. VoiceGesture, WiVo, LipPass and WISE have no such a requirement. However, the first three approaches do not consider the safety of the sensing channel, while we consider the safety for both the voice command and sensing two channels in WISE. More detailed comparison is summarized in Table 3, wherein the voice replay attack refers to replay a recorded voice command from the legal user and then mimic the legal user' mouth movement.

Various types of wireless signals have been adopted to sense different targets recently [10, 16, 21, 31, 37, 49], such as the activity recognition, user identification, object sensing, human skeleton recovery or motion tracking, *etc.* In WISE, we also utilize wireless signals to sense the user, while our design focuses more on protecting the credential information in the wireless sensing signals. The rational behind our design for RFID is to add some random noise (*i.e.*, Gaussian white noises [39]) to "pollute" the sensing signal [13], and only the legal receiver knows how to recover it. A recent work [41] explores adding noises in RFID signals, but we propose a different channel estimation method for the EPC recovery. On the other hand, as Wi-Fi employs a different communication protocol (*i.e.*, 802.11 standards), we introduce a constellation masking design, which has been explored to reduce the peak-to-average power ratio (PAR) for OFDM systems before [11, 12, 28], while we further employ and customize it to protect the Wi-Fi sensing signal. Moreover, we also implement an acoustic-based transmission protocol following the 802.11 standard, so that the transceiver sheltering design for Wi-Fi can be seamlessly transplanted to the acoustic sensing signals.

8 CONCLUSION

In this paper, we present an anti-spoofing system design WISE to accept voice commands from live legal users, which could supplement existing speech recognition systems significantly and enable new application potentials in practice. WISE also provides a generic interface to be compatible to a variety of wireless signals to form the sensing channel, including RFID, Wi-Fi and inaudible sounds. We develop a prototype system and examine the performance of WISE using all three types of the sensing-channel signals in six different real-world environments under a variety of system settings.

ACKNOWLEDGMENTS

This work was supported by National Key R&D Program of China 2020YFB1707700, project JCYJ20190808183203749 supported by the Science Technology and Innovation Committee of Shenzhen Municipality, the GRF grant from Hong Kong RGC (CityU 11217817), CityU SRG-Fd 7005274, the NSFC Grant No. 61832008, 62072367, 61772413, 61802299, 62002284 and Key Research and Development Program of Shaanxi (Program No.2012KTCL01-11).

REFERENCES

[1] Muhammad Ejaz Ahmed, Il-Youp Kwak, Jun Ho Huh, Iljoo Kim, Taekkyung Oh, and Hyoungshick Kim. 2020. Void: A Fast and Light Voice Liveness Detection System. In *Proc. of USENIX Security Symposium*. 2685–2702.

[2] RG Bachu, S Kopparthi, B Adapa, and BD Barkana. 2008. Separation of Voiced and Unvoiced Using Zero Crossing Rate and Energy of the Speech Signal. In *Proc. of ASEE*. 1–7.

[3] Chao Cai, Rong Zheng, and Menglan Hu. 2019. A Survey on Acoustic Sensing. *arXiv preprint arXiv:1901.03450* (2019).

[4] Tao Chen, Longfei Shangguan, Zhenjiang Li, and Kyle Jamieson. 2020. Metamorph: Injecting Inaudible Commands into Over-the-air Voice Controlled Systems. In *Proc. of NDSS Symposium*.

[5] Han Ding, Jinsong Han, Yanyong Zhang, Fu Xiao, Wei Xi, Ge Wang, and Zhiping Jiang. 2018. Preventing Unauthorized Access on Passive Tags. In *Proc. of IEEE INFOCOM*. 1115–1123.

[6] Serife Kucur Ergünay, Elie Khoury, Alexandros Lazaridis, and Sebastien Marcel. 2015. On the Vulnerability of Speaker Verification to Realistic Voice Spoofing. In *Proc. of IEEE BTAS*.

[7] Huan Feng, Kassem Fawaz, and Kang G. Shin. 2017. Continuous Authentication for Voice Assistants. In *In Proc. of ACM Mobicom*.

[8] EPC Global. 2005. Specification for RFID Air Interface. *EPC Radio-Frequency Identity Protocols Class-1 Generation-2 UHF RFID Protocol for Communications* 860 (2005), 1–94.

[9] Dai-fei Guo, Wei-Hong Zhu, Zhen-Ming Gao, and Jian-qiang Zhang. 2000. A Study of Wavelet Thresholding Denoising. In *Proc. of IEEE ICSP*. 329–332.

[10] Unsoo Ha, Junshan Leng, Alaa Khaddaj, and Fadel Adib. 2020. Food and Liquid Sensing in Practical Environments using RFIDs. In *Proc. of USENIX NSDI*.

[11] Chenggao Han and Takeshi Hashimoto. 2016. Coded Constellation Rotated Vector OFDM with Almost Linear Interleaver. In *Proc. of IEEE WCNC*.

[12] Chenggao Han, Takeshi Hashimoto, and Naoki Suehiro. 2010. Constellation-Rotated Vector OFDM and Its Performance Analysis over Rayleigh Fading Channels. *IEEE Transactions on communications* 58, 3 (2010), 828–838.

[13] Haitham Hassanieh, Jue Wang, Dina Katabi, and Tadayoshi Kohno. 2015. Securing RFIDs by Randomizing the Modulation and Channel. In *Proc. of USENIX NSDI*.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proc. of IEEE CVPR*. 770–778.

[15] Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao, Yaqing Wang, Ye Yuan, Hongfei Xue, Chen Song, Xin Ma, Dimitrios Koutsonikolas, Wenyao Xu, and Lu Su. 2018. Towards Environment Independent Device Free Human Activity Recognition. In *Proc. of ACM Mobicom*.

[16] Kiran Joshi, Dinesh Bharadia, Manikanta Kotaru, and Sachin Katti. 2015. WiDeo: Fine-Grained Device-Free Motion Tracing Using RF Backscatter. In *Proc. of USENIX NSDI*.

[17] Tomi Kinnunen, Md Sahidullah, Ivan Kukanov, Héctor Delgado, Massimiliano Todisco, Achintya Sarkar, Nicolai Bæk Thomsen, Ville Hautamäki, Nicholas Evans, and Zheng-Hua Tan. 2016. Utterance Verification for Text-Dependent Speaker Recognition: A Comparative Assessment Using the RedDots Corpus. (2016).

[18] Serkan Kiranyaz, Onur Avci, Osama Abdeljaber, Turker Ince, Moncef Gabbouj, and Daniel J Inman. 2019. 1D Convolutional Neural Networks and Applications: A Survey. *arXiv preprint arXiv:1905.03554* (2019).

[19] Eleanor Lawson, Jane Stuart-Smith, James M Scobbie, Satsuki Nakai, David Beavan, Fiona Edmonds, Iain Edmonds, Alice Turk, Claire Timmins, J Beck, et al. 2015. Dynamic Dialects: An Articulatory Web Resource for the Study of Accents. (2015).

[20] Li Lu, Jiadi Yu, Yingying Chen, Hongbo Liu, Yanmin Zhu, Yunfei Liu, and Minglu Li. 2018. Lippass: Lip Reading-Based User Authentication on Smartphones Leveraging Acoustic Signals. In *Proc. of IEEE INFOCOM*. 1466–1474.

[21] Wenguang Mao, Jian He, Huihuang Zheng, Zaiwei Zhang, and Lili Qiu. 2016. High-Precision Acoustic Motion Tracking: Demo. In *Proc. of ACM MobiCom*.

[22] Seshashyama Sameeraj Meduri and Rufus Ananth. 2012. A Survey and Evaluation of Voice Activity Detection Algorithms.

[23] Yan Meng, Zichang Wang, Wei Zhang, Peilin Wu, Haojin Zhu, Xiaohui Liang, and Yao Liu. 2018. Wivo: Enhancing the Security of Boice Control System via Wireless Signal in IOT Environment. In *Proc. of ACM MobiHoc*. 81–90.

[24] F. Mavromatis N. Kargas and A. Bletsas. 2015. Fully-Coherent Reader with Commodity SDR for Gen2 FM0 and Computational RFID. *IEEE Wireless Communications Letters* 4, 6 (2015), 617–620.

[25] Jayant M Naik. 1990. Speaker Verification: A Tutorial. *IEEE Communications Magazine* 28, 1 (1990), 42–48.

[26] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The Kaldi Speech Recognition Toolkit. In *Proc. of IEEE 2011 workshop on ASRU*.

[27] Ramjee Prasad. 2004. *OFDM for Wireless Communications Systems*. Artech House.

[28] S. Prabhu Raghavendra and Grayver Eugene. 2009. Active Constellation Modification Techniques for OFDM PAR Reduction. In *Proc. of IEEE Aerospace conference*.

[29] Yanzhen Ren, Zhong Fang, Dengkai Liu, and Changwen Chen. 2019. Replay Attack Detection Based on Distortion by Loudspeaker for Voice Authentication. *Multimedia Tools and Applications* 78, 7 (2019), 8383–8396.
[30] Md. Sahidullah, Dennis Alexander Lehmann Thomsen, Rosa Gonzalez Hautamaki, Tomi Kinnunen, Zhenghua Tan, Robert Parts, and Martti Pitkanen. 2018. Robust Voice Liveness Detection and Speaker Verification Using Throat Microphones. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26, 1 (2018), 44–56.
[31] Muhammad Shahzad and Shaohu Zhang. 2018. Augmenting User Identification with WiFi Based Gesture Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–27.
[32] Wei Shang and Maryhelen Stevenson. 2010. Score Normalization in Playback Attack Detection. In *Proc. of IEEE ICASSP*.
[33] Sayaka Shiota, Fernando Villavicencio, Junichi Yamagishi, Nobutaka Ono, Isao Echizen, and Tomoko Matsui. 2016. Voice Liveness Detection for Speaker Verification based on a Tandem Single/Double-channel Pop Noise Detector. In *Odyssey*. 259–263.
[34] Yi-Sheng Shiu, Shih Yu Chang, Hsiao-Chun Wu, Scott C-H Huang, and Hsiao-Hwa Chen. 2011. Physical Layer Security in Wireless Networks: A Tutorial. *IEEE wireless Communications* 18, 2 (2011), 66–74.
[35] Adrian P Simpson. 2001. Dynamic Consequences of Differences in Male and Female Vocal Tract Dimensions. *The journal of the Acoustical society of America* 109, 5 (2001), 2153–2164.
[36] Bronson Syiem, Sushanta Kabir Dutta, Juwesh Binong, and Lairenlakpam Joyprakash Singh. 2020. Comparison of Khasi Speech Representations with Different Spectral Features and Hidden Markov States. *Journal of Electronic Science and Technology* (2020), 100079.
[37] Yonglong Tian, Guang-He Lee, Hao He, Chen-Yu Hsu, and Dina Katabi. 2018. RF-Based Fall Monitoring Using Convolutional Neural Networks. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–24.
[38] Vibha Tiwari. 2010. MFCC and its Applications in Speaker Recognition. *International journal on emerging technologies* 1, 1 (2010), 19–22.
[39] C Van Den Broeck. 1983. On the Relation between White Shot Noise, Gaussian White Noise, and the Dichotomic Markov Process. *Journal of Statistical Physics* 31, 3 (1983), 467–483.
[40] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. 2018. Generalized End-to-End Loss for Speaker Verification. In *Proc. of IEEE ICASSP*. 4879–4883.
[41] Ge Wang, Haofan Cai, Chen Qian, Jinsong Han, Xin Li, Han Ding, and Jizhong Zhao. 2018. Towards Replay-resilient RFID Authentication. In *Proc. of ACM Mobicom*.
[42] Quan Wang, Hannah Muckenhirn, Kevin Wilson, Prashant Sridhar, Zelin Wu, John Hershey, Rif A Saurous, Ron J Weiss, Ye Jia, and Ignacio Lopez Moreno. 2018. Voicefilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking. *arXiv preprint arXiv:1810.04826* (2018).
[43] Inc Wikimedia Foundation. 2019. "Voice Frequency". https://en.wikipedia.org/wiki/Voice_frequency.
[44] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li. 2015. Spoofing and Countermeasures for Speaker Verification: A Survey. *Speech Communication* 66 (2015), 130–153.
[45] Dong Yu and Li Deng. 2016. *AUTOMATIC SPEECH RECOGNITION*. Springer.
[46] Chi Zhan, Dongyu She, Sicheng Zhao, Ming-Ming Cheng, and Jufeng Yang. 2019. Zero-Shot Emotion Recognition via Affective Structural Embedding. In *Proc. of IEEE/CVF ICCV*.
[47] Chunlei Zhang, Kazuhito Koishida, and John HL Hansen. 2018. Text-Independent Speaker Verification Based on Triplet Convolutional Neural Network Embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26, 9 (2018), 1633–1644.
[48] Linghan Zhang, Sheng Tan, and Jie Yang. 2017. Hearing Your Voice is Not Enough: An Articulatory Gesture Based Liveness Detection for Voice Authentication. In *Proc. of ACM CCS*. 57–71.
[49] Mingmin Zhao, Yonglong Tian, Hang Zhao, Mohammad Abu Alsheikh, Tianhong Li, Rumen Hristov, Zachary Kabelac, Dina Katabi, and Antonio Torralba. 2018. RF-based 3D Skeletons. In *Proc. of ACM SIGCOMM*.
[50] Bing Zhou, Mohammed Elbadry, Ruipeng Gao, and Fan Ye. 2017. BatMapper: Acoustic Sensing Based Indoor Floor Plan Construction Using Smartphones. In *Proc. of ACM MobiSys*. 42–55.
[51] Bing Zhou, Jay Lohokare, Ruipeng Gao, and Fan Ye. 2018. EchoPrint: Two-Factor Authentication using Acoustics and Vision on Smartphones. In *Proc. of ACM MobiCom*. 321–336.
[52] Man Zhou, Zhan Qin, Xiu Lin, Shengshan Hu, Qian Wang, and Kui Ren. 2019. Hidden Voice Commands: Attacks and Defenses on the vcs of Autonomous Driving Cars. *IEEE Wireless Communications* 26, 5 (2019), 128–133.