Secure mmWave-Radar-Based Speaker Verification for IoT Smart Home

Yudi Dong, *Graduate Student Member, IEEE*, and Yu-Dong Yao, *Fellow, IEEE*

***Abstract*—Voice assistant devices function as interaction gateways in the Internet-of-Things (IoT) smart home. By using voice assistants, users are able to control smart homes via speech commands. However, voice assistants introduce potential security risks and privacy disclosures. For example, malicious actors could impersonate genuine users to send smart home speech commands. Speaker/user verification thus becomes a critical issue for smart home security. This article proposes a secure method for speaker verification in IoT smart homes using millimeter-wave (mmWave) radar. Specifically, we utilize the radar to capture both vocal cord vibration (VCV) and lip motion (LM) as multimodal biometrics for identifying speakers. Traditional voice-based speaker verification methods are vulnerable to impostor attacks, such as replay attacks and voice synthesis attacks, that use recorded or imitated voice audio to spoof the system. Our approach is able to protect IoT smart homes from these attacks by continuously detecting the liveness of the user using mmWave sensing and deep learning techniques. Extensive experiments show that the proposed approach can achieve high verification accuracy and be more robust against imposter attacks.**

***Index Terms*—Deep convolutional neural network (CNN), Internet of Things (IoT), lip motion (LM) biometrics, millimeter-wave (mmWave) radar, smart home security, vocal cord vibration (VCV) biometrics.**

I. Introduction

DUE TO the rapid proliferation and development of the Internet of Things (IoT), smart homes have moved into a new paradigm. Devices in an IoT smart home, such as lights, locks, and security cameras, are connected and managed remotely over the Internet. With the advent of voice assistants [1] (e.g., Google Home, Amazon Echo, and Apple Siri), voice control has been deployed as the standard human–machine interaction interface for smart home devices [2], as illustrated in Fig. 1. Although voice assistants provide the speed, efficiency, and convenience for smart home users, many concerns regarding privacy and security are constantly rising. It often involves security-sensitive information while using voice assistants in, for instance, the control of security cameras and door/window locks. Malicious attackers with physical access to your voice assistants potentially modify the IoT devices' settings to their benefits [3]. Thus, speaker/user verification has been a task of fundamental importance to ensure secure access and build more security for IoT smart homes.

Traditional methods for speaker verification are based on voice biometrics extracted from audio signals [4], which can be divided into two categories: 1) text-dependent and 2) text-independent. Text-dependent implies that predetermined passphrases are used for speaker verification. For example, Google devices utilize the text-dependent speaker verification [5] with the global password "OK Google." Microsoft also focuses on the text-dependent authentication method [6] using a different global password "Hey Cortana." As we can see, the text-dependent method verifies the identity only one time at login, which cannot validate the speaker throughout the entire period of conversation in a session. To resolve this problem, a text-independent method [7] is presented to authenticate users without constraint on the utterance content, which is more convenient and can continuously verify the speaker, thereby it is able to prevent potential attacks during the whole session.

However, voice biometrics-based speaker verification, either text-dependent or text-independent, suffers from various impostor attacks, for instance, replay attacks [8] that utilize prerecorded audio from a legitimate user to spoof the speaker verification system and voice synthesis attacks [9] that attempt to spoof the system by playing the computer-generated audio through loudspeakers. The countermeasures of these spoof attacks rely on analyzing acoustic characteristics of the input audio to distinguish between the genuine audio of live users and the played speech [10]–[12]. However, such methods are only effective for the input audio with blemishes; for example, the prerecorded audio contains ambient noises, or the synthetic audio has no pop signal caused by breathing [13]. Thus, such methods of defending against replay attacks easily fail if an adversary uses a superb recorder to obtain high-fidelity recordings [14]. Also, as the development of artificial intelligence and synthesis techniques [15], synthetic speech is becoming more and more genuine that is hard to be recognized based on acoustic characteristics.

Recently, speaker verification systems based on other biometrics [16]–[18] have drawn considerable attention as they can counter these attacks by detecting the speaker behaviors [e.g., lip motions (LMs) and throat vibrations] instead of exploring acoustic characteristics of the input audio. Liu and Cheung [16] proposed to explore the characteristics of LMs to identify speakers. However, the verification accuracy of this system is limited because the LM signals are extracted

Manuscript received May 12, 2020; revised August 5, 2020; accepted September 2, 2020. Date of publication September 10, 2020; date of current version February 19, 2021. *(Corresponding author: Yu-Dong Yao.)*

The authors are with the Electrical and Computer Engineering Department, Stevens Institute of Technology, Hoboken, NJ 07030 USA (e-mail: yyao@stevens.edu).

Digital Object Identifier 10.1109/JIOT.2020.3023101

Fig. 1. Voice control for devices in an IoT smart home.

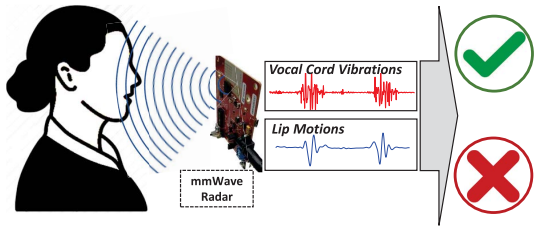


Fig. 2. Speaker verification using the mmWave radar.

from the video, which can only capture partial LMs in a 2-D plane. In addition, Sahidullah *et al.* [18] utilized the throat microphone to obtain the throat vibrations for voice liveness detection and speaker verification. While effective, it requires users to wear the body-conducted sensor, which introduces constraints and inconvenience.

To address these issues, as illustrated in Fig. 2, we introduce a millimeter-wave (mmWave) radar-based speaker verification system that leverages both vocal cord vibration (VCV) and LM as multimodal biometrics to authenticate speakers and detect spoof attacks. First, it can not only achieve accurate speaker verification by using the new biometrics but also it is performed through noncontacting sensing, which does not require users to attach any sensors on the body. Second, it is highly effective in distinguishing between live users and machine-played audio. In particular, our system utilizes a 77-GHz mmWave radar to sense and extract VCVs and LMs while a speaker has utterance. These small articulatory gestures can be measured through computing the phase change of radar signals by taking a fast Fourier transform (FFT). The phase values are unwrapped to amend their absolute jump values. Then, we perform the phase difference calculation on the unwrapped phase to enhance the signal. The algorithm for body motion influence elimination is performed on phase differences to reduce the impact of body movements. Next, the onset/offset detection algorithm is applied to segment speaking-related signals. To extract the VCV signal and LM signal from the segments, we, respectively, design an IIR filter by using the aliasing effect. The mel-frequency cepstrum coefficients (MFCCs) are utilized for exploring unique biometric features from VCV signals. To obtain distinct biometrics embedded in LM signals, the fuzzy wavelet packet transform (Fuzzy WPT) is performed. We finally utilize a deep convolutional neural network (CNN) with two convolutional layers and five fully connected layers to verify the identity of a speaker and detect spoof attacks. Extensive experiments involving six subjects (3616 speaking sentences) are conducted to evaluate our approach. Also, we perform replay attacks to our system to examine its effectiveness in detecting attacks. The results show that our system enables the speaker verification with high accuracy and defends against spoof attacks very effectively.

Our contributions in this work are summarized as follows.

1) We explore new noncontact VCV-based biometrics for speaker verification. Also, we utilize LM-based biometrics as additional features to obtain more accurate verification.
2) We propose to leverage mmWave radar to perform speaker verification, which can not only authenticate speakers in a noncontact and unobtrusive manner as in the voice-based method but also can better detect various spoof attacks.
3) Extensive experiments with different scenarios are performed to evaluate our approach, including verification with live speaking people, verification with unadjusted sensor placements, verification with different classifiers, and verification under replay attacks, which demonstrate that our system can achieve accurate and robust speaker verification and great resistance to spoof attacks for the application of IoT smart home.

The remainder of this article is organized as follows. Section II reviews the related work. Section III explores the human phonation mechanism and the feasibility of using VCV and LM as biometrics and illustrates the measuring principle of mmWave radar. In Section IV, we design methods for extracting VCV and LM signals from radar signals and using them as multimodal biometrics for CNN-based authentication. Then, Section V presents experimental results to evaluate the performance and robustness of our approach. Finally, we discuss and conclude our work in Sections VI and VII.

II. Related Work

A. Speaker Verification Using Microphone

Voice biometrics is the most popular manner for speaker verification. The traditional approach of voice-based speaker verification utilizes the joint factor analysis [19] and linear discriminant analysis [4], which is able to, respectively, extract *i*-vector features from utterances and discriminate speakers in reduced dimensionality. But this approach is not tractable due to its complex optimization analysis. Recently, the deep learning model, such as CNN and long short-term memory recurrent neural network (LSTM) [5], [6], becomes an attractive alternative to substitute the complicated analysis. Google [5] proposes to use an LSTM to extract the speaker representation of an utterance and then use simple logistic regression to discriminate the speaker representation, which is a simple and efficient method requiring little domain-specific knowledge. Similarly, Microsoft [6] utilizes two CNN models to respectively, extract speaker discriminative information and the phonetic information from an utterance for better identification. However, voice biometrics is vulnerable under imposter attacks, where an attacker plays recorded audio from an authentic speaker or synthetic audio to spoof the speaker verification system. Kinnunen *et al.* [20] evaluate 49 popular voice-based speaker verification systems in terms of imposter attack discrimination. The average equal error rate (EER) of these systems is up to 26.01%.

B. Speaker Verification Using Other Sensors

To better counter such imposter attacks, researchers explore other sensors, such as cameras [16], [17] or throat microphones [18], [21], to detect speaking behaviors (e.g., LM and vocal cord vibration) and perform speaker verification. Liu and Cheung [16], [17] proposed to use a camera to capture LMs of speakers saying the passwords for authentication, which can achieve lower EER comparing with voice-based methods. But this method requires the speaker to provide the correct private password information. Some researchers utilize throat microphone to capture the VCV for speaker verification [18], [21]. These systems can effectively counter imposer attacks by detecting the voice liveness. However, they require speakers to wear the throat microphone on their necks, which cause inconvenience or difficulty in using these systems.

To achieve a noncontacting scheme, Zhang *et al.* [22] proposed a smartphone-based acoustic sensing system for the speaker verification, called VoiceGesture, which uses a loudspeaker and a microphone to detect the user's articulatory gestures and achieves a good performance of replay attack detection. However, the application scenario of VoiceGesture is restricted to smartphones, where it works when users place the phone by the ear or in front of the mouth. Also, VoiceGesture focuses on the mouth motions, such as LM, tongue motion, and jaw motion, which does not explore VCV biometrics. In addition, Meng *et al.* [23] proposed a system named WiVo, which utilizes ambient WiFi signals to sense vocal gestures for speaker verification. Similar to VoiceGesture [22], WiVo makes use of mouth motions and does not investigate VCV biometrics. In addition, ambient WiFi signals are not stable and easy to be influenced. The effective sensing range of WiFi is also small, where users need to be close to the WiFi antenna (i.e., 20 cm).

Different from the aforementioned studies [22], [23], our system uses mmWave radar to detect both VCV and mouth motions for speaker verification, which can be more accurate and robust by exploiting the multimodal vocal gestures. In addition, the common concern in [22] and [23] is that the sensing distance of both acoustic sensing [22] and WiFi sensing [23] is short, which restricts them in a typical IoT smart home usage scenario that requires a few meters sensing distance. The mmWave radar has a larger sensing range due to its adjustable transmission power and frequency modulated continuous wave. Our system, equipped with a 77-GHz mmWave radar, is able to detect replay attacks with high accuracy in the range of 2 m, which is presented in Section V-J.

C. Speech Acquisition Using mmWave Radar

Most related to our work is mmWave radar-based speech acquisition, which acquires high-quality voice signals for speech recognition through detecting articulatory gestures using mmWave radar [24]–[26]. Li *et al.* [24] and Chen *et al.* [25], respectively, utilized phase spectrum compensation and empirical mode decomposition to reconstruct the speech signals from 94-GHz mmWave signals. Xu *et al.* [26] developed a noise-resistant speech sensing system based on mmWave, which uses a deep LSTM network to perform exhaustive extraction of speech information from the mmWave spectrogram. Also, VCV sensing using mmWave attracts attention in the academic community [27]–[29]. However, all of them do not study further the biometrics of articulatory gestures or VCV for speaker verification. Our system lays the first attempt for VCV biometrics based on mmWave sensing. We propose a mmWave-based speaker verification system, which is inspired by these previous studies but fundamentally different from them in exploring multimodal biometrics of LM signals and VCV signals through noncontact mmWave sensing.

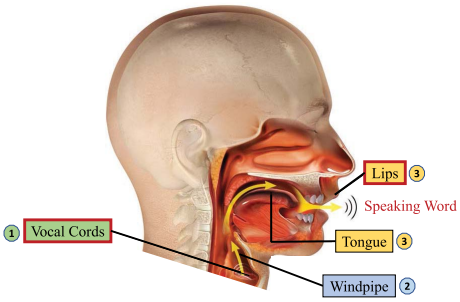


Fig. 3. Mechanism of human voice production.

III. Preliminary

In this section, we introduce the background of the physiology of human voice production and explain the potential of the VCV and LM as the distinguishing biometrics. We describe the operating principles of the mmWave radar. In addition, we present the common imposter attacks in a speaker verification system.

A. Feasibility Study

The voice mechanism of human speaking consists of three processes: 1) voiced sound; 2) resonance; and 3) articulation [30]. As illustrated in Fig. 3, each process, which involves different parts of the human body, has specific roles in voice production. In particular, vocal cords first vibrate through the pushed air from the lungs, which produces the basic sound or voiced sound. The voiced sound is then amplified by the resonances of the vocal tract resonators, such as the windpipe and throat. Finally, the recognizable speaking words are produced by modifying the voiced sound leveraging the vocal tract articulators (i.e., the tongue and lips).

Vocal cords consist of two mucous membranes that are controlled by vocal muscles [31]. These components are entirely different from person to person, which are as unique as fingerprints and are also impossible to counterfeit [32]. Thus, VCV intrinsically is a valid biometric for each individual [32]. Similarly, the human lips are also supposed to vary from person to person even with their simple structure [33]. LM can be a valid biometric for speaker verification. In this article, we consider VCV as the main biometrics and utilize LM as the auxiliary biometrics to perform speaker verification.

B. mmWave Radar

The mmWave radar used in our system is an AWR1642 radar sensor integrating a digital signal processor (DSP)

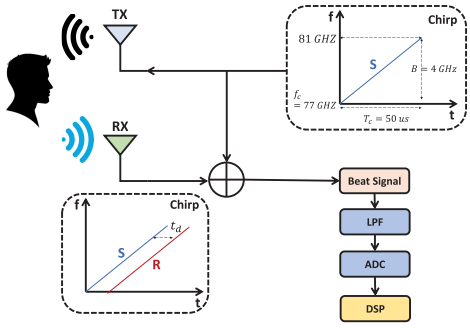


Fig. 4. Sensing scheme of the FMCW radar.

from Texas Instruments (TIs) Company, which is an integrated single-chip FMCW radar sensor capable of operation in the 76–81 GHz band with a 4-GHz bandwidth [34]. The AWR1642 radar has a high resolution, which can detect the tiny displacement from 0.02 to 0.1 mm [34]. This displacement sensitivity is high enough for our work since the amplitude of the VCV is from 0.3 to 1.5 millimeters [35]. In addition, the mmWave readily penetrates certain materials, such as clothing and plastic [36], and is reflected from the human body. By taking advantage of this feature, mmWave radar is able to detect the tiny skin vibrations induced by VCV without being affected by external influences such as clothing.

As shown in Fig. 4, the radar synthesizer generates a periodic linearly increasing frequency chirp (i.e., FMCW signal) transmitted by the TX antenna. The transmitting rate is set as 100 Hz. The transmitted chirp $S(t)$ is characterized by a start frequency f_c, bandwidth B, and duration T_c as follows:

$$S(t) = e^{j(2\pi f_c t)+\pi \frac{B}{T_c} t^2}. \tag{1}$$

The chirp is reflected off an object and receives by the RX antenna as a delayed version of the transmitted chirp

$$R(t) = e^{j\left(2\pi f_c (t-t_d)+\pi \frac{B}{T_c}(t-t_d)^2\right)} \tag{2}$$

where t_d is the time delay of receiving.

The TX chirp $S(t)$ and RX chirp $R(t)$ are then mixed to get a beat signal [37]

$$B(t) = S(t) \oplus R(t) = e^{j(f_b t+\phi_b)} \tag{3}$$

where f_b is the frequency of the beat signal which is equal to the difference of frequencies of $S(t)$ and $R(t)$. Similarly, ϕ_b referred to as a phase is equal to the phase difference of the transmitted and received chirps. In particular, the relationship between the received phase and the range of the object is as follows:

$$\phi_b = \frac{4\pi}{\lambda} R \tag{4}$$

where λ is the wavelength of the radar signal and R refers to the range between the object and the radar sensor. Any movements of an object can be reflected in the phase changes of the beat signal. The beat signal is finally digitized using a low-pass filter (LPF) and an analog-to-digital converter (ADC) for further processing on a DSP. We will further describe the use of the beat signal for deriving the LM signal and VCV signal in Sections IV-A and IV-C.

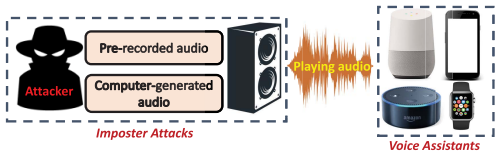


Fig. 5. Illustration of impostor attacks for voice assistants.

C. Imposter Attacks

The current speaker verification system is vulnerable to imposter attacks. As shown in Fig. 5, an adversary attempts to play the prerecorded voice of a genuine client or the computer-generated audio to spoof the speaker verification system in our voice assistants. There are typically two types of imposter attacks, including replay attacks and voice synthesis attacks.

Replay Attacks: These attacks utilize a sample digital voice recorder to obtain the voice audio of a legitimate user and then play the prerecorded audio to attack the system. It is quite easy to carry out and it requires a recorder and a loudspeaker to play the voice audio but it archives a very high attack success rate. In addition, traditional voice-based methods for replay attack detection can only achieve an EER of 26.01% on average [20]. Thus, replay attacks pose serious threats to the speaker verification system.

Voice Synthesis Attacks: Similar to the replay attacks, voice synthesis attacks also use a loudspeaker to play the audio to trick the system. The difference is that the audio used here is artificially synthesized based on acoustic characteristics of genuine one by using voice conversion or speech synthesis techniques. These attacks, which require specific expertise, are more sophisticated than replay attacks to generate.

IV. Design Method

In this section, we present our methods of data processing of radar signals and neural network for speaker verification. As depicted in Fig. 6, our data processing algorithm mainly consists of three modules. The first is the module of radar signal processing. Once we acquire the beat signal from the receiver, we measure the change in phase by using an FFT. The phase values are then unwrapped and enhanced through phase unwrapping and phase difference, respectively. The values of phase differences are processed for eliminating body motion influence and ambient noise through the motion influence elimination algorithm. Next, in the second module, the processed radar signals are segmented via onset detection and offset detection. By identifying the duration of every segment, we can pick up speaking-related segments. Finally, in the third module, we use an IIR filter based on the aliasing analysis to filter out the VCV signal, which is then reconstructed into biometric features using the mel-frequency cepstrum coefficient (MFCC). Also, the LM signal is extracted by using the IIR filter and reconstructed through fuzzy WPT.

For speaker verification in Fig. 6, we use the MFCC-based features from the VCV signal and fuzzy WPT-based features from the LM signal to build the profile for the legitimate user and train a deep CNN for identifying speakers. Once VCV and

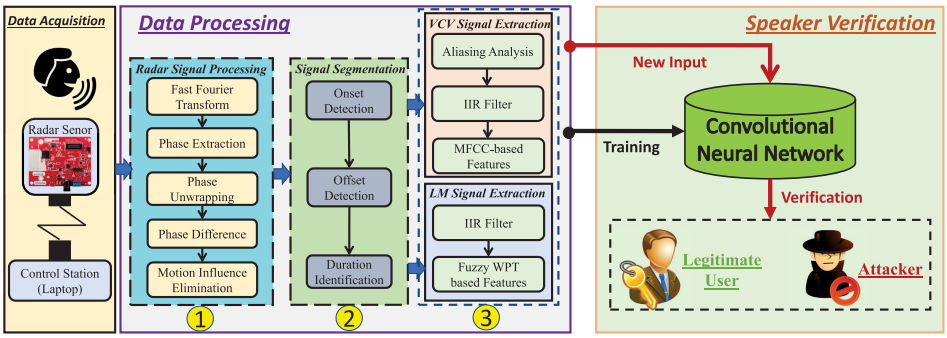


Fig. 6. Flow of the designed method.

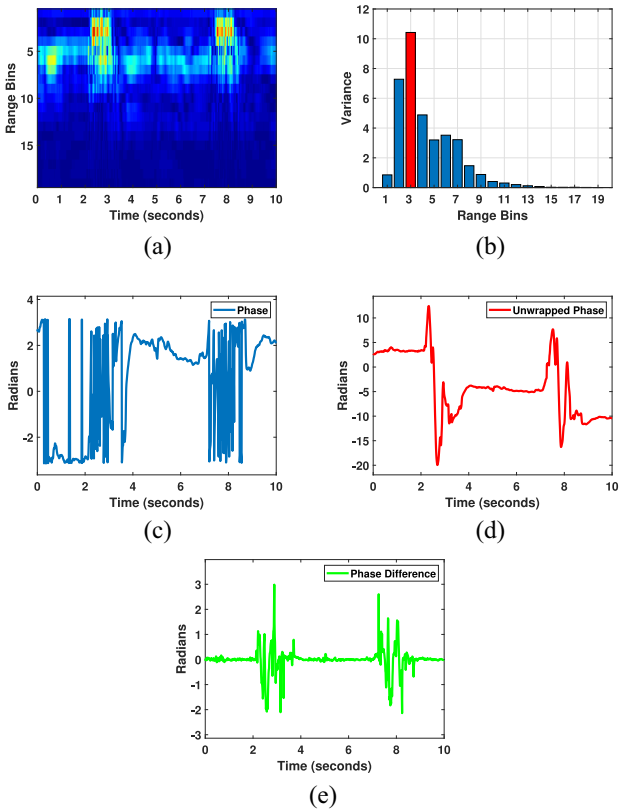


Fig. 7. Steps of radar signal processing. (a) Magnitude of FFT results of beat signals. (b) Range selection. (c) Phase extraction. (d) Phase unwrapping. (e) Phase difference.

LM related signals of an unknown subject are detected in our system, our trained CNN would tell whether it is a legitimate user or not.

A. Radar Signal Processing

To measure the LM and VCVs, we compute the phase change $\Delta\phi_b$ of the digitalized beat signal mentioned previously in Section III-B. The movement distance of an object ΔR has the following relation with the phase change:

$$\Delta\phi_b = \frac{4\pi}{\lambda}\Delta R \tag{5}$$

and we can use the phase change of the radar signal to measure the small-scale motions of the human body.

The phase is computed by performing an FFT of the beat signal $B(t)$. For example, in Fig. 7(a), the magnitude of FFT results of beat signals at all 19 range bins are shown when someone sits in front of the radar and speaks two words. The highlighted parts of Fig. 7(a) around the third second and eighth second signify the speaking activities. To select the target range bin, we compute the variance of the beat signal magnitude at each range bin and choose the range bin with the greatest variance. Here, we can see from Fig. 7(b) that the third rang bin is our target.

Fig. 7(c) shows the phase information at the target bin, which has many absolute value jumps. To obtain smoother phase information, we correct the radian phase angles by adding multiples of $\pm 2\pi$ when the absolute value jumps between consecutive elements are greater than or equal to the jump tolerance of π radians, which is called phase unwrapping. Fig. 7(d) depicts the unwrapped phase and we can see that fluctuations in two places clearly show the two speaking activities. To subtract successive phase values, as shown in Fig. 7(e), we finally perform the phase difference calculation by subtracting successive values of the unwrapped phase, which can enhance the signal and remove any phase drifts.

In addition, some large body movements while speaking, such as head movements and shoulder movements, are able to cause large fluctuations in the phase of radar signals, which impacts the measurements of LM and VCV. To eliminate this influence, we compute the energy (i.e., the mean of phase values) every ten samples (i.e., 0.1 s) in real time. If the energy of one segment exceeds a defined threshold, the values of these ten samples are discarded.

B. Signal Segmentation

After eliminating the influences of large body motions, any fluctuations in the waveform of phase differences are supposed to be caused by speaking activity. We thus segment these speaking-related fluctuations for further processing. First, we detect all the onsets and offsets in the phase difference signal. Onset and offset are the timings of transients in signals. The difference is that the onset is the beginning of a fluctuation after silence and the offset is the ending of the fluctuation followed by silence. The onset and offset can be detected by computing the spectral flux based on the spectrogram of the signals. The details of onset/offset detection algorithms are refereed to [38].

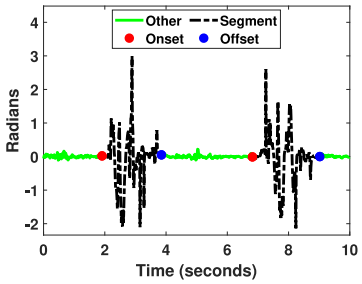


Fig. 8. Illustration of signal segmentation.

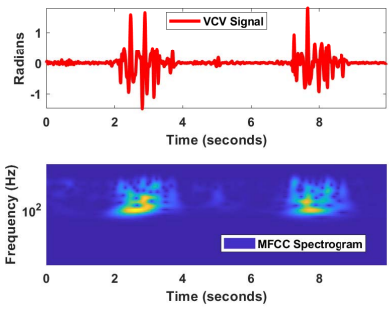


Fig. 9. Visualization of VCV signals and features.

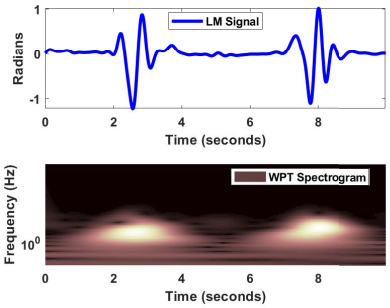


Fig. 10. Visualization of LM signals and features.

As illustrated in Fig. 8, we can observe that there are two speaking activities showing two fluctuations in the signal of phase differences. To segment these two speaking-related fluctuation signals, we only need to extract the signals between the detected onset and offset. In addition, some outliers caused by hardware noise can be falsely identified/segmented. The duration of these false segments is very short, which is typically less than 0.2 s based on our observation. Thus, to avoid these wrong segments getting into the next step, we set a threshold (i.e., 0.2 s) for identifying the duration of segments, where the duration of the segment greater than 0.2 s can be passed for further processing.

C. VCV/LM Signal Extraction

The vibration frequencies of lips and vocal cords are in the range of [0.2 Hz, 3 Hz] and [90 Hz, 200 Hz] [39], respectively. To extract the LM signals from the segmented phase difference values, we directly operate one band-pass IIR filters with range [0.2 Hz, 3 Hz]. According to the Nyquist theorem, since the sampling rate of the radar device is only 100 Hz (less than twice VCV frequency), the sampled VCV signal has aliasing. The alias frequency of the VCV signal after sampling using the radar is given as [40]

$$F_a = |F - R_s * R_{\text{int}}| \tag{6}$$

where F refers to the frequency of the original signal that is the VCV signal with the frequency range [90 Hz, 200 Hz], and R_s is our sampling rate that is 100 Hz, and R_{int} is the closest integer between multiple of R_s and F, which is equal to one. We can then calculate the range of the alias frequency from 10 to 100 Hz. Thus, a low pass IIR filter with a threshold (i.e., 10 Hz) is operated on the segmented phase values to extract the VCV related signals.

Next, we use MFCC and Fuzzy WPT to extract features from VCV and LM signals, respectively.

1) MFCC-Based Features of VCV: MFCC is a very efficient method for feature extraction of 1-D signals and it has been widely utilized in speech recognition. We derive MFCCS in four main procedures. First, the VCV signal is taken a short-time Fourier transform to obtain the power spectrum. Second, we use 25 overlapping triangular windows to map the power spectrum onto a nonlinear mel scale, where we can get an MFCC log spectrogram. As shown in Fig. 9, we visualize the extracted VCV signals and the corresponding MFCC spectrogram where the highlight portions around 3 and 7 s indicate the speaking activities. Then, we calculate the log power at each of the mel frequencies and take a linear cosine transform of them. Finally, 25 MFCCs can be obtained by computing the amplitudes of the transformed log power spectrum of the speaking activities. Also, the first-order and second-order derivatives of the 25 MFCCs are calculated, respectively. In total, for each VCV segment, 25×3 MFCC features are extracted.

2) Fuzzy WPT-Based Features of LM: To extract features that highly correlate with lips motions, we perform fuzzy WPT [41] on each LM signal segment. The wavelet packet decomposition is an expansion of the discrete wavelet transform whereby both the approximation and detail subspaces are decomposed. The efficacy of the wavelet packet transform relies on choosing a proper wavelet basis, which determines whether the decomposed subspaces are highly distinguishable among individuals. To select the best wavelet basis, a fuzzy-entropy-based mutual information (MI) method [41] is applied. We finally perform five-level fuzzy wavelet packet decomposition, which obtains $\sum_{j=0}^{7} 2^j$ (i.e., 25) wavelet subspaces. Fig. 10 shows the LM signals and the 5th level wavelet packet spectrogram that indicates two speaking activities around 3 and 7 s. We then empirically choose four statistic features of each subspace signal as the representative features for each subspace, which are *mean*, *standard deviation*, *mode*, and *skewness*. Thus, for each LM segment, a total of 25×4 fuzzy wavelet packet-based features are extracted.

We notice that the MFCC-based features and fuzzy WPT-based features may not be on the same scale, which would cause the classification bias. To avoid this issue, we perform standardization normalization on features by subtracting the mean and dividing the standard deviation of feature values.

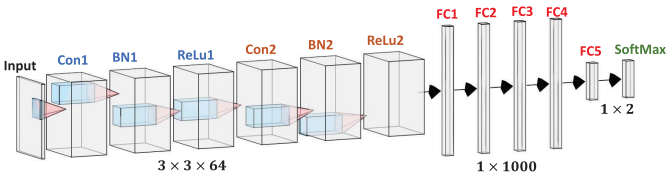


Fig. 11. Architecture of our deep CNN.

D. CNN-Based Speaker Verification

For speaker verification, we design a deep convolutional network. The configuration of this network is as shown in Fig. 11. Specifically, the first layer is the input layer used for reading the data from a 25×7 matrix, where the data of the first three columns are VCV MFCC-based features and the remaining four columns represent LM WPT-based features. Then, we use two convolution layers with a 64-channel filter with a size of 3×3 and the same padding. One batch normalization layer used for normalizing the data and one rectified linear unit (ReLU) layer used as the activation function are added behind each convolution layer. Then, four fully connect layers with 1000 neurons are added in the network. For the last two layers, a fully connected layer with two neurons and a softmax layer are utilized for classifying two classes. In our case, one class is the legitimate user of the speaker verification system and another class consists of other people.

For countering imposter attacks that use a loudspeaker to spoof the system, we use the same designed data processing methods and the same CNN-based classifier, which can effectively detect these attacks.

In addition, the deep learning model is inclined to have an overfitting problem, especially with a limited number of samples. We address this issue in three aspects as follows.

1) *Cross-Validation Scheme:* Cross validation is a powerful measure to prevent overfitting for machine/deep learning [42]. We use a fivefold cross-validation scheme in our CNN-based classifier, which is described in Section V-C.
2) *Feature Extraction:* Deep learning usually does not require feature extraction as the neural layers can extract features automatically. However, with a limited number of training samples, the deep learning features may not be sufficiently representative, which may lead to overfitting. Therefore, we manually extract two distinct features (i.e., MFFC-based features and WPT-based features), which can help to prevent overfitting due to the coarse-grained inputs.
3) *Data Diversity:* We increase the data diversity as much as possible to prevent overfitting. We performed 30 rounds of experiments to collect data, which lasts about three months. We also ensure that the training data and testing data are from different days of collection, which largely avoids overfitting during the training.

Furthermore, to tackle the robustness problem in the deep learning model, some studies propose to use a denoising autoencoder [43] or adversarial networks [44], which could be applied to our system in the future. In Section VI, we discuss how these potential methods make our system more robust.

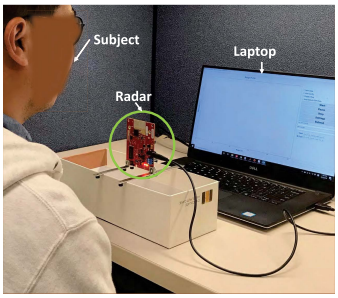


Fig. 12. Experimental setup.

V. Experiments and Evaluation

In this section, we evaluate the performance of our proposed method for both speaker verification and replay attack detection. We examine the robustness of our system considering various placement angles of the radar. Also, we explore the impact of the number of subjects. Finally, we compare the performance of the system under different classifiers.

A. Experimental Setup

We use the AWR1642 77 GHz-band mmWave radar from TIs Inc., which operates at the frequency of 77 GHz with a bandwidth of 4 GHz. Each transmitting frame is configured to have two chirps. The frame rate is set to 100 Hz, which is the same as the sampling rate. The chirp duration is 50 ms based on the beat sampling rate of 2 MHz. In addition, the radar has two TX antennas and four RX antennas. However, one single TX-RX antenna pair is currently used in our experiments for the faster processing speed and shorter delay. Moreover, the AWR1642 radar and a DSP chip have been integrated into a small board (i.e., 5 cm $\times$ 8 cm), which is portable and convenient for us to use under different environments and situations. The board is powered by 5-V voltage with 2.5-A current. The experimental setup is shown in Fig. 12 and the radar sensor is horizontally placed (i.e., 0 degree) in front of a sitting participant with a distance of 0.2 m. The radar sensor is connected with a laptop for control and data processing. The laptop is also used for recording the voice from speakers. A professional digital voice recorder is placed beside the laptop to record the voice for the experiments of detecting replay attacks.

B. Data Collection

Six subjects are involved in the experiments. Five of them are live users and another subject is a loudspeaker which can play the prerecorded voice from the voice recording devices (i.e., laptop microphone and professional digital voice recorder). One of the live users is regarded as a legitimate user, and all other subjects are unauthorized users or attackers. We performed 30 rounds of experiments. In every round, each subject will read or play 20 sentences with the same experimental setup to collect one data set. That is to say, the sample number is 3616 as combining of six objects, 30 rounds, and 20 sentences each round. After data processing, there are in total 30 sets of data containing 3616 VCV and LM samples. In the evaluation, 20 data sets are randomly selected for training and the remaining ten are used for testing.

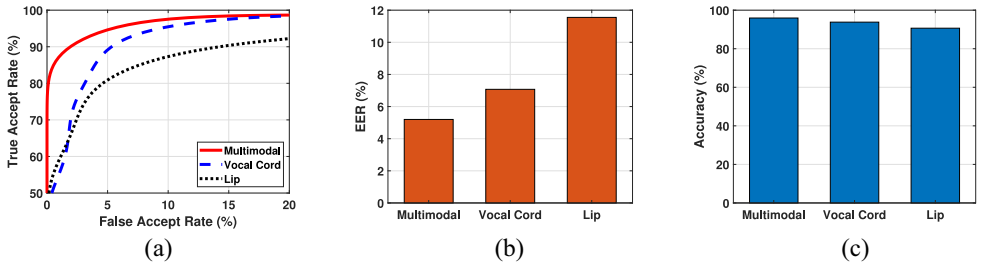


Fig. 13. System performance of speaker verification. (a) ROC Curve. (b) EER. (c) Accuracy.

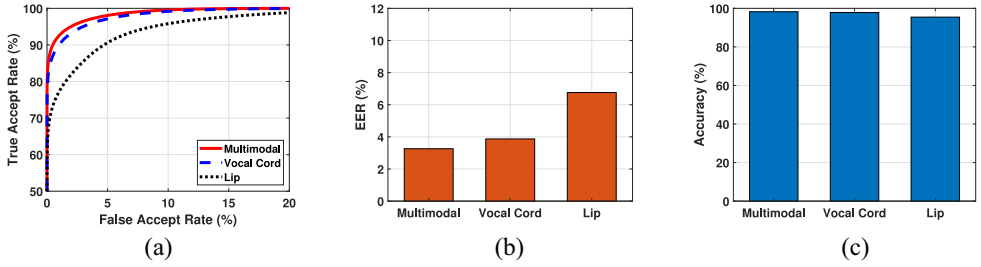


Fig. 14. System performance of detecting replay attacks. (a) ROC curve. (b) EER. (c) Accuracy.

C. Classifier Training

We train the CNN-based classifier with a fivefold cross-validation scheme, which is able to prevent overfitting. Specifically, the training data are randomly divided into five subsets. Four subsets are selected for training the classifier and the remaining one subset is used for validation. This process is repeated five times with different validation data from five subsets. Finally, we average the model parameters of five cross-validation results to obtain the optimal model.

D. Evaluation Metrics

We adopt three metrics, including accuracy, receiver operating characteristic (ROC) curves, and EER to evaluate the performance of our system.

Accuracy is the fraction of correct verification samples in our system, which is defined as

$$\text{Accuracy} = \frac{\text{Correct verification samples}}{\text{Total verification samples}}. \tag{7}$$

The ROC curve is generated by plotting the true accept rate (TAR) against the false accept rate (FAR) at various threshold settings, which illustrates the performance of a binary classifier system as its discrimination threshold is varied. A good classifier has a high TAR with a low FAR. A system is commented to have better performance when its ROC curve covers a larger area.

EER is an acknowledged index for evaluating the biometric security system. EER is the point on the ROC curve that corresponds to have an equal probability of miss-classifying an acceptance or rejection sample. This point is obtained by intersecting the ROC curve with a diagonal of a unit square. The value of EER indicates that the proportion of false acceptances is equal to the proportion of false rejections. The lower the EER value, the better the performance of the biometric system.

E. Performance of Verifying Users

The speaker verification system is mostly used for identifying the live person. We first present the performance of our system in verifying live users. Fig. 13(a) depicts the ROC curves of our system using different biometric features. We can see that the system using multimodal biometrics has the best performance achieving 95% TAR at 5% FAR, where its ROC curve is the highest in the plot. Although the vocal cord biometrics method does not perform as good as the multimodal biometrics method, it is still satisfactory. If we only use lip biometrics in the system, performance degradation is severe. As shown in Fig. 13(b), EER of the multimodal, vocal cord, and lip biometrics is 5.2%, 7.07%, and 11.55%, respectively. Similarly, in Fig. 13(c), we observe that multimodal biometrics can achieve an accuracy of 95.93%, whereas the vocal cord biometrics alone achieves an accuracy of 93.79% and the accuracy of lip biometrics is only around 90%. The above results demonstrate the effectiveness of our system in verifying live speaking person. Also, we show that our proposed vocal cord biometrics is valid for speaker verification. It can achieve better performance if we combine vocal cord biometrics with lip biometrics to complement with each other.

F. Performance of Detecting Replay Attacks

We evaluate our system under imposter attacks. We choose to perform replay attacks in the evaluation, which is the most common imposter attack type. Specifically, we use a loudspeaker to play the prerecorded audio of the legitimate user to spoof the system. The experimental setup is the same as shown in Fig. 12, but the live speaking person is replaced with a loudspeaker.

Fig. 14(a) shows the ROC curves of different biometrics under replay attacks. We observe that the multimodal biometrics gives the best performance. The vocal cord biometrics also performs well, which has a very similar ROC curve with the

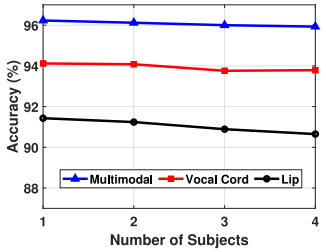


Fig. 15. Performance of different numbers of subjects.

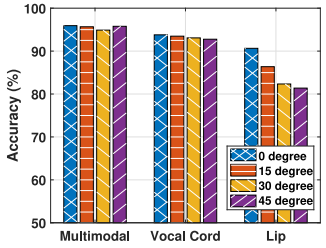


Fig. 16. Performance under different angle degrees of the radar placements.

multimodal biometrics. They all achieve 98% TAR with 5% FAR. When using lip biometrics, TAR is only around 90% with 5% FAR. In addition, Fig. 14(b) and (c), respectively, shows the performance in terms of EER and accuracy under replay attacks. With the vocal cord biometrics, the system can achieve 97.83% accuracy at ERR of 3.87%. The accuracy of lip biometrics is lower than the vocal cord biometrics, which only achieves 95.45% with 6.76%. We can see that the multimodal biometrics has the highest accuracy of 98.25% with the lowest ERR of 3.26%, which is far better than the traditional voice-based methods where the average EER is 26.01% [20]. The results show that our proposed multimodal biometrics is very effective in detecting replay attacks and outperform traditional voice-based methods.

G. Impact of the Number of Subjects

Here, we investigate the performance of our system with different numbers of subjects. There are five participants in our experiments. One of them is regarded as the legitimate user and the other subjects are unauthorized speakers. We use the data set that includes one unauthorized speaker, two unauthorized speakers, three unauthorized speakers, or four unauthorized speakers to perform speaker verification, respectively. Fig. 15 depicts the verification accuracy with different numbers of subjects. We observe that the accuracy of multimodal biometrics is slightly dropped from 96.23% to 95.93%. The effect of increasing the number of subjects is not significant. We believe that our system can still maintain high verification accuracy to handle a large number of subjects.

H. Impact of Radar Placement

In practical applications, the placement of the radar senor might vary, which leads to different relative positions of the radar and a user. The experimental setup in Fig. 12 shows a 0° placement, where the radar is parallel to the plane of the human body. We, respectively, turn the radar to 15°, 30°, and 45° to compare the performance under different placement angles.

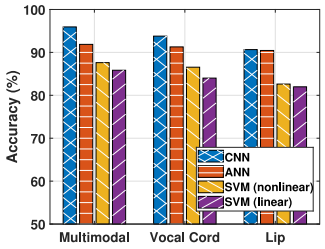


Fig. 17. Comparison of the performance of different classifiers.

Fig. 16 presents the accuracy of our system under different angle degrees of the radar placements. We observe that our system is highly effective even the radar placement is changed. In particular, multimodal biometrics can achieve an accuracy of around 95% regardless of radar placement angles. Similarly, different radar placements do not impact the performance of the system using the vocal cord biometrics. However, the lip biometrics method is significantly affected by the change of radar placement, which implies that the lip biometrics alone is not that robust compared with the other methods. These results show that our system based on multimodal biometrics does not require the user to put the radar at a specific position, which indicates that our system is robust while using in different placement scenarios.

I. Impact of Classifiers

Next, we investigate the impact of different classifiers on system performance. In the experiments presented in Sections V-E–V-H, we use our CNN (Fig. 11) as the classifier. We then compare it with three other classifiers, including an artificial neural network (ANN) that has no convolutional layers, support vector machine (SVM) with a nonlinear kernel, and SVM with a linear kernel. Fig. 17 shows the comparison of accuracy using different classifiers. We observe that the designed CNN has the best performance and ANN ranks second in accuracy, which suggests that the convolutional layer makes a difference to the performance of the neural network. Both SVM with a nonlinear kernel and SVM with a linear kernel fall behind the neural network classifiers and they achieve the accuracy below 88%. At a high level, we can see that deep learning methods (i.e., CNN and ANN) outperform machine learning methods (i.e., SVM). In particular, the results show that the designed neural network with convolutional layers enables us to achieve the best performance.

J. Impact of Distance

To explore the effective sensing range of the proposed system, we evaluate the impact of distance on performance, where the distance refers to the range between the mmWave radar and subjects. Due to the hardware restriction of the AWR1642 mmWave radar used in experiments, the transmission power can be increased in a limited way. Thus, with the maximum transmission power, the effective sensing distance of our radar device is about 2 m. Fig. 18(a) shows the

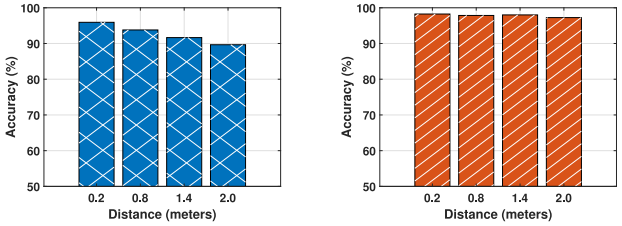


Fig. 18. System performance of different radar–subject distances. (a) Verification accuracy. (b) Accuracy of detecting replay attacks.

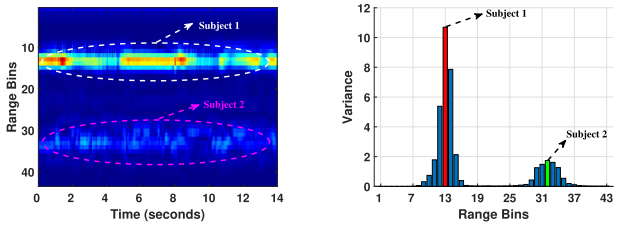


Fig. 19. Investigation of multiple subjects. (a) Magnitude of FFT results of beat signals. (b) Range selection.

verification accuracy values for different radar–subjects distances. As expected, the performance drops with the increase of distance. That is because the amplitude of phase change values becomes diminishing with the limited transmission power in a long distance, which makes our system harder to detect the VCVs. However, we believe that the verification accuracy can be good if the transmission power could be adjusted to a reasonable value. In addition, as illustrated in Fig. 18(b), the accuracy of replay attack detection stays high even when the distance increases to 2 m. This result shows that our system is robust against replay attacks with the impact of different distances. In Section VI, we discuss the approach of extending the sensing range in future work.

K. Investigation of Multiple Subjects

The mmWave radar has the capability to work with multiple subjects. Multiple subjects in the sensing range can be detected accurately. Radar beat signals of different subjects could be extracted and processed separately in our system. We perform experiments to verify if our system works when multiple subjects are in the different locations of the sensing range. Two subjects have a 14-s speaking activity simultaneously at the distance of 0.6 and 1.5 m away from the radar. The beat signals of all 43 range bins are illustrated in Fig. 19(a). The radar sensing range is adjusted to 2 m, so each rang bin includes the beat signals of every 0.046 m [i.e., $(2\ \text{m}/43) = 0.046\ \text{m}$]. We can see that highlighted parts indicating speaking activities are around 13th range bin (i.e., $13 \times 0.046\ \text{m} = 0.59\ \text{m}$) and 32th range bin (i.e., $32 \times 0.046\ \text{m} = 1.47\ \text{m}$), which exactly correspond to the locations of two subjects. Then, as shown in Fig. 19(b), we compute the variance of the beat signal magnitude at each range bin and find two local maximums at the 13th range bin and 32nd range bin. The beat signals of these two range bins can be extracted for the following steps (e.g., phase extraction, signal segmentation, feature extraction, and CNN-based classification), which are described in Section IV. Thus, the experimental results show that our system is able to deal with the situation of multiple subjects.

VI. Discussion

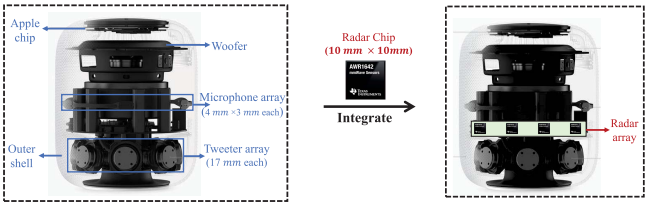
In this section, we discuss some situations where our system is applied to the real-world usage scenarios.

Integrating mmWave Radar to Voice Assistant Device: The device illustrated in Fig. 12 is an embedded system broad that involves a mmWave radar chip, a DSP chip, and a lot of interfaces, which is 5 cm $\times$ 8 cm in size. To apply the proposed method to real-world IoT smart homes, we need to integrate the mmWave radar chip to a voice assistant device instead of integrating a whole embedded system broadly, since voice assistant devices all come with their own DSP chips and interfaces. We can use the Apple Homepod as an example shown in Fig. 20(a) to see the internal chips and components of a voice assistant device. Moreover, the size of the mmWave radar chip is between the size of a microphone and the size of a tweeter in a voice assistant device. For instance, the size of an AWR1642 radar chip [45] used in our work is 10 mm $\times$ 10 mm. The size of a microelectromechanical system [46] (MEMS) microphone, that is used in voice assistant devices (e.g., Google Home and Apple Home Pod), is 4 mm $\times$ 3 mm. The tweeter size in a typical voice assistant device is 17 mm in diameter [47].

Therefore, in terms of the mmWave radar size and the hardware context in voice assistant devices, we believe that integrating radar chips into a voice assistant device is entirely feasible. Fig. 20(b) illustrates a diagram about the integration of the mmWave radar into an Apple Homepod, where we show a possible setup integrating a 4-mmWave-radar array to an Apple Homepod.

Extending Effective Sensing Range of mmWave Radar: We have performed experiments and discussed the impact of sensing distance in Section V-J. Due to the limit of transmission power in hardware, the effective sensing range of the AWR1642 radar is around 2 m. However, the mmWave radar is capable of sensing much farther by increasing the transmission power. Therefore, the mmWave radar rivals the sensing range of microphones. It is thus desirable to apply the mmWave radar in the voice assistant devices for IoT smart homes.

Eliminating Impact of Nearby Individuals or Objects: As demonstrated in Section V-K, our system could detect multiple objects in the radar sensing range. The range resolution is 0.046 m, which means that our system can distinguish between objects that are very close at a distance of 0.046 m. Therefore, if there are moving individuals or objects (e.g., cats or dogs) nearby users, our radar can separately capture and process their signals, and utilize the deep CNN model to determine whether these signals belong to legitimate users or not. Furthermore, as using a microphone array to extend the effective sensing range in the current voice assistant devices, our system is also able to adopt the mmWave radar array instead of using one signal radar. For instance, we can integrate four mmWave radar chips to four cardinal points of a voice assistant device, which can achieve the all-around sensing ability.



(a) (b)

Fig. 20. Integration of mmWave radar chips into an Apple Homepod voice assistant. (a) Internals of an Apple Homepod. (b) Integration with mmWave radar.

Using Advanced Deep Learning Model to Enhance System Robustness: Our current classifier is a deep CNN model, which outperforms ANN and SVM models and achieves high accuracy for verification and replay attack detection. To cope with the complex noises and situations in the real-world IoT smart homes, our system may utilize a more advanced deep learning model in the future.

As mentioned before, VCV has a high frequency (i.e., [90 Hz, 200 Hz]), which is hard to have any consistency with other movements. The IIR filter can accurately isolate the VCV signal from other movement signals. However, the frequency of LM is between 0.2 and 3 Hz, which is easy to have overlaps with some human movements, such as breathing and heartbeat. Also, a user could hold a cat in arms when speaking voice commands, where the movements of the cat are also captured by the radar together with the user's LMs. Therefore, the LM signals extracted using the IIR filter contain harmonics caused by these movements. It is one of the reasons that lip biometrics are not as good as vocal cord biometrics in our system.

To address this issue, we can adopt a deep denoising autoencoder [43] in our classifier to reduce the noises and harmonics of inputs. Also, we can model the noises using an adversarial network [44] and jointly train our classifier with the adversarial network, which can counter the various noises.

Using Data Fusion of Multiple Antennas to Improve System Performance: In this study, we use the data from one single TX–RX antenna pair due to the memory limit of the DSP unit. With a high-performance DSP, we are considering to use data fusion technologies (e.g., the Kalman filter [48]) to leverage more information from other antenna pairs, which may further improve the performance of our approach.

VII. Conclusion

In this article, we introduced a new multimodal biometrics system for noncontact speaker verification based on the mmWave radar for IoT smart home applications. Our system measures the unique VCV and LM of individuals with regard to the exclusive structure of vocal cord and special articulatory gestures in individuals. The usage of the mmWave radar makes our system not only be noncontact and unobtrusive but also counter imposter attacks very effectively. The evaluation with extensive experiments shows that our system has a high verification accuracy and low EER under various conditions, such as with different placements of radar and under replay attacks. Our system can achieve verification accuracy of over 95% and detect replay attacks with over 98% accuracy and the EER at around 3%.

References

[1] (Aug. 2019). *7 Key Predictions for the Future of Voice Assistants and AI: Clearbridge*. [Online]. Available: https://clearbridgemobile.com/7-key-predictions-for-the-future-of-voice-assistants-and-ai/

[2] (Feb. 2019). *Analyst: 8 Billion Voice Assistants by 2023*. [Online]. Available: https://searchengineland.com/analyst-8-billion-voice-assistants-by-2023-312035

[3] (Jul. 2019). *Inside the Smart Home: IoT Device Threats and Attack Scenarios*. [Online]. Available: https://www.trendmicro.com/vinfo/us/security/news/internet-of-things/inside-the-smart-home-iot-device-threats-and-attack-scenarios

[4] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[5] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Shanghai, China, 2016, pp. 5115–5119.

[6] S.-X. Zhang, Z. Chen, Y. Zhao, J. Li, and Y. Gong, "End-to-end attention based text-dependent speaker verification," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, San Diego, CA, USA, 2016, pp. 171–178.

[7] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Process. Mag.*, vol. 32, no. 6, pp. 74–99, Nov. 2015.

[8] Z. Wu, S. Gao, E. S. Cling, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *Proc. IEEE Asia–Pac. Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA)*, Siem Reap, Cambodia, 2014, pp. 1–5.

[9] Z. Wu and H. Li, "Voice conversion and spoofing attack on speaker verification systems," in *Proc. IEEE Asia–Pac. Signal Inf. Process. Assoc. Annu. Summit Conf.*, Kaohsiung, Taiwan, 2013, pp. 1–9.

[10] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Commun.*, vol. 66, pp. 130–153, Feb. 2015.

[11] M. Todisco and N. Delgado, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Proc. Odyssey Conf.*, vol. 45, 2016, pp. 283–290.

[12] F. Alegre, A. Amehraye, and N. Evans, "Spoofing countermeasures to protect automatic speaker verification from voice conversion," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Vancouver, BC, Canada, 2013, pp. 3068–3072.

[13] S. Shiota, F. Villavicencio, J. Yamagishi, N. Ono, I. Echizen, and T. Matsui, "Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification," in *Proc. 6th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 239–243.

[14] M. Smiatacz, "Playback attack detection: The search for the ultimate set of antispoof features," in *Proc. Int. Conf. Comput. Recognit. Syst.*, 2017, pp. 120–129.

[15] A. Hannun *et al.*, "Deep speech: Scaling up end-to-end speech recognition," 2014. [Online]. Available: arXiv:1412.5567.

[16] X. Liu and Y.-M. Cheung, "Learning multi-boosted HMMs for lip-password based speaker verification," *IEEE Trans. Inf. Forensics Security*, vol. 9, pp. 233–246, 2014.

[17] Y.-M. Cheung and X. Liu, "Lip-password based speaker verification system," U.S. Patent 9 159 321, Oct. 2015,

[18] M. Sahidullah *et al.*, "Robust voice liveness detection and speaker verification using throat microphones," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 1, pp. 44–56, Jan. 2018.

[19] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Proc. Odyssey Conf.* , 2010, p. 14.

[20] T. Kinnunen *et al.*, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. Int. Speech Commun. Assoc. (ISCA)*, 2017, pp. 2–6.

[21] D. Ishac, A. Abche, G. Nassar, E. Karam, and D. Callens, "Speaker identification based on vocal cords 'vibrations' signal: effect of the window," in *Proc. 3rd Int. Conf. Elect. Electron. Eng. Telecommun. Eng. Mechatronics (EEETEM)*, 2017, pp. 131–135.

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

[23] Y. Meng *et al.*, "WiVo: Enhancing the security of voice control system via wireless signal in IoT environment," in *Proc. 18th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, 2018, pp. 81–90.
[24] S. Li *et al.*, "A 94-GHz millimeter-wave sensor for speech signal acquisition," *Sensors*, vol. 13, no. 11, pp. 14248–14260, 2013.
[25] F. Chen *et al.*, "A novel method for speech acquisition and enhancement by 94 GHz millimeter-wave sensor," *Sensors*, vol. 16, no. 1, p. 50, 2016.
[26] C. Xu *et al.*, "WaveEar: Exploring a mmwave-based noise-resistant speech sensing for voice-user interface," in *Proc. 17th Annu. Int. Conf. Mobile Syst. Appl. Serv.*, 2019, pp. 14–26.
[27] C.-S. Lin, S.-F. Chang, C.-C. Chang, and C.-C. Lin, "Microwave human vocal vibration signal detection based on doppler radar technology," *IEEE Trans. Microw. Theory Techn.*, vol. 58, no. 8, pp. 2299–2306, Aug. 2010.
[28] H. Hong *et al.*, "Time-varying vocal folds vibration detection using a 24 GHz portable auditory radar," *Sensors*, vol. 16, no. 8, p. 1181, 2016.
[29] F. Chen, S. Li, Y. Zhang, and J. Wang, "Detection of the vibration signal from human vocal folds using a 94-GHz millimeter-wave radar," *Sensors*, vol. 17, no. 3, p. 543, 2017.
[30] M. Young, *Singing the Body Electric: The Human Voice and Sound Technology*. Abingdon, U.K.: Routledge, 2016.
[31] I. R. Titze, "The human vocal cords: A mathematical model," *Phonetica*, vol. 28, nos. 3–4, pp. 129–170, 1973.
[32] S. Kiritani, H. Hirose, and H. Imagawa, "High-speed digital image analysis of vocal cord vibration in diplophonia," *Speech Commun.*, vol. 13, nos. 1–2, pp. 23–32, 1993.
[33] M. Chora, "Human lips as emerging biometrics modality," in *Proc. Int. Conf. Image Anal. Recognit.*, 2008, pp. 993–1002.
[34] S. Rao, A. Ahmad, J. C. Roh, and S. Bharadwaj, *77 GHz Single Chip Radar Sensor Enables Automotive Body and Chassis Applications*, Texas Instrum., Dallas, TX, USA, 2018.
[35] J. Lohscheller, J. G. Švec, and M. Döllinger, "Vocal fold vibration amplitude, open quotient, speed quotient and their variability along glottal length: Kymographic data from normal subjects," *Logopedics Phoniatrics Vocol.*, vol. 38, no. 4, pp. 182–192, 2013.
[36] D. M. Sheen, D. L. McMakin, and T. E. Hall, "Three-dimensional millimeter-wave imaging for concealed weapon detection," *IEEE Trans. Microw. Theory Techn.*, vol. 49, no. 9, pp. 1581–1592, Sep. 2001.
[37] K. Ramasubramanian, *Using a Complex-Baseband Architecture in FMCW Radar Systems* Texas Instrum., Dallas, TX, USA, 2017.
[38] Y. Dong, J. Liu, Y. Chen, and W. Y. Lee, "SalsaAsst: Beat counting system empowered by mobile devices to assist salsa dancers," in *Proc. IEEE 14th Int. Conf. Mobile Ad Hoc Sens. Syst. (MASS)*, Orlando, FL, USA, 2017, pp. 81–89.
[39] I. R. Titze and E. J. Hunter, "Normal vibration frequencies of the vocal ligament," *J. Acoust. Soc. Amer.*, vol. 115, no. 5, pp. 2264–2269, 2004.
[40] C. E. Efstathiou. *Signal Sampling: Nyquist-Shannon Theorem*. Accessed: Aug. 2, 2020. [Online]. Available: http://195.134.76.37/applets/AppletNyquist/Appl_Nyquist2.html
[41] R. N. Khushaba, S. Kodagoda, S. Lal, and G. Dissanayake, "Driver drowsiness classification using fuzzy wavelet-packet-based feature-extraction algorithm," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 1, pp. 121–131, Jan. 2011.
[42] B. Gambäck and U. K. Sikdar, "Using convolutional neural networks to classify hate-speech," in *Proc. 1st Workshop Abusive Language Online*, 2017, pp. 85–90.
[43] X. Feng, Y. Zhang, and J. Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Florence, Italy, 2014, pp. 1759–1763.
[44] H. Wang and C.-N. Yu, "A direct approach to robust deep learning using adversarial networks," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–15.
[45] *AWR1642 mmWave Radar*, Texas Instrum., Dallas, TX, USA, Apr. 2020. [Online]. Available: https://www.ti.com/product/AWR1642
[46] *IM69D130 Mems Microphone*, Infineon Technol., Neubiberg, Germany, Dec. 2017. [Online]. Available: https://www.infineon.com/cms/en/product/sensor/mems-microphones/im69d130/
[47] C. G. Shannon Liao. (Oct. 2017). *Battle of Smart Speakers*. [Online]. Available: https://www.theverge.com/2017/10/5/16425142/google-home-mini-vs-amazon-echo-dot-max-apple-homepod
[48] Q. Gan and C. J. Harris, "Comparison of two measurement fusion methods for Kalman-filter-based multisensor data fusion," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 37, no. 1, pp. 273–279, Jan. 2001.

Yudi Dong (Graduate Student Member, IEEE) received the B.Eng. degree in optoelectronic information engineering from the University of Shanghai for Science and Technology, Shanghai, China, in 2014, and the M.Sc. degree in electrical engineering from the Stevens Institute of Technology, Hoboken, NJ, USA, in 2017, where he is currently pursuing the Ph.D. degree.

His current research interests include deep learning, mobile security, and wireless communication.

Yu-Dong Yao (Fellow, IEEE) received the B.Eng. and M.Eng. degrees in electrical engineering from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 1982 and 1985, respectively, and the Ph.D. degree in electrical engineering from Southeast University, Nanjing, in 1988.

He was a Visiting Student with Carleton University, Ottawa, ON, Canada, from 1987 to 1988. He has been with the Stevens Institute of Technology, Hoboken, NJ, USA, since 2000, and served as the Department Chair of Electrical and Computer Engineering from 2007 to 2018. From 1989 to 2000, he was with Carleton University, Spar Aerospace Ltd., Montreal, QC, Canada, and Qualcomm Inc., San Diego, CA, USA. He is also the Director of the Stevens' Wireless Information Systems Engineering Laboratory. He holds one Chinese patent and 13 U.S. patents. His research interests include wireless communications, cognitive radio, and machine learning, and deep learning.

Dr. Yao received the Master of Engineering (Honoris Causa) from the Stevens Institute of Technology in 2018. He was an Associate Editor of the IEEE COMMUNICATIONS LETTERS from 2000 to 2008, and the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY from 2001 to 2006, and an Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2001 to 2005. For his contributions to wireless communications systems, he was elected a Fellow of the National Academy of Inventors in 2015, and the Canadian Academy of Engineering in 2017.