VocalPrint: A mmWave-Based Unmediated Vocal Sensing System for Secure Authentication

Huining Li, Chenhan Xu, Aditya Singh Rathore, Zhengxiong Li, *Member, IEEE*, Hanbin Zhang, *Member, IEEE*, Chen Song, *Member, IEEE*, Kun Wang, *Senior Member, IEEE*, Lu Su, *Senior Member, IEEE*, Feng Lin, *Senior Member, IEEE*, Kui Ren, *Fellow, IEEE*, and Wenyao Xu, *Senior Member, IEEE*

Abstract—With the continuing growth of voice-controlled devices, voice metrics have been widely used for user identification. However, voice biometrics is vulnerable to replay attacks and ambient noise. We identify that the fundamental vulnerability in voice biometrics is rooted in its indirect sensing modality (e.g., microphone). In this paper, we present *VocalPrint*, a resilient mmWave interrogation system which directly captures and analyzes the vocal vibrations for user authentication. Specifically, *VocalPrint* exploits the unique disturbance of the skin-reflect radio frequency (RF) signals around the near-throat region of the user, caused by the vocal vibrations. The complex ambient noise is isolated from the RF signal using a novel resilience-aware clutter suppression approach for preserving fine-grained vocal biometric properties. Afterward, we extract the vocal tract and vocal source features and input them into an ensemble classifier for authentication. *VocalPrint* is practical as it allows the effortless transition to a smartphone while having sufficient usability due to its non-contact nature. Our experimental results from 41 participants with different interrogation distances, orientations, and body motions show that *VocalPrint* achieves over 96 percent authentication accuracy even under unfavorable conditions. We demonstrate the resilience of our system against complex noise interference and spoof attacks of various threat levels.

Index Terms—mmWave sensing, voice authentication, biometrics

1 Introduction

Due to the growth of voice-controlled devices and services, the use of vocal-based biometrics for user authentication has surged [1], [2]. Voiceprint is a strong physiological and behavioral combined biometrics [3], considered to be just as biologically unique in individuals as a fingerprint [4]. There is a sizable literature on user identification by analyzing voice, including speech [5] and non-speech [6] vocal data. Commodity voice-controlled devices, such as the Amazon Echo and Google Home, have integrated speaker identification functions to secure the user information [7].

However, there are several major security limitations for adopting voice biometric technologies in real-world applications [8]. For example, fraudsters may eavesdrop the legitimate user's speech samples or utilize a variety of artificial intelligence technologies to generate synthetic voice data [9], and then launch a "replay attack" against voice-based authentication systems. How to defend against the playback attack has a long and rich history, and is a core research problem in biometric security [10], [11], [12]. Researchers have studied sets of software-based solutions based on liveness distinction between human and loudspeakers, including challenge-response protocols [13], ultrasonic reflections of mouth motion [14], time-difference-of-arrival (TDoA) of phenome sounds to two microphones [15], sound field difference [16], etc. Although these approaches could alleviate the security risk under some circumstances, they need user's active cooperation and also assume that replaying cannot generate identical sound waves. We discover that the fundamental vulnerability in voice biometrics is rooted in its *indirect sensing modality*. Currently, voice biometric systems mainly employ a microphone sensor. When the user speaks, the vocal folds vibrate. The generated sound propagates in the air media, and the air vibration is captured by a microphone. This kind of indirect voice sensing modality through a media creates an inevitable attack surface in the physical world (e.g., replay attacks using a high-definition loudspeaker or bionic loudspeaker arrays), which is hardly addressed by software-based approaches. Moreover, indirect voice sensing is prone to interference from ambient noises which decrease the usability of voice biometric systems, as it leads to false positives during authentication and is insensitive to minute alteration in fake voice input during replay attacks.

- *Huining Li, Chenhan Xu, Aditya Singh Rathore, Zhengxiong Li, Hanbin Zhang, and Wenyao Xu are with the Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY 14261 USA. E-mail: {huiningl, chenhanx, asrathor, zhengxio, hanbinzh, wenyaoxu}@buffalo.edu.*
- *Chen Song is with the Department of Computer Science, San Diego State University, San Diego, CA 92182 USA. E-mail: csong@sdsu.edu.*
- *Kun Wang is with the Department of Electrical and Computer Engineering, University of California, Los Angeles, CA 90095 USA. E-mail: wangk@ucla.edu.*
- *Lu Su is with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA. E-mail: lusu@purdue.edu.*
- *Feng Lin and Kui Ren are with the Department of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China. E-mail: {flin, kuiren}@zju.edu.cn.*

Manuscript received 19 Dec. 2020; revised 27 Apr. 2021; accepted 24 May 2021. Date of publication 31 May 2021; date of current version 5 Dec. 2022.
(Corresponding author: Huining Li.)
Digital Object Identifier no. 10.1109/TMC.2021.3084971

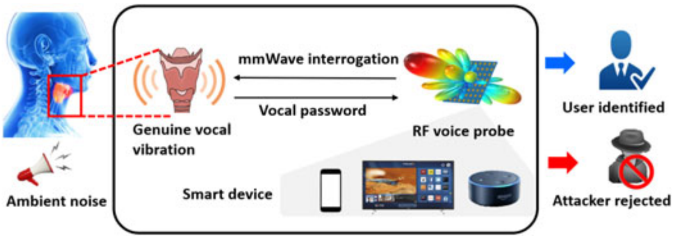


Fig. 1. A mmWave biometric interrogation system leveraging the vocal vibrations for accurate user authentication.

It is a fact that voiced sound is determined by vocal fold vibration, which is the root of voiceprint uniqueness [4]. On the basis of this argument, we propose that the most secure and attack-resistant voice sensing approach to user identification is to *directly* acquire and analyze the user's vocal fold vibration. Radio-frequency (RF) signals, such as millimeter wave (mmWave), shows immense potentials in sensing micron-level skin displacement [17], [18] due to their directional beamforming and skin-reflectance properties. A recent study demonstrated the feasibility of acquiring the vocal vibrations that occur at the range of 2-3mm via mmWave radar [19]. Motivated by these works, a *non-contact* and *unmediated* (direct) biometric mmWave interrogation system can be developed to capture the unique vocal vibrations for secure user authentication.

To realize our system, we need to address the following challenges: (1) How to suppress the complex noise clutters arising from static and dynamic objects in the environment and motion artifact for preserving fine-grained voice biometric properties in received mmWave response? (2) How to extract and identify the intrinsic features that can perfectly capture the vocal tract and vocal source information to maximize the system performance? (3) How to validate the resilience of our system against spoof attacks?

To this end, we present our system, *VocalPrint*, to facilitate a resilient mmWave interrogation system for secure and non-contact voice authentication, illustrated in Fig. 1. We leverage a low-cost, portable, and high-resolution 77 GHz Frequency Modulated Continuous Wave (FMCW) radar to identify the user from the dynamic environment and non-invasively sense the minute vocal vibrations. The displacement in the vocal vibrations is inferred from the phase shift of the peak corresponding to the human target in the intermediate frequency (IF) signals. To help reserve fine-grained voice biometric properties in the RF voice signals, we develop a resilient-aware assembled clutter suppression scheme to isolate random motion artifact and ambient noise clutter. Once the precise vocal vibration signals are obtained, we extract text-independent vocal source and vocal tract features, respectively, which closely relate to the human speech articulation. Finally, these text-independent biometric descriptors are fed into a fine-tuned feature selection module and an ensemble classifier for user authentication. To intensively evaluate our system, we recruit 41 volunteers with results showing that *VocalPrint* can enable a reliable authentication with over 96 percent accuracy. Furthermore, we validate the resilient security of *VocalPrint* against ambient interference (e.g., acoustic noise, dynamic environment, human obstruction) and spoofing attacks (e.g., counterfeit, mimicry, signal-based) to show its significant potential as an enhancement to voice authentication in real-world applications.

The contribution of our work has three-fold:

- We perform the first study to identify that the fundamental vulnerability in voice biometrics is rooted in its indirect sensing modality. We also explore an unmediated (direct) mmWave sensing approach to acquire and analyze the user's vocal fold vibration in a secure and attack-resistant manner.
- We develop *VocalPrint*, an end-to-end biometric system to facilitate resilient security of voice authentication. We first design a novel resilience-aware clutter suppression model to obtain precise vocal vibration data that reserves fine-grained biometric properties, and then extract intrinsic features that depict vocal source and vocal tract information for user identification.
- We demonstrate the effectiveness and robustness of *VocalPrint* through extensive experiments with results showing superior authentication accuracy even under unfavorable conditions. We conduct comprehensive studies to validate the resilience of *VocalPrint* against complex noise interference and spoof attacks of various threat levels.

2 Theory and Preliminaries

2.1 Unmediated Vocal Sensing

Microphone is widely employed in voice sensing, but its sensing mechanism requires air medium which can result in the vulnerability issues in voice biometrics system. Specifically, vocal fold vibrates and generates sound waves, and then the sound waves propagate in the air. They finally reach the microphone transducers, and are converted into electrical signals. Contrast to microphone, mmWave emission is the electromagnetic field propagation and do not need any other medium [20]. Therefore, acoustic noise that propagates in the air medium and results air vibration cannot enter into the mmWave channel. Moreover, such unmediated mmWave sensing modality can directly capture vocal fold vibration that is the root of voiceprint uniqueness and leverage anti-spoofing information (i.e.,throat physiological intrinsic) to defend against replay attacks. In addition, unmediated mmWave sensing also have other advantages over microphone sensing. For example, it can sense up to 50 meters and sense the voice through wall [21]. As for the price, although a mmWave sensor costs hundreds of dollars, but its price will go down in mass production.

2.2 Voice Biometrics Rationale

Voice can be regarded as physiological and behavioral combined biometrics, which contains unique and permanent information of individuals [4]. Specifically, voice permanence is derived from the fixed physical shape of individual's lung, vocal cords, and vocal tract. Voice uniqueness stems from the precise and coordinated vibration of the vocal cords and vocal tract [22], [23]. When a person speaks, the air flow is first expelled from the lungs and then traverses through the vocal cords. The vocal cords with the glottis constrict to block the air flow and the resulting vibrations in air produce voiced

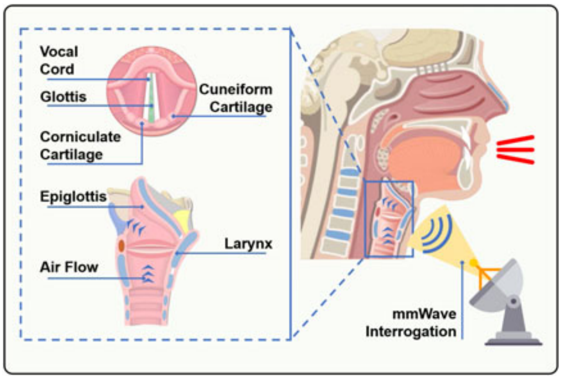


Fig. 2. The vocal vibrations result from dynamics of articulatory organs and can be sensed by the mmWave radar from near-throat region.

signals, including the vowels and some consonants such as *[b]*, *[v]*. In contrast, when the vocal cords with the glottis dilate, the air flow is allowed to pass through without heavy vibrations, thereby generating unvoiced signals. Afterward, both voiced and unvoiced signals are resonated and reshaped by the vocal tract consisting of multiple articulatory organs (e.g., epiglottis, corniculate cartilage, cuneiform cartilage, shown in Fig. 2). The movement of articulatory organs forms a path with specific geometrical shapes (i.e., articulatory gesture) for the air flow [24], which manipulates the amplitude and frequency of vocal vibrations. Although different people may share the same type of articulatory gesture when pronouncing the same phoneme, the movement speed and intensity vary from person to person and contain distinctive information. Moreover, the larynx modulates the tension on vocal cords to produce fine-tuned vocal vibrations, which further adds the uniqueness to an individual voice. Therefore, this uniqueness of the human voice is intrinsically sourced in vocal vibrations.

2.3 A Preliminary Study

There is a significant growing interest in human sensing applications using RF sensing [25], [26], [27]. Specifically, WaveEar [28] is one recent representative work on investigating speech recognition using mmWave technologies. To examine the feasibility of acquiring vocal biometric features in mmWave sensing, we conduct a preliminary study using a mmWave-band FMCW probe.

Preliminary Data Collection. In the preliminary experiment, we leverage a beamforming mmWave probe to sense

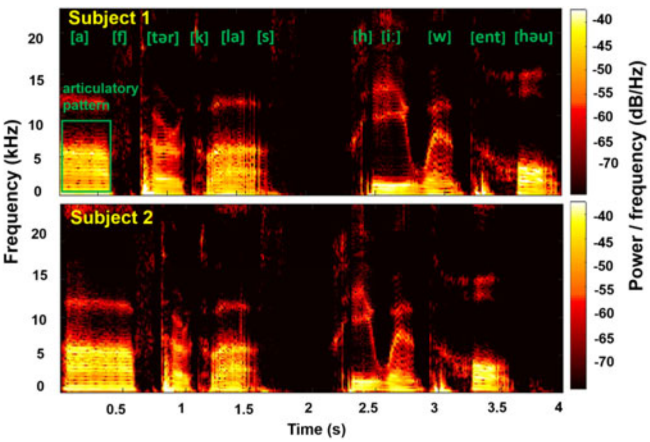


Fig. 3. Spectrogram of reconstructed voice.

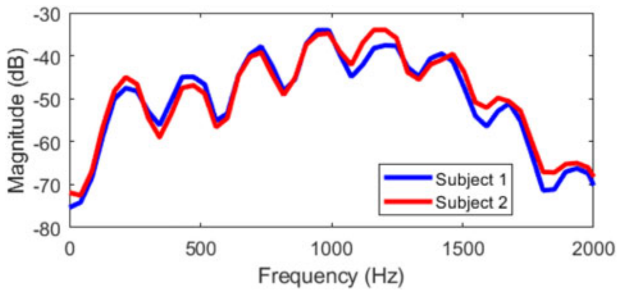


Fig. 4. Short-term spectral envelop.

the subject's vocal vibration and collect received mmWave signals. Specifically, two subjects are asked to sit in the same position and pronounce the sentence, "After class, he went home". For the ease of analysis, we align the mmWave probe in the direction of the subject's throat. The distance between the subject and the probe is 20 cm.

Auditory Analysis in mmWave Sensing. Speech recognition [29] and speaker identification [30] are both voice-based applications. However, underlying mechanisms and associated technologies are distinct. Speech recognition utilizes the temporal cues and envelopes in voice data and it is critical to capture and parse coarse-grained (e.g., hundreds of milliseconds to seconds) articulatory features (e.g., up/down/back/forth movement). Speaker identification exploits spectral information in voice data. For example, fine-grained (e.g., tens of milliseconds) spectral envelopes contain the resonance properties of vocal tracts and timber, which are the pivotal features in speaker identification. We adopted the analytical scheme in WaveEar [28] to reconstruct voice signals of both speakers. As shown in Fig. 3, both reconstructed voices have a similar spectrum and envelope (with a segment of 100 ms). The voice data can be successfully processed by the commodity speech recognition software kit [31]. However, as shown in Fig. 4, the short-term (10 ms) spectral envelopes in both reconstructed voices have a low resolution, and spectral poles in both spectrums are nearly the same. Biometric traits are lost in the mmWave-reconstructed voice data.

Summary. A new analytical scheme in processing mmWave signals is investigated for speaker identification. Particularly, short-term properties in vocal patterns need to be reserved and augmented. In the following sections, we will present (A) high-definition mmWave interrogation, (B) a resilience-aware clutter removal scheme using a novel assembled model, and (C) robust feature extraction and matching methods.

3 VocalPrint System Overview

In this paper, we introduce *VocalPrint*, a resilience-aware mmWave biometric interrogation system. The end-to-end system overview is shown in Fig. 5.

VocalPrint Hardware. A high-resolution mmWave probe is leveraged to accurately sense the vocal vibrations in a non-contact manner. Specifically, the probe first transmits a frequency modulated continuous wave towards the throat of the user and then receives the skin-reflect response signal which comprises sufficient information of the vocal vibrations when the user is speaking. The received signal is

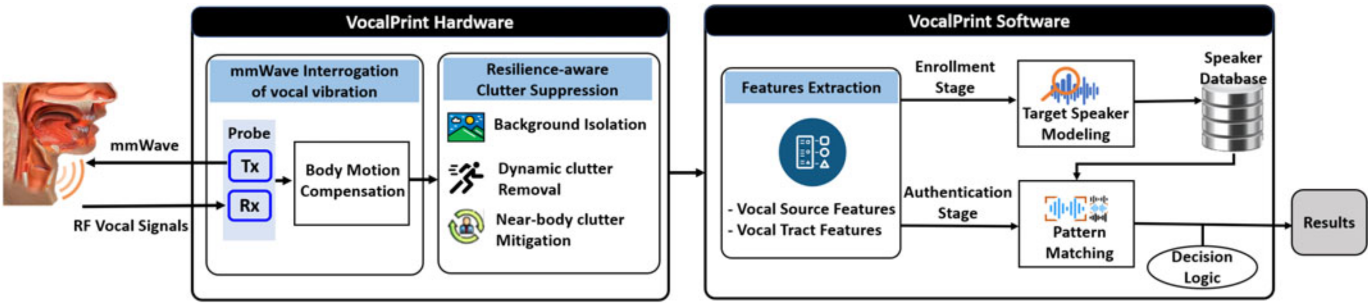


Fig. 5. The overview of *VocalPrint* mainly consisting of a mmWave interrogation module to sense the vocal vibrations, a resilience-aware clutter suppression module to remove the complex noise, and an authentication module to identify the legitimate user against imposters.

transmitted to a resilience-aware clutter suppression model for isolating the noise from the surrounding environment and even the dynamic obstruction caused by multiple human subject interference.

VocalPrint Software. Once the precise vocal vibration data is acquired, *VocalPrint* extracts optimal biometric features that depict vocal source and vocal tract information. After that, the vocal biometric descriptors are input to a fine-tuned authentication model that consists of a feature selection module and an ensemble classifier for identifying the legitimate user against imposters.

4 mmWave Interrogation of Vocal Vibrations

4.1 mmWave Probe Design and Integration

The Continuous Wave (CW) is increasingly used in sensing various vital signs, such as breathing and heartbeat, due to its ability to capture near-field motion and displacement [32]. However, CW is not accurate in range measurement because it lacks the timing mark (the frequency is fixed). Besides, CW cannot differentiate between two or more reflecting objects because the reflected signals and clutters are all mixed up in both the time and frequency domains. Therefore, we conclude that CW is not capable of authenticating a person at the non-pre-known position in a complex environment. In *VocalPrint*, we leverage FMCW, which can detect both accurate range and minute displacement. Moreover, FMCW enables a low-frequency received signal processing by the mixed IF signal, which considerably reduces the loading of designing and realizing the circuit. In the next part, we give the formal description of the continuous vocal vibration interrogation utilizing the FMCW mmWave. Based on the interrogation theory, we give more discussion about the mmWave probe parameters in Section 7.1.

4.2 Continuous Vocal Vibration Interrogation

To enable the continuous vocal vibration interrogation, FMCW modulates a saw-tooth baseband (used as timing mark) to the high-frequency mmWave carrier. Specifically, the periodic chirp signal $T(t)$ transmitted to the speaking person's throat is defined as

$$T(t) = \exp\left[j\left(2\pi f_0 t + \int_0^t 2\pi\rho t\ dt\right)\right], \tag{1}$$

where $0 < t < T_r$, f_0 is the carrier frequency, T_r is one chirp cycle, B is the bandwidth of one chirp, and $\rho = B/T_r$ is the chirp rate. Assume that the distance between the radar and human throat is $X(t) = X_0 + d(t)$, where X_0 is the original distance, $d(t)$ represents the minute skin displacement caused by vocal vibration. With the round trip delay t_d lagged behind the transmitted chirp signal, the received signal consists of the clutter components $R_{\text{clutter}}(t)$ and the vocal component $R(t)$ carrying the vocal vibration, which is

$$R(t) = \Gamma\ \exp[j(2\pi f_0(t - t_d) + \int_0^{t-t_d} 2\pi\rho t\ dt)], \tag{2}$$

where $t_d < t < T_r + t_d$, Γ denotes the amplitude normalized to the transmitted chirp signal, and $t_d = \frac{2[X_0+d(t)]}{c}$. The clutter suppression is studied further in Section 5.

For every chirp, the valid time period for mixing is (t_d, T_r), and thereby the IF signal for a chirp after mixing can be obtained as

$$H(t) = T(t) \times R^*(t) \approx \Gamma\ \exp[j(2\pi\rho t_d t + 2\pi f_0 t_d)], \tag{3}$$

where $*$ represents a conjugate transpose operation, $\times$ is the mixer, the mathematical term related to t_d^2 is left out due to $t_d^2 \ll t_d t,\ \ t_d < t < T_r$. From Eq. (3), we can see that the mixed IF signal is directly related to the skin displacement caused by vocal vibration $d(t)$. Since $d(t)$ is very small during one chirp, we track the IF signal across a sequence of M chirps. Substituting $t_d = \frac{2[X_0+d(t)]}{c}$, the IF signal for the m-th chirp period can be formulated as

$$\begin{aligned} &H(mT_r + t) = \\ &\Gamma\ \exp\left\{\left[\frac{4\pi\rho X_0}{c}t + \frac{4\pi f_0 X_0}{c} + \left(\frac{4\pi\rho t}{c} + \frac{4\pi f_0}{c}\right)d(mT_r)\right]\right\}, \end{aligned} \tag{4}$$

where c denotes the light speed. Because of $\rho t \ll f_0$ $(t \in (t_d, T_r))$ in typical FMCW radars, $\frac{4\pi\rho t}{c}$ can be neglected. Then, $H(mT_r + t)$ can be obtained as

$$\begin{aligned} &H(mT_r + t) = \Gamma\ \exp[j(w_H t + \psi_m)], \\ &\quad w_H = \frac{4\pi\rho X_0}{c},\ \ \psi_m = \frac{4\pi f_0 X_0 + 4\pi f_0 d(mT_r)}{c}. \end{aligned} \tag{5}$$

Therefore, the vibration displacement during the m-th chirp period $d(mT_r)$ can be calculated as

$$d(mT_r) = \frac{c}{4\pi f_0}\Delta\psi_m, \tag{6}$$

where $\Delta\psi_m$ can be achieved by conducting Fast Fourier Transform (FFT) on the IF signals for a sequence of M chirps.

4.3 Body Motion Compensation

Random body motion from users is the main interference in the received mmWave signals. For every FMCW chirp, the throat-reflected mmWave signals mixed with the interference are naturally distributed into various bins corresponding to range information in the frequency domain, i.e., range profile. Body motion introduces undesired shifts (misalignment) among range profiles of consecutive chirps. Since the location of the target user is unknown and may change over time, it is difficult to identify (changeable) range bin related to the human body. Consequently, we solve the misalignment problems based on the whole range-Doppler-matrix (RDM). A conventional countermeasure is to leverage a digital filter and compensate for the body motion. However, such a solution can introduce cumulative errors over time that are hardly dealt with. To address this issue, we develop a fine-grained range profile alignment solution.

We first define some variables. $S_m(l)$ is denoted as the m-th achieved range profile, $\widetilde{S}_m(l)$ is denoted as the m-th aligned range profile, where $m = 0, \ldots, M-1$, $l = 0, \ldots, L-1$, M is the number of the range profiles (i.e., the number of chirps), and L is the number of range bins. χ_m represents the range bin shift applied to the $S_m(l)$. We also define the reference range profile by using the knowledge of the latest aligned range profiles, formulated as

$$Q_{m+1}(l) = \frac{m}{m+1} Q_m(l) + \frac{1}{m+1} |\widetilde{S}_m(l)|, \tag{7}$$

where $Q_m(0) = \widetilde{S}_m(0) = S_m(0)$ at the beginning.

Then, we formulate the envelope correlation function between the motion-shifted range profile and its corresponding reference range profile, written as

$$\Pi(\chi_{m+1}) = \sum_{l=0}^{L-1} |Q_{m+1}(l)| \cdot |S_{m+1}(l - \chi_{m+1})|. \tag{8}$$

The maximum value of the envelope correlation function indicates the optimum alignment between the shifted and the reference range profile. To achieve fine-grained correlation function optimization, there are mainly two stages. First, we calculate the integer part of the shift to maximize the value of the correlation function. Second, we use the Nelder-Mead algorithm [33] iteratively to find the optimum range shift for achieving local maximum, and the previous integer part is regarded as the initial guess for the iteration. Finally, we can obtain the aligned range profile as

$$\begin{aligned} \widetilde{S}_{m+1}(l) &= S_{m+1}(l - \chi_{m+1}^{opt}) \\ &= \mathrm{FFT}\left\{\exp\left(j(2\pi \frac{\chi_{m+1}}{L} \triangle\right) \mathrm{IFFT}\{S_{m+1}(l)\}\right\}, \end{aligned} \tag{9}$$

where $\triangle$ is the vector $[0, 1, \ldots, L-1]^\top$. After finishing this process, the next range profile will be aligned in the iteration.

Preliminary Results. In our preliminary work, we collected the data of the body motion artifacts from a mmWave hardware platform and simulated the proposed method. The subject is asked to sit in the direction of the mmWave probe and randomly wobble his upper body when speaking. Other experiment settings are the same as we mentioned in Section 2.2. Fig. 6 shows the range-Doppler-matrix (RDM) before (a) and after (b) body motion removal. RDM is the result of the frequency domain analysis among multiple range profiles, which can illustrate noises and signal-of-target. It indicates that our proposed method can eliminate interference from random body motion, but the clutter on the background still exists. Next, we will introduce the resilience-aware clutter suppression approaches to enhance voice features in RF voice data.

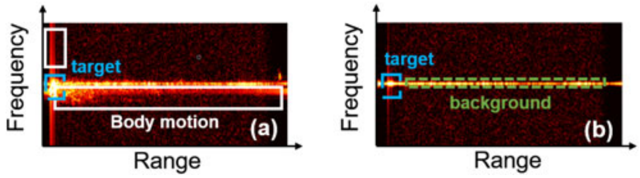


Fig. 6. The RDM before (a) and after (b) the body motion compensation.

5 Resilience-Aware Clutter Suppression

Undesirable backscatters caused by static/dynamic surrounding objects are main barriers in acquiring precised voice because they may disturb short-term spectral properties. Therefore, we investigate a resilience-aware clutter removal scheme using a novel assembled model to preserve voice biometric features in RF streaming signals.

5.1 Background Clutter Isolation

The background clutter in reflected mmWave signals is more complicated than that in conventional acoustic signals. Specifically, outdoor (e.g., snow and rain) and indoor (e.g., furniture and computer) background are both able to reflect the high-frequency mmWave [34]. Due to various reflection rates and the multipath interference, it is hard to isolate these objects by simply applying a threshold or training a classifier. Therefore, we regard the background clutter as the accumulation of the reflected signal by many small parts of the background, such as table legs, chairs, and monitors. Since the Weibull model provides the potential to accurately characterize the real clutter distribution over a much wider range of conditions than either the log-normal or Rayleigh model [35], we model the background clutter using Weibull distribution. The amplitude and phase of these reflections are random and its spectrum envelope a obeys to the following Weibull distribution [36] about the clutter:

$$p(a) = \frac{na^{n-1}}{\mu^n} \exp\left[-\left(\frac{a}{\mu}\right)^n\right], \tag{10}$$

where n and μ are the shape and scale parameters of the distribution, respectively.

To isolate the background clutter, we first arrange the range profiles achieved in Eq. (9) to a matrix (M rows $\times$ L columns) and perform the second FFT chirp-wise (slow-time FFT) for obtaining range-Doppler-matrix (RDM). Then, we conduct target searching (isolation) on the log-normalized RDM. Here, we define a resilient matrix as $u = \mathrm{C}(\mathcal{R})$, where $\mathcal{R}$ represents the RDM. Due to the complexity of the background clutter, the resilient function C is an $M \times L$ matrix of functions, its elements are generated from the normalization process based on the unbiased estimation of mean spectrum envelope energy and the standard deviation calculated by area element-averaging, formulated as

$$c_{ij} = \frac{\ln|\mathcal{R}_{ij}| - \hat{\mathbb{E}}(a)}{\mathrm{std}_a}, \tag{11}$$

where $\mathrm{std}_a = \frac{\pi}{n\sqrt{6}}$. $\hat{\mathbb{E}}(a)$ is the unbiased estimation of the $\mathbb{E}(a)$ [37], which can be calculated by

$$\hat{\mathbb{E}}(a) = \frac{1}{10}\left(\sum_{i-12,j-3}^{i+12,j+3}|\mathcal{R}_{ij}| - \sum_{i-10,j-2}^{i+10,j+2}|\mathcal{R}_{ij}|\right), \tag{12}$$

where ± 12 and ± 3 characterizes the window size, ± 10 and ± 2 characterizes the size of guard cell. We conduct a preliminary experiment to select the appropriate window size and the guard cell size. Specifically, one subject is asked to speak towards the probe in a fixed position with some desks and chairs as background. After achieving RF voice signals, we adjust the window length from ± 12 to ± 16, window width from ± 3 to ± 5, guard cell length from ± 8 to ± 10, and guard cell width from ± 0 to ± 2 in the background clutter isolation module. By comparing the signal-to-noise ratio (SNR) of RDM with different window sizes and guard cell sizes, we select the appropriate values. Finally, we set a resilient threshold u_0 to isolate the background clutter and update the RDM by assigning 0 to the RDM element whose corresponding element value in resilient matrix greater than u_0. According to the Eqs. (10) and (11), the clutter isolation rate p_c is formulated as

$$p_c = \exp\left[-\exp\frac{\pi}{\sqrt{6}}u_0 - \gamma\right], \tag{13}$$

where γ is the Euler-Mascheroni constant. The above equation indicates that the clutter isolation rate depends on the resilient threshold, which is a constant false alarm rate (CFAR). Here we set p_c to 10^{-6} and u_0 to 2.5 in the *VocalPrint* accordingly.

5.2 Dynamic Clutter Removal

Although background clutter is isolated, the moving objects, such as passersby and vehicles, can cause dynamic clutter that cannot be addressed by applying the resilient threshold on RDM. This is because its amplitude of spectrum envelop does not obey the aforementioned Weibull distribution. Intuitively, if an object keeps moving during consecutive chirps, it will move for a distance and shift to the next range bin, and thereby making certain element's amplitude among consecutive RDMs change to zero. Based on this theory, we make an element-wise comparison among D consecutive RDMs $\mathcal{R}^i (i \in [1, D])$, and check each element's amplitude. If $\exists i, j, k : \mathcal{R}^i_{jk} = 0$ is true, the $\mathcal{R}^1_{jk}$ will be regarded as the clutter and we update $\mathcal{R}^1_{jk} \leftarrow 0$. The required number of RDMs to detect the moving objects with velocity v_i is formulated as

$$MT_r\frac{1}{D}\sum_i^D v_i \geq \Delta RES, \tag{14}$$

where $\Delta RES = \frac{c}{2B}$ is the range resolution. If we set $D = 16$, the moving objects with average speed $\frac{1}{D}\sum_i^D v_i = 0.11$ m/s that is much slower than human walking speed can be removed. Since the body motion compensation module is able to remove large body motions in the raw signals, the target subjects who are moving will not be regarded as dynamic clutters.

5.3 Near-Body Clutter Mitigation

After removing background clutter and dynamic clutter, the mmWave signal mainly includes vocal fold biometric information. However, the mmWave signal reflected by the near-body object is within the same range bin of the vocal vibration, which will still interfere with the phase estimation formulated in Eq. (5).

To mitigate the near-body clutter, we denote the composite amplitude and the initial phase of all the near-body clutter as A_0 and θ_0, respectively. Then, the phasor scatters on the complex phasor diagram can be formulated by the IF signal

$$H(mT_r + nT_s) = A_0\exp(j\theta_0) + \sum_{k=1}^K A^k_{m,n}\exp(j\theta^k_{m,n}), \tag{15}$$

where T_s is the sampling time interval, $A^k_{m,n}$ and $\theta^k_{m,n}$ represent the amplitude and the initial phase of the kth tone in the vocal vibration at $mT_r + nT_s$. Considering that the chirp cycle time T_r is far less than the duration of one phoneme, we can rewrite Eq. (15) as

$$H(mT_r + nT_s) = A_0\exp(j\theta_0) + \bar{A}\exp(j\theta_m). \tag{16}$$

We estimate the phasor amplitude $\widetilde{A}_m$ by maximizing the likelihood

$$\widetilde{A}_m = \left|\frac{1}{N}\sum_{n=0}^{N-1} H(mT_r + nT_s)\exp(-j\omega_H nT_s)\right|. \tag{17}$$

According to Eq. (16), the phasor scatter set $\mathcal{S} = \{\widetilde{A}_m\angle\psi_m\}$, $(m = 1, 2, \ldots, M)$ in the phasor diagram satisfies

$$\|\widetilde{A}_m\angle\psi_m - A_0\angle\theta_0\|_2 = \bar{A}. \tag{18}$$

Finding the composite amplitude A_0 and initial phase θ_0 of the clutter is equivalent to solving the following optimization problem:

$$\min_{A_0,\angle\psi_0,\bar{A}} \sum_{m=0}^M \{\|\widetilde{A}_m\angle\psi_m - A_0\angle\theta_0\|_2^2 - \bar{A}^2\}. \tag{19}$$

By denoting y as $[2A_0\cos\theta_0, 2A_0\sin\theta_0, \bar{A}^2 - oo^\top]$, the closed-form solution to the above minimization problem can be written as $y = (G^\top G)^{-1}G^\top d$, where $G^\top = [g_m]$ is a $3 \times M$ matrix with each column $g_m = [a_m\cos\psi_m, a_m\sin\psi_m, 1]^\top$, and $d = [\|\widetilde{A}_m\angle\psi_m\|_2^2]$, $m \in [1, M]$.

Finally, the near-body clutter mitigation is performed by removing the component $A_0e^{j\theta_0}$ from Eq. (15) and updating the ψ_m as

$$\psi_m = \arctan\frac{a_m\sin\psi_m - A_0\sin\theta_0}{a_m\cos\psi_m - A_0\cos\theta_0}. \tag{20}$$

Preliminary Results. In our preliminary experiment, we collect RF voice data that contains both static and dynamic clutters and verify our proposed model-centric clutter

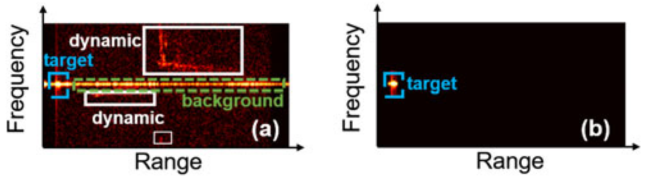


Fig. 7. The RDM before (a) and after (b) leveraging clutter suppression scheme.

suppression scheme. Two subjects are asked to sit towards mmWave probe in an uncontrolled outdoor environment with moving backgrounds (e.g., vehicles, and passersby), and pronounce "Ahhh" for around 5 seconds. Other experiment settings are the same as we mentioned in Section 2.2.

- *RDM analysis:* Fig. 7 shows the RDMs before and after leveraging clutter suppression scheme, and we observe that the dynamic and background clutters shown in Fig. 7a are all removed in Fig. 7b. Note that, since near-body clutter is within the same range bin of target vocal vibration signal, the mitigation effect cannot be observed from RDM. The results indicate that our proposed model-centric signal processing scheme is able to mitigate the impact of clutters in RF streaming and obtain precise vocal vibration data.
- *Spectral features analysis:* We further extract spectral centroid and crest from precise vocal vibration signals. The spectral centroid is the indication of the center of gravity of the spectrum, so it can locate large peaks corresponding to approximate formants' positions and pitch frequencies. The spectral crest represents the peakiness of the spectrum that can be used for quantifying the tonality of the signal. They are both typical spectral descriptors that can discriminate between different speakers. As shown in Fig. 8, we observe that both spectral centroid and crest possess a great difference between these two subjects in terms of local extremum, mean value, and variation trend. The results indicate that fine-grained spectral properties in vocal patterns are well preserved in RF voice data with the help of clutter suppression and can be used for identification.

6 Vocal Authentication

In this section, we explore and identify the optimal biometric features that characterize vocal source and vocal tract information (shown in Fig. 9) and input them to an ensemble classifier for robust user authentication.

6.1 Vocal Biometric Features Extraction

Vocal Source Features. The vocal source signal characterizes the muscle structure and tension of the vocal cords, and the related glottal pulse parameters, e.g., closing instants rate, opening duration, and opening degree of the glottis [23]. The vibration pattern of the vocal cords not only provides a voicing source for speech production but also characterizes unique nonlinear flow patterns. The resulting periodic pulse-like epoch shape varies among speakers. Therefore,

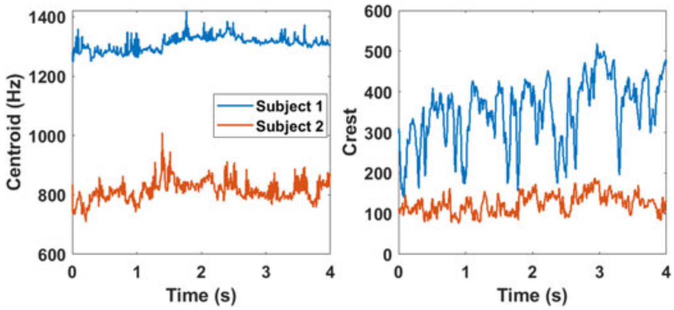


Fig. 8. Spectral centroid and crest that are extracted from two subjects' vocal vibration data after clutter suppression.

features derived from the vocal source provide unique physiological information for user identification.

To extract glottal flow cepstrum coefficients (GFCC) that represent the spectral magnitude characteristics of a speaker's glottal excitation pattern, we use the iterative adaptive inverse filtering (IAIF) method to estimate the glottal waveform of speech signal and then perform mel-spaced cepstral analysis [38]. We also derive residual phase Cepstrum coefficients (RPCC) [39] to characterize the phase information of the underlying excitation waveform. Moreover, to measure the underlying energy required for speech production, we compute the Teager phase cepstrum coefficients (TPCC) [40] that capture phase characteristics of the Teager nonlinear energy model of the speech production [41]. The process for the extraction is two-stepped. First, we apply the Teager-Kaiser energy operator to a band-pass filtered speech signal for calculating excitation energy contour and perform the Hilbert transformation to acquire a fine energy structure. Second, the cepstrum of the fine energy structure is computed and warped to the Mel frequency scale followed by a log and discrete cosine transform (DCT) operation to obtain TPCC.

Vocal Tract Features. Vocal tract system that consists of multiple articulatory organs (e.g., epiglottis, corniculate cartilage, cuneiform cartilage) works as a filter to resonate and reshape vocal source signals. The motion of relevant articu-

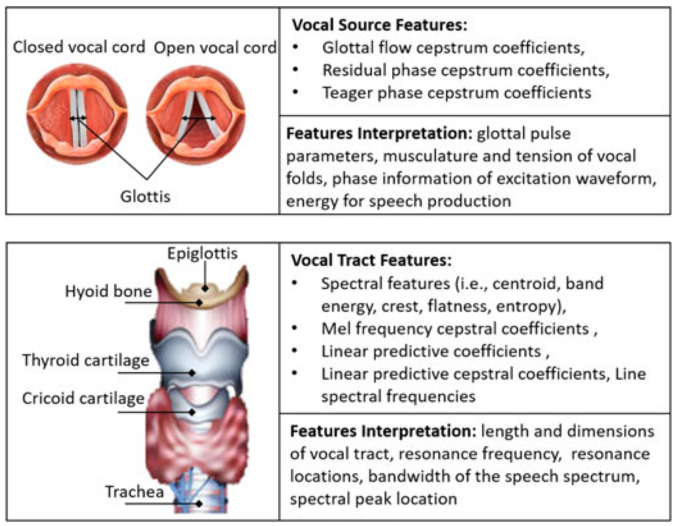


Fig. 9. The illustration of extracted biometric vocal features and corresponding interpretation.

latory organs generates associated articulatory gestures for the flow, but the movement speed and intensity vary from person to person. Therefore, we extract vocal tract features for speaker identification.

We first derive some spectral features, i.e., centroid, band energy, crest, flatness, entropy as the descriptors of the short-term spectral envelope [42], which are the acoustic correlate of timbre. Since coefficients on a linear/nonlinear Mel-scale of frequency can characterize the spectral envelope of a quasi-stationary signal segment, we further extract Mel frequency cepstral coefficients (MFCC) [43] to reflect the resonance properties of the vocal tract system. Specifically, we first convert pre-processed RF vocal biometric signals into a set of mel-frequency spectrums and then employ Triangular band-pass filters to make the signals adhere to the attenuation characteristics of the Mel scale. After the logarithmic compression and DCT, 12-dimensional MFCCs is acquired. To complement MFCCs, linear predictive coefficients (LPC) [44] is selected to characterize formants, i.e., a resonance frequency of the vocal tract. We adopt linear prediction methods to infer the filter coefficients equivalent to the vocal tract by minimizing the mean square error between the input vocal signals and estimated vocal signals. Based on the extracted LPC, we deduce linear predictive cepstral coefficients (LPCC) [45] by performing Cepstral analysis on LPC calculated spectral envelope. We also derive line spectral frequencies (LSF) [46] from LPC, since it can characterize bandwidths and resonance locations and emphasize the spectral peak location.

6.2 Fine-Tuned Authentication

Biometric Feature Selection. The majority of deep learning-based feature extraction/selection methods consume large computation resources and lack interpretability, which is not suitable for our exploratory study to derive biometric traits from RF voice signals. Therefore, we adopt the Fisher Score-based feature selection method [47] since it selects vocal features that are more distinct between different speakers and consistent within one speaker using small computation effort. However, it fails to select some features that have relatively low individual scores but a very high score if they are combined with others as a whole. To overcome this shortage, we employ cutting plane algorithm [48] to select a subset of features simultaneously that maximize the lower bound of conventional Fisher score. In each iteration, multivariate ridge regression and projected gradient descent are adopted alternatively to solve a multiple kernel learning problem [49]. After the feature selection, the initial vocal biometric feature vector is reduced to 39 descriptors and then fed to the classification model.

Classifiers Fusion. As the first exploratory study to derive vocal biometric traits from skin-reflected mmWave, we employ the following widely used speaker identification classifiers:

- *Gaussian Mixture Model-Universal Background Model (GMM-UBM):* The use of GMM for modeling speaker identity is because the Gaussian components approximate spectral features and Gaussian mixtures can model arbitrary densities [50]. To guarantee reliable system performance without increasing model complexity, we further introduce UBM to help develop the speaker identification model [51]. The UBM model is trained with expectation-maximization (EM) algorithm on a large amount of data gathered from the background population (i.e., the NIST 2001 one-speaker detection database [52]), then the target speaker model is adapted from the UBM model utilizing training data based on maximum a posteriori (MAP) principle [51]. We calculate the difference of log-likelihood between the target speaker model and the UBM model to determine whether selected features are originated from the genuine speaker or not.
- *Support Vector Machine (SVM):* It is a classification and regression method based on statistical learning theory [53]. We adopt SVM in speaker identification because it can achieve superior generalization performance in classifying unseen data [54]. With the help of kernel functions, the SVM optimizer can find a maximum-margin hyperplane that separates training samples from the genuine speaker and impostor subjects.
- *Hidden Markov Model (HMM)*: It is a statistical tool that describes a Markov process with unobserved states. We select HMM for speaker identification because the states of an HMM characterize the vocal configuration of a speaker and the changes of vocal configuration may duplicate in pronunciation [55]. We use the Baum-Welch algorithm [56] to determine the parameters of an HMM. Then, speaker identification is performed by a Viterbi algorithm to compute likelihood scores for each signal [57].

Finally, we combine the output scores of these three classifiers by weighted sum and optimize the fusion weights based on logistic regression. The BOSARIS Toolkit [58] is employed for implementing fusion and determining the genuine speaker.

7 System Implementation and Evaluation Setup

7.1 System Implementation

The selection criteria for mmWave probe hardware depends upon the desired waveform characteristics that take two major factors into consideration. First, the chirps and frames generated by the probe should be able to capture the high-resolution vocal vibration from the target user. Second, the mmWave probe should guarantee the minimum signal-to-noise ratio (SNR) so that the proposed data processing techniques can distinguish among vocal vibrations and the clutter. Therefore, we carefully design the mmWave waveform configuration as shown in Table 1. This configuration enables the range resolution of 3.75 cm, displacement resolution around 1 mm [18], which satisfies the requirements for sensing the vocal vibrations. Our design of the mmWave waveform can be generated effortlessly by any off-the-shelf mmWave probe [59], [60] or customized hardware [61], thereby facilitating affordability (less than $70), portability (100g) and energy-efficiency (135 mW) in real-world setups. In our work, we leverage a Texas Instruments AWR1642 mmWave probe (TX Power = 12.5 dBm, RX Gain = 30 dB)

TABLE 1
mmWave Waveform Design

Frequency Slope	71.5 MHz/μs	Bandwidth	4 GHz
ADC Samples/Second	5000K	Idle Time	10 μs
Chirp Cycle Time	65.8 μs	Chirps/Frame	128
Frame Periodicity	9 ms	Samples/Chirp	256

[62] to emit the signal and capture the data. The range profiles are generated on-board and then transferred to the laptop for further processing.

7.2 Evaluation Setup

Experiment Preparation. We conduct extensive experiments to confirm the capability of *VocalPrint* for user authentication. Fig. 10 shows the experimental setup used for evaluation. A subject is asked to sit in a chair in a relaxed position. We position the customized mmWave probe in front of the subject and align it in the direction of subject's throat, i.e., the subject orientation is 0° with respect to the probe. The distance between subject and the probe is 20 cm. In each authentication trial, all subjects are required to sit in the same position, unless specified in the evaluation. The mmWave probe connects to a 5.0 V power supply, and the working current is 2A. We deploy two laptops with the Windows 10 operating system. One laptop is used for collecting the signals from the receiving terminal of a mmWave probe, using the network interface card. The other laptop is employed to display the reading materials for the subjects. The training processes are performed by a workstation equipped with an Intel Xeon CPU E5-1620 v4 @ 3.50 GHz.

Data Collection. Our biometric study is approved by IRB. 41 subjects (21 males and 20 females) are asked to read *The North wind and the sun passage* (113 words) and the first two sentences of *The Grandfather Passage* (37 words) following a prompter to guarantee the same reading time. On average, each subject takes around 51 seconds for *The North wind and the sun* and 14.6 seconds for the first two sentences of *The Grandfather Passage*. The collected data are anonymous and stored locally to protect the subject's privacy.

Partition. To evaluate the performance with text-independent features, we use the received signals corresponding to *The north wind and the sun* for training and *The Grandfather Passage* for testing. The received signals are segmented evenly with a 50 percent overlapping rate and then filtered by an efficient speech detection mechanism based on the Zero Cross Rate and Root Mean Square in time domain [63] to isolate non-speech segments. The segment lengths are varied as 5 ms, 10 ms, 15 ms, 20 ms, 25 ms, and 30 ms, respectively, for performance analysis. Based on segment length, we finally collect 111720–672000 samples and use 71400–428400 samples for training and the rest for testing. Among the overall 41 subjects, each acts as a genuine user once while the remaining 40 subjects act as imposters to access the system. Therefore, the genuine subjects and imposters ratio is 1:40 for every authentication trial.

Evaluation Metrics. We introduce F-score, balanced accuracy (BAC), receiver operating characteristics (ROC) curve, equal error rate (EER) as metrics in our evaluation since

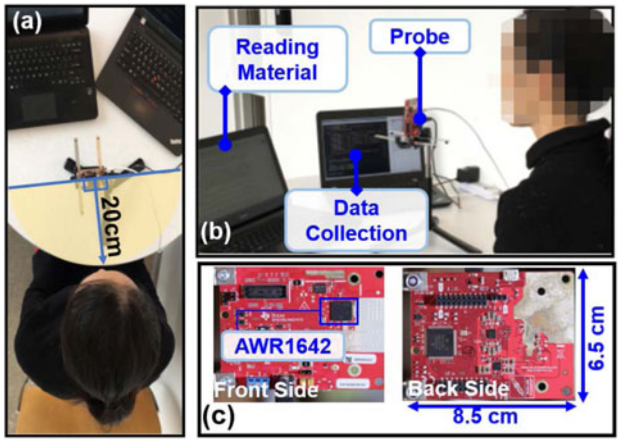


Fig. 10. The evaluation setup: (a) subject's sitting position; (b) in a lab environment; (c) mmWave probe architecture.

these are non-sensitive to class distribution for evaluating authentication systems [32], [64].

8 Performance Evaluation

In this section, we evaluate the performance and robustness of *VocalPrint* for authentication. All the results are obtained after the body motion compensation except the one specified as "before body motion compensation" in the evaluation of subjects in motion.

8.1 Overall System Performance

We first examine the effectiveness of our proposed ensemble classifiers. Specifically, We train three classification models (i.e., SVM, HMM, GMM-UBM) and compare their authentication performance with the ensemble classifiers. The segment length for collected data is set as 100 ms. F-score and balanced accuracy (BAC) are selected as performance metrics in our evaluation. Fig. 11 shows that the ensemble classifiers can achieve up to 98.9 percent BAC, and 96.8 percent F-score. By contrast, the F-score of SVM, HMM, GMM-UBM are all less then 95 percent. In conclusion, the ensemble classifiers employed in our system can achieve superior authentication performance.

To maximize the applicability of *VocalPrint* in real-world scenarios, it is important that the system can not only differentiate between the legitimate users and imposters but also perform the authentication in a timely fashion. The authentication time is defined as the total time elapsed to make a final prediction and is dependent on the segment length

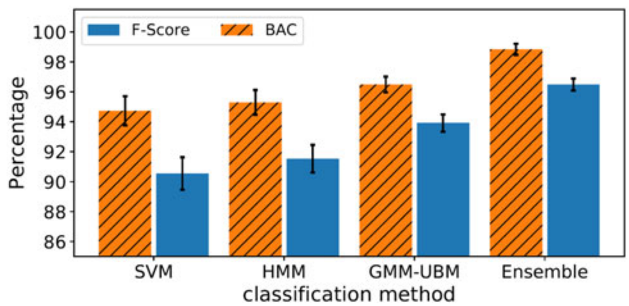


Fig. 11. The overall performance of *VocalPrint* with different classification models.

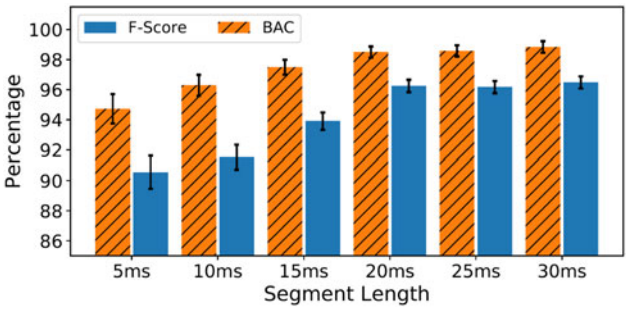


Fig. 12. The overall performance of *VocalPrint* with different segment lengths.

needed to authenticate users. To determine an optimal segment length of mmWave signal for precise authentication, we evaluate the system performance with segment lengths as 5 ms, 10 ms, 15 ms, 20 ms, 25 ms, 30 ms, respectively.

Fig. 12 illustrates the F-score and BAC measure for 41 subjects with different segment lengths. We observe that when the segment length is less than 15 ms, it does not contain sufficient information for accurate authentication, indicated by the low BAC and F-score, and high standard deviation (STD). The performance is improved when the length of the segment is increased, however, the improvement in F-score and BAC is not significant after the segment length is increased from 20 ms to 30 ms. Specifically, BAC achieves 98.52, 98.58, and 98.85 percent with the STD of 0.37, 0.37 and 0.38 percent for 20 ms, 25 ms and 30 ms, respectively. F-score reaches 96.27, 96.18, and 96.46 percent with the STD of 0.41, 0.4 and 0.4 percent for 20 ms, 25 ms and 30 ms.

For a more concrete analysis, we also plot the ROC curves and calculate the corresponding area-under-curve (AUC) with different segment lengths, as shown in Fig. 13. Although the 30ms segment achieves the best performance, the performance is not improved significantly compared with the 20 ms segment. The corresponding EER are given as 9.91, 10.18, 9.08, 4.97, 4.99, and 4.92 percent, respectively. These results are consistent with BAC and F-score.

Based on the above observations, we conclude that the segment length of 20 ms is most appropriate for training and testing. With such a segment length, the total time needed to verify a user is 340 ms. The total authentication time is stable even if the environment introduces more ambient clutters and noise. This is because we only search on the range-Doppler-matrix once to locate the background clutters and dynamic clutters no matter the amounts of clutters. Moreover, we establish a mathematics model and solve a circle fitting optimization problem to find the composite amplitude and initial phase of the near-body clutters, which means the time complexity is stable. The results also demonstrate the effectiveness of *VocalPrint* for reliable user authentication. For the remainder of this paper, we use the segment length of 20 ms during the performance analysis.

8.2 Robustness and Usability Analysis

3D Orientations. To maximize the user experience, *VocalPrint* should require minimum user cooperation and be tolerant to unfixed sensing orientations. Therefore, it is necessary to investigate whether the performance of

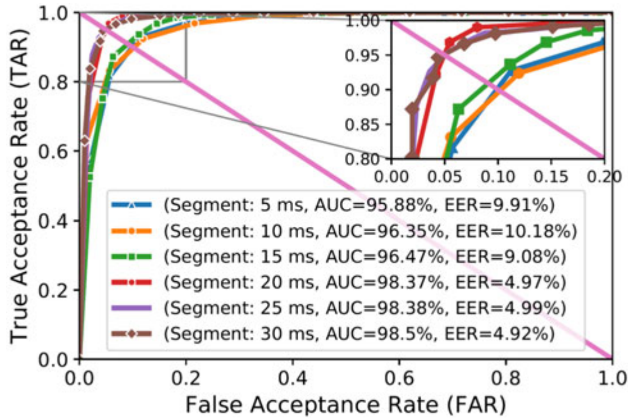


Fig. 13. The ROC and EER with different segment lengths.

VocalPrint will be affected by changeable 3D orientations. In the experiment, we collect users' RF voice data from different angles of the human throat with respect to the probe. Specifically, the azimuth is varied from 0° to 60°, and altitude is varied from -20° to $+20^\circ$. The experiment results are shown in Fig. 14. We observe that the BAC reaches up to 96 percent when azimuth is within 45° and altitude is within 20°. This may be because the vibrations from neck surfaces (besides the throat region) also contain vocal biometric information. Moreover, although the near-throat vibrations such as mouth motion and heart beat are within the same range bin of the vocal vibration, they almost cannot affect the system performance due to the near-body clutter mitigation module.

Variant Distances. Due to mmWave attenuation, the *VocalPrint* performance drops as the sensing distance increases. Therefore, we want to further explore *VocalPrint* can work in what kind of application scenarios and at which level of authentication accuracy, when extending the distance. Specifically, we evaluate *VocalPrint* performance in subdivided daily-life scenarios: 1) body field (0-0.5 m): communication with smartphone and wearable device; 2) social distancing field (0.5 m-2 m): interaction with car and desktop device; 3) local field (2 m-5 m): interaction with the smart home appliance. Fig. 15 shows that *VocalPrint* can achieve over 98 percent BAC in the body field and over 95 percent BAC in the social distancing field. As the distance increases to 460 cm, the authentication accuracy is around 91.7 percent.

Body Posture and Motion. To enhance usability, *VocalPrint* should facilitate accurate user authentication at all times, without requiring users to stop their ongoing activities (e.g., driving) or put down any object held in their other hand. As

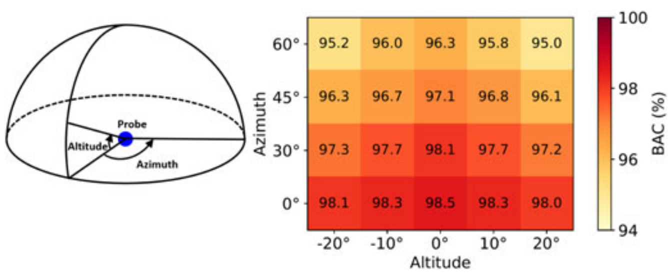


Fig. 14. The performance of *VocalPrint* with changeable 3D orientations.

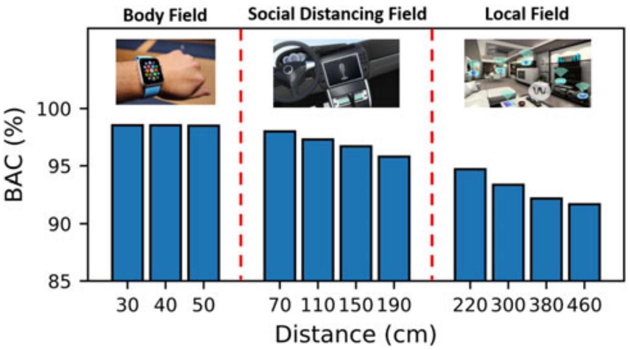


Fig. 15. The performance of *VocalPrint* in the body field, social distancing field, and local field.

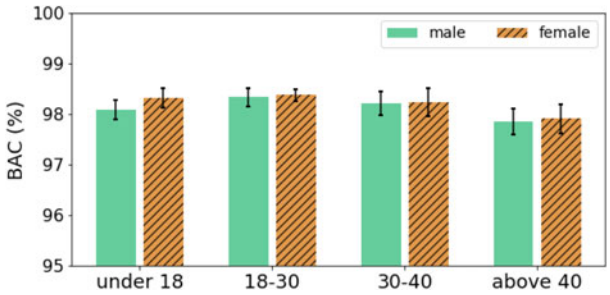


Fig. 17. The performance of *VocalPrint* with different users' demographics.

a result, we investigate the performance of *VocalPrint* under the effect of body posture and motion. Specifically, we study (1) posture and motion that may shelter the near-throat skin surface from mmWave sensing; (2) periodic body posture and motions. In our experiment, while reading the first two sentences of *The Grandfather Passage*, each subject is asked to continuously perform four common daily-life activities, including rhythmic movements during listening to music, combing hair, mimic driving, and writing, thereby exhibiting minute to large-scale body motion. Note that we distribute reading materials printed on a paper to the subjects so that they can write on them while reading simultaneously in the experiment. With the current experimental setting, the BAC results before and after body motion compensation are shown in Fig. 16. With the body motion compensation model, we observe that the BAC corresponding to the subject performing a rhythmical movement and combing hair during authentication reaches above 98 percent, while the BAC for sheltering motions (i.e., writing and mimic driving) varies between 96 and 98 percent. The results demonstrate that body motion resulting in sheltering of the near-throat skin surface from mmWave sensing can affect system performance to some extent. This is due to the limited penetration capability of 77 GHz mmWave that is leveraged in this work [65]. Meanwhile, without the body motion compensation method proposed in our work, the BAC gets reduced to an average performance of 73 percent. Regardless of body motion, *VocalPrint* shows a reliable performance in user authentication.

Users' Demographics. To explore whether or not vocal-based authentication will be affected by users' demographics, we evaluate *VocalPrint* performance under different user background, i.e., age and gender. We recruit both male and female subjects from four age groups, i.e., under 18, 18-30, 30-40, above 40. They are arranged to read the first two sentences of *The Grandfather Passage* with the same experiment settings. As observed in Fig. 17, the BAC of different groups are around 98 percent, and the authentication performance for users above 40 years old drops slightly. This is because our system is trained based on a limited number of young and educated subjects and make our model lack of generality to some extent. We plan to invite more subjects with diverse background to participate in our study. Overall, *VocalPrint* performance is robust against different age groups and gender groups.

Speech Language. Since *VocalPrint* is a text-independent system, we are curious about whether our system is robustness to the language of utterance. Subjects are asked to enroll with English sentences and test with Chinese sentences, and vice versa. We also test the speaking speed in either condition. The subjects are requested to read the testing utterance in a slow, normal, and fast speed, respectively. The speed is controlled by the subjects themselves. Fig. 18 shows that the average BAC values are between 98 and 98.6 percent in each language condition no matter the speaking speed varies. The results indicate that *VocalPrint* can be used in multilingual applications because we extract intrinsic vocal source and tract properties of users for training the model, rather than the idiolect and conversation-level characteristics.

Wearable Accessories. In our daily life, it is common for users to wear accessories around the throat. Therefore, we are motivated to examine whether the authentication performance will be affected by the wearable accessories which are made of different materials (e.g., metal, cotton, wool, plastic) and pose partial or full occlusion to the throat. Specifically, the subjects are asked to wear necklaces, shirt collar, scarf, and earbuds, respectively while reading the first

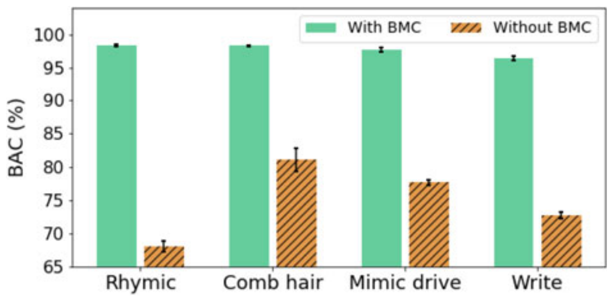


Fig. 16. The BAC comparison with and without body motion compensation.

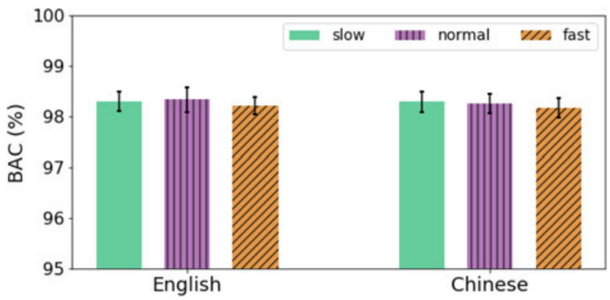


Fig. 18. The performance of *VocalPrint* with different languages.

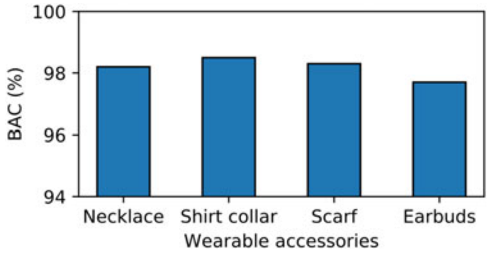


Fig. 19. The performance of *VocalPrint* with different wearable accessories.

two sentences of *The Grandfather Passage*. Fig. 19 shows that *VocalPrint* achieves more than 98 percent BAC with necklace, shirt collar, and wool scarf around neck, and 97.7 percent BAC with earbuds. Therefore, *VocalPrint* is robust to wearable accessories.

Longitudinal Study. For any biometric method, permanence is a critical factor. We examine the permanence of vocal vibrations to show the potential of *VocalPrint* as an enhancement to voice authentication. 20 subjects (10 males and 10 females) participate in the long-term study lasting 30 days. In every period of three days, each subject reads the first two sentences of *The Grandfather Passage*, and mmWave signals are obtained. The training set is generated based on the collected mmWave signals on the first day of enrollment. As Fig. 20 shown, the average values of BAC are between 98 and 99 percent, and the STDs are between 0.37 and 0.39 in the 30-day duration. We can conclude that there is no notable decreasing and ascending tendency on average BAC results, which indicates that *VocalPrint* is robust to the time change.

9 Ambient Resilience Study

Many voice authentication systems suffer from notable attenuation in performance due to complex ambient noise. In this section, we investigate the impact of ambient factors on *VocalPrint* performance.

9.1 Acoustic Noise

In practice, there are two primary types of acoustic noise with different spectral characteristics: pop music and presentation. To evaluate the authentication performance in presence of acoustic noise, a loudspeaker is placed next to the user and plays the recorded music and presentation sound at different decibel levels (i.e., volume varies as 0, 25, 50 and, 75 percent). At the same time, each subject reads the first two sentences of *The Grandfather Passage*. With the current experiment setting, we obtain the BAC and F-score under different volumes of music and presentation sound, as shown in Fig. 21. We observe that the authentication accuracy does not exhibit much difference under music and

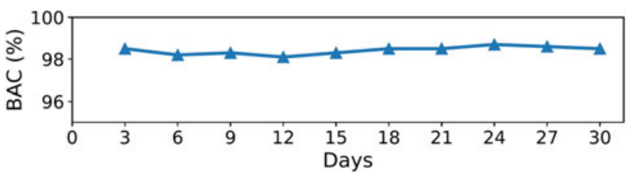


Fig. 20. A 30-day longitudinal study.

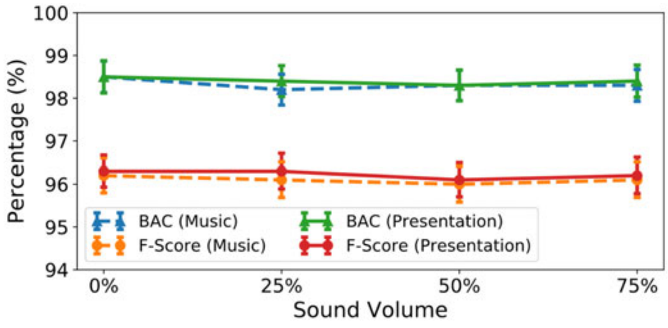


Fig. 21. Performance under acoustic noise.

presentation sound. Even when the loudspeaker volume increases from 0 to 75 percent, the values of BAC and F-score remain stable. To be specific, the BAC values are between 98 and 99 percent, and the F-score values are between 96 and 97 percent. These results validate that *VocalPrint* is immune from different types and volumes of acoustic noises. This is consistent with the fact that *VocalPrint* employs an electromagnetic channel that is not affected by acoustic noise in the ambient environment.

9.2 Environmental Dynamics

It is a known fact that variations in the sensing environment can significantly affect the quality of the received signal, increasing the false acceptance rate of the authentication system. Specifically, mmWave signals may be affected by the stationary and moving objects in the environment. To evaluate the capability of our proposed resilience-aware suppression model, we select three ambient conditions: (1) snowy outdoor with environmental temperature as $-5°C$ ($23°F$) and no human obstruction; (2) student lounge with environmental temperature as $20°C$ ($68°F$) and periodic human obstruction; (3) three people continuously walking around the mmWave probe within 2m distance, depicting constant human obstruction. The subjects are asked to read the first two sentences of *The Grandfather Passage* in different ambient conditions as mentioned above. Fig. 22 shows the authentication results. In each ambient condition, the BAC reaches over 98 with around 0.4 percent STD, and the F-score values are more than 96 percent with approximately 0.4 percent STD. These results indicate that the resilience-aware clutter suppression approach can effectively remove the clutters in the usual authentication scenarios. Therefore, *VocalPrint* is resilient against environmental dynamics and can be applied in real-world scenarios.

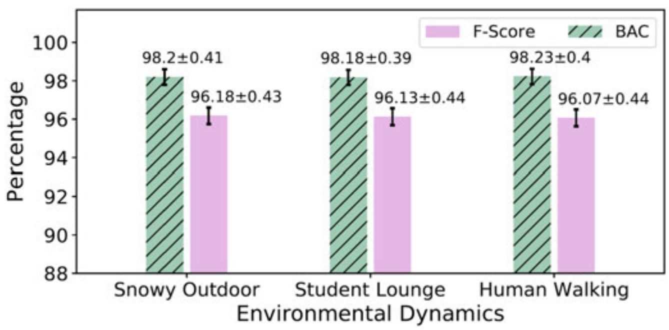


Fig. 22. Performance under environmental dynamics.

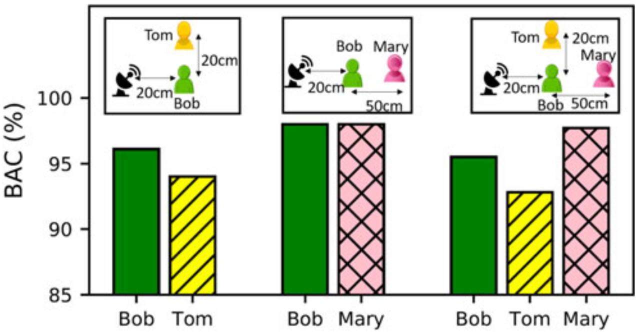


Fig. 23. The performance of *VocalPrint* with multiple speakers.

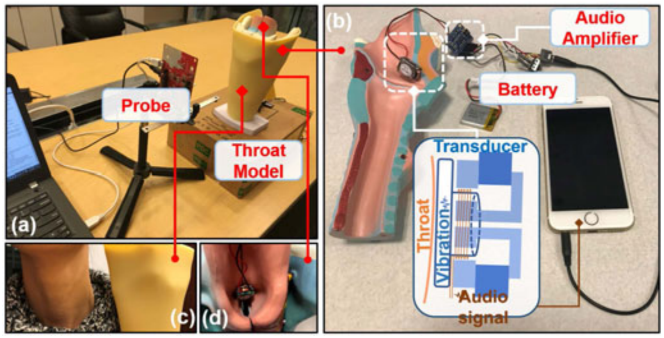


Fig. 24. Counterfeit attack experiment setup. The adversary counterfeits vocal vibrations using an audio transducer and 1:1 throat model.

9.3 Multiple Human Speakers

In real practice, multiple human speakers may present in the background of the legitimate user and their vocal vibration data may disturb our desired target data. Therefore, we are curious about whether *VocalPrint* can identify legitimate user among multiple human speakers. we recruit 3 subjects (namely, Bob, Tom, Mary) and arrange them into three groups (e.g., Bob-Tom, Bob-Mary, Bob-Tom-Mary). Specifically, the orientation and distance between Bob's throat and mmWave probe are 0° and 20 cm as mentioned in the experiment setup. Tom sits rightwards Bob with a distance of 20 cm and his orientation is 45° with respect to the probe. Mary sits behind Bob at a distance of 50 cm. In order to avoid blockage by Bob, Mary's orientation is around 10° with respect to the probe. Subjects in the same group are requested to pronounce the first two sentences of *The Grandfather Passage* simultaneously. The results of identifying each speaker in these three groups are illustrated in Fig. 23. We observe that the BAC can reach up to 98 percent in Bob-Mary group where the distance between the subjects is 50 cm. The performance drops significantly if the subjects are too close, e.g., within 20 cm in Bob-Tom group. when extending the number of subjects in a group, the subject who is 50 cm away from other subjects can still achieve around 98 percent BAC. Therefore, we can conclude that *VocalPrint* is resilient against multiple human speakers in common social distancing field (i.e., more than 50 cm) due to current waveform design.

10 Spoofing Resilience Study

10.1 Counterfeit Attack

We assume that the attacker (1) knows that the uniqueness of human voice is intrinsically sourced in vocal vibration; (2) observes that when a person speaks, vocal cord vibrations caused by air pressure are propagated through the vocal tract and can be measured on the skin surface. Based on this knowledge, the attacker may forge the target's vocal vibration to spoof *VocalPrint*. To verify if a human's vocal vibration can be simulated, we construct a counterfeit attack model, as shown in Fig. 24a. We place an audio transducer inside a throat model to replay a pre-recorded passphrase of the target user. As illustrated in Fig. 24b, the transducer is used to simulate the vocal source excitation signal (i.e., vocal cord vibrations caused by air pressure). When the audio signal passes the coil of the transducer, a dynamic electro-magnetic field is generated, which makes the actuator vibrate the throat model. The supralaryngeal vocal tract in our throat model acts as reshaping the source signal, as shown in Fig. 24d. Finally, the forged vocal vibration is reflected by a readily available bionic skin material (i.e., Silicone [66]) covering the throat model (see Fig. 24c).

To overcome this counterfeit attack, we implement a body motion detector in the random motion compensation module (see Section 4.3) to judge whether the reflected vocal vibration is originated from a live user or a model. Specifically, the detector examines the value of range shift χ_m in Eq. (8) when the envelop correlation function reaches maximum. If $\chi_m = 0$, it implies that the range bins misalignment issue does not exist, i.e., there is no random body motion. To evaluate the effectiveness, we place the forged throat model at a distance of 20 cm to *VocalPrint*, and try 200 trials. The results show that merely 1.5 percent forged vocal vibration data match with the legitimate recording.

10.2 Vibration Replay Attack

We also consider a challenging scenario where the audio transducer (i.e., vibration part) is stuck to the actual human throat. We launch 200 attacks, only 9 (4.5 percent) forged models are misclassified as legitimate users. This is because the non-linear response [67] of audio transducer may affect the vibration fingerprint in RF streaming signals, and thereby cannot match with biometric vocal pattern registered in the database. Non-linear response effect shows that the circuit inside an e-device works as a passive signal modulator that can reflect back RF signals with inherent identity information, and is well-studied for hidden electronics recognition. To summarize, experiment results indicate that *VocalPrint* can combat this counterfeit attack.

10.3 Mimicry Attack

Some attackers may intend to compromise the *VocalPrint* by mimicking the speaker. To verify whether *VocalPrint* can defend against mimicry attack, 10 volunteers are invited to mimic the target speaker. These volunteers are face-to-face with the target speaker and observe how they pronounce speech. After the volunteers repeatedly practice pronunciation by mimicking 1) the articulatory movement of upper and lower lips, tongue and jaw; 2) speaking speed, intonation, rhythm, conversation-level characteristics (e.g., "uh-huh", "oh yeah", etc.) of the speaker, they initiate the mimicry attacks in front of the *VocalPrint*. Each volunteer mimics 5 target speakers for 10 trials. In all, 500 mimicry attacks are launched to *VocalPrint* but every attack fails. The results are

TABLE 2
A Comparison of Voice-Based Authentication Methods

System	Liveness Detection Principle	Sensing Mechanism	Sensing Orientation	Acoustic Noise Sensitivity	User Cooperation	Test Distance
VAuth [2]	Body vibration	Bone conduction	N/A	Resistive	Yes	Contact
Chen *et al.* [68]	Magnetic field	Magnetic	Directional	Resistive	Yes	<10 cm
CaField [16]	Sound field	Acoustic	Directional	Sensitive	Yes	<50 cm
VoicePop [69]	Pop noise	Acoustic	Non-directional	Sensitive	Yes	<10 cm
VoiceLive [15]	TDoA of phenome sounds	Acoustic	Non-directional	Sensitive	Yes	<50 cm
VoiceGesture [14]	Ultrasonic reflection of mouth motion	Acoustic	Non-directional	Sensitive	Yes	<50 cm
WiVo [70]	RF reflection of mouth motion	Radio frequency	Directional	Resistive	No	<50 cm
VocalPrint	RF reflection of vocal vibration	**Radio frequency**	**Directional**	**Resistive**	**No**	**0-200 cm**

as expected because the human voice is individually unique and cannot be entirely mimicked.

10.4 Signal Replay Attack

We assume that the attacker knows the details of the mmWave probe that we use to sense the skin-reflected signals. Understanding this fact, the attacker can first eavesdrop the communication between the target speaker and the mmWave probe in *VocalPrint* so as to record the skin-reflected signals from the speaker. After implementing the eavesdropping process, in a new phase, the attacker can deploy devices to absorb mmWave signal transmitted by *VocalPrint*, and then replay the pre-recorded skin-reflected signals to spoof *VocalPrint*. To defend this signal replay attack, during $(t+i)$ phase, *VocalPrint* randomly selects three chirps as r^{t+i}-*th*, s^{t+i}-*th*, q^{t+i}-*th* in the emitted signal with a randomized chirp rate ρ, denoted as ρ_r^{t+i}, ρ_s^{t+i}, ρ_q^{t+i}, respectively. By altering the chirp rate of selected chirp in the emitted signal, the frequency shift value τ of range profile varies after performing FFT operation on IF signal, denoted as τ_r^{t+i}, τ_s^{t+i}, τ_q^{t+i} in *VocalPrint*, respectively. During $(t+i)$ phase, when an attacker replays signal pre-recorded in t phase, such signal can be easily refused by comparing the parameters in $(t+i)$ phase recorded by *VocalPrint* with the parameters in t phase recorded by the attacker. To examine the effectiveness, we record the reflected signals from different target speakers and use mmWave signal generator to send the imitation signals to *VocalPrint*. The experiment results show that *VocalPrint* can recognize all the imitation signals.

11 Related Work

Voice Authentication. Voice authentication is a historical topic in biometrics and has been studied well [70], [71]. Existing studies show that most voice authentication solutions are vulnerable against spoofing attacks [72], [73]. To defense attacks, many liveness detection approaches have been proposed based on the distinction between human and loudspeaker [74], [75]. For example, mouth motion in speaking is distinct from the manner that the loudspeaker vibrates the diaphragm. Based on it, some studies leveraged RF reflections [70] and ultrasonic reflections [14] of the mouth motion for liveness detection, but mouth motion is observable and can be mimicked potentially. Other studies exploited characteristics exclusive to human speakers or loudspeakers, such as magnetic field emitted from the loudspeaker [68], pop music in human utterances [69], time-difference-of-arrival from two microphones [15], and sound field [16]. However, these microphone-based solutions are intrinsically sensitive to ambient noise and also assume that replaying cannot generate identical sound waves. By introducing a non-sound based sensing modality, *VocalPrint* can protect the system even if attackers generate identical sound waves. Other solutions leveraged contact-based sensors for voice authentication, such as VAuth [2], Vocal Resonance [63] which can immune ambient noise. However, these bone-conduction solutions require skin-contact and sacrifice usability. According to the literature (see Table 2), no solution exists in addressing all these issues in voice authentication.

mmWave-Based Human Sensing. Recent advances have demonstrated that mmWave accurately detects minute variations caused by a human without body contact [76], [77]. Some works leverage mmWave into activity recognition [78], [79], [80] and emotion recognition [81]. Some other works focus on the detection of biometrics [81], [82]. For example, Petkie *et al.* [83] employed a 228 GHz heterodyne radar to measure the respiration and heart rates at a distance of 10 meters. Lin and Song *et al.* [32] implemented Cardiac Scan, a non-contact and continuous sensing system for user authentication. Recent workS [19], [28] leverage mmWave to sense voice-related information to facilitate voice-user interface. Compared with these works, our work explores vocal vibration as a continuous and non-contact biometric identification and captures anti-spoofing features (i.e., throat physiological intrinsic) to defend against malicious attacks.

12 Dicussion

Long Sensing Distance. As the distance increases, velocity resolution of the range profile will correspondingly degrade, thereby the performance of background clutter isolation will be affected and finally lead to decreased authentication accuracy. To extend effective sensing distance for remote voice biometric-based application, we can increase the bandwidth of IF signals in mmWave waveform design [84] and enhance the signal to noise ratio (SNR) of mmWave signal. For example, Chen *et al.* [85] developed an algorithm combining empirical mode decomposition and mutual information entropy to reduce noise component.

Sensing Orientation. VocalPrint's performance is affected by the user orientation especially in long-distance scenarios. Moreover, when the user orientation (i.e., Azimuth) extends to 90°, the system performance drops significantly, with less than 70 percent BAC. To enhance the robustness of *VocalPrint*, one possible solution is to collect the user's vocal vibration from different orientations with respect to the probe in the enrollment phase, because vocal sounds travel through the bone and the resulting vibration on the neck surface (besides the throat region) is also valuable for authentication [63]. Specifically, we plan to collect 5,000 samples for each orientation, i.e., 60°, 90°, 120°, 150°, to train the model. This is similar to the enrollment phase of FaceID which requires the user to move his/her head slowly to complete a circle [86].

Compatibility With IoT Devices. 5G smartphone has been introduced in the market, and the smart IoT devices that equip 5G baseband will be pervasive in the future [87]. In the hardware aspect, the new phased array antenna and beamforming module in the architecture of mmWave-capable smart device can support continuous interrogation of user's vocal vibration. In the software aspect, signal processing based resilience-aware clutter suppression and vocal authentication can be implemented with the help of the high-speed DSP in the smart device [88]. The potential power consumption is around $2{\sim}3$ w when applying the *VocalPrint* in 5G smartphone [89].

mmWave-Voice Side Channel. mmWave-voice can be a new side channel that is possible to be hacked by malicious applications or attackers. Specifically, once the malicious application takes control of the mmWave module of the smart device, it will search the victim's voice in the nearby and extract sensitive audio information. To solve this issue, advanced permission control function should be introduced in the applications for accessing mmWave base band.

Voice-Related Diseases. The symptoms of cold and flu may cause changes in glottal excitation and vocal tract damping. Therefore, our authentication system based on vocal source and vocal tract features are not robust to verify the user who has a sore throat or catches a cold. In future work, we plan to recruit some participants (requested to wear masks) with symptoms of cold and flu in the enrollment phase for training a model that can tolerate changes within subjects.

13 Conclusion

Existing voice authentication systems are vulnerable to noise interference and spoof attacks. In this paper, we introduce a novel biometric system, *VocalPrint*, for resilient security of voice authentication. Specifically, *VocalPrint* is on the basis of a 77 GHz FMCW probe to sense the minute vocal vibrations in near-throat region of users and leverage the skin-reflect mmWave signals. A novel resilience-aware clutter suppression approach is proposed to isolate the complex ambient noise and body motion from the mmWave signals and allow further extraction of unique vocal tract and vocal source features. Extensive experiments indicate that the authentication accuracy of *VocalPrint* exceeds 96 percent even under unfavorable conditions. We also show the ambient resilience and spoof resilience of *VocalPrint* to show its practicality in real-world setups. In future work, we plan to evaluate *VocalPrint* with more people suffering from speech disorders and improve system accuracy.

Acknowledgments

This work was supported by the National Science Foundation under Grants ECCS-2028872 and CNS-1718375.

References

[1] J. A. Markowitz, "Voice biometrics," *Commun. ACM*, vol. 43, no. 9, pp. 66–73, 2000.

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

[3] M. Abuhamad, A. Abusnaina, D. Nyang, and D. Mohaisen, "Sensor-based continuous authentication of smartphones' users using behavioral biometrics: A contemporary survey," *IEEE Internet Things J.*, vol. 8, no. 1, pp. 65–84, Jan. 2020.

[4] L. G. Kersta, "Voiceprint identification," *Nature*, vol. 196, no. 4861, pp. 1253–1257, 1962.

[5] J. P. Campbell, "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, Sep. 1997.

[6] J. Silovský and J. Nouza, "Speech, speaker and speaker's gender identification in automatically processed broadcast stream," *Radioengineering*, vol. 15, pp. 17366–17373, 2006.

[7] S. Larson, "Google home now recognizes your individual voice," *CNN Money, San Francisco, CA, USA*, 2017. [Online]. Available: https://money.cnn.com/2017/04/20/technology/google-home-voice-recognition/index.html

[8] C. S. Smith, "Alexa and siri can hear this hidden command. You can't," 2018. [Online]. Available: http://www.nytimes.com/2018/05/10/technology/alexa-siri-hidden-command- audio-attacks.html

[9] X. Yuan *et al.*, "Commandersong: A systematic approach for practical adversarial voice recognition," in *Proc. 27th USENIX Conf. Secur. Symp*, 2018, pp. 49–64.

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

[11] Z. Ba, S. Piao, X. Fu, D. Koutsonikolas, A. Mohaisen, and K. Ren, "ABC: Enabling smartphone authentication with built-in camera," in *Proc. 25th Annu. Netw. Distrib. Syst. Secur. Symp.*, 2018, pp. 1–15.

[12] C. Jung, J. Kang, A. Mohaisen, and D. Nyang, "Digitalseal: A transaction authentication tool for online and offline transactions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*, 2018, pp. 6956–6960.

[13] R. M. Bolle, J. H. Connell, and N. K. Ratha, "System and method for liveness authentication using an augmented challenge/response scheme," U.S. Patent 6,851,051, Feb. 1, 2005.

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

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

[16] C. Yan, Y. Long, X. Ji, and W. Xu, "The catcher in the field: A fieldprint based spoofing detection for text-independent speaker verification," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2019, pp. 1215–1229.

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

[18] S. Scherr, S. Ayhan, B. Fischbach, A. Bhutani, M. Pauli, and T. Zwick, "An efficient frequency and phase estimation algorithm with CRB performance for FMCW radar applications," *IEEE Trans. Instrum. Measure.*, vol. 64, no. 7, pp. 1868–1875, Jul. 2015.

[19] C. Lin, S. Chang, C. Chang, and 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.

[20] F. K. Schwering, E. J. Violette, and R. H. Espeland, "Millimeter-wave propagation in vegetation: Experiments and theory," *IEEE Trans. Geosci. Remote Sens.*, vol. 26, no. 3, pp. 355–367, May 1988.

[21] Z. Wang, Z. Chen, A. D. Singh, L. Garcia, J. Luo, and M. B. Srivastava, "UWHear: Through-wall extraction and separation of audio vibrations using wireless signals," in *Proc. 18th Conf. Embedded Netw. Sensor Syst.*, 2020, pp. 1–14.

[22] N. Roy and R. Roy Choudhury, "Listening through a vibration motor," in *Proc. 14th Annu. Int. Conf. Mobile Syst., Appl., Serv.*, 2016, pp. 57–69.

[23] G. Fant, *Acoustic Theory of Speech Production: With Calculations Based on X-ray Studies of Russian Articulations*. Tubingen, Germany: Walter de Gruyter, 1970, no. 2.

[24] B. J. Kröger, G. Schröder, and C. Opgen-Rhein, "A gesture-based dynamic model describing articulatory movement data," *J. Acoustical Soc. Amer.*, vol. 98, no. 4, pp. 1878–1889, 1995.

[25] B. Lohman, O. Boric-Lubecke , V. Lubecke, P. Ong, and M. Sondhi, "A digital signal processor for doppler radar sensing of vital signs," *IEEE Eng. Med. Biol. Mag.*, vol. 21, no. 5, pp. 161–164, Sep./Oct. 2002.

[26] C. Li, V. M. Lubecke, O. Boric-Lubecke , and J. Lin, "A review on recent advances in doppler radar sensors for noncontact healthcare monitoring," *IEEE Trans. Microw. Theory Techn.*, vol. 61, no. 5, pp. 2046–2060, May 2013.

[27] G. Wang, J.-M. Munoz-Ferreras , C. Gu, C. Li, and R. Gómez-García, "Application of linear-frequency-modulated continuous-wave (LFMCW) radars for tracking of vital signs," *IEEE Trans. Microw. Theory Techn.*, vol. 62, no. 6, pp. 1387–1399, Jun. 2014.

[28] 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.

[29] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, pp. 303–304, 1995.

[30] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[31] D. Povey *et al.*, "The kaldi speech recognition toolkit," in *Proc. IEEE Workshop Automat. Speech Recognit. Understanding*, 2011, pp. 1–4.

[32] F. Lin, C. Song, Y. Zhuang, W. Xu, C. Li, and K. Ren, "Cardiac scan: A non-contact and continuous heart-based user authentication system," in *Proc. 23rd Annu. Int. Conf. Mobile Comput. Netw.*, 2017, pp. 315–328.

[33] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright, "Convergence properties of the Nelder–Mead simplex method in low dimensions," *SIAM J. Optim.*, vol. 9, no. 1, pp. 112–147, 1998.

[34] M. L. Attiah, M. Ismail, R. Nordin, and N. F. Abdullah, "Dynamic multi-state ultra-wideband mm-wave frequency selection for 5G communication," in *Proc. IEEE 12th Malaysia Int. Conf. Commun*, 2015, pp. 219–224.

[35] D. Schleher, "Radar detection in Weibull clutter," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-12, no. 6, pp. 736–743, Nov. 1976.

[36] J. D. Park and W. J. Kim, "An efficient method of eliminating the range ambiguity for a low-cost FMCW radar using VCO tuning characteristics," *IEEE Trans. Microw. Theory Techn.*, vol. 54, no. 10, pp. 3623–3629, Oct. 2006.

[37] Y. Xu, S. Wu, C. Chen, J. Chen, and G. Fang, "A novel method for automatic detection of trapped victims by ultrawideband radar," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 8, pp. 3132–3142, Aug. 2012.

[38] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 5, pp. 569–586, Sep. 1999.

[39] J. Wang, "Physiologically-motivated feature extraction methods for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 1690–1694.

[40] H. A. Patil and P. N. Baljekar, "Classification of normal and pathological voices using TEO phase and Mel cepstral features," in *Proc. Int. Conf. Signal Process. Commun.*, 2012, pp. 1–5.

[41] S. Das and J. H. Hansen, "Detection of voice onset time (VOT) for unvoived stops (/p/,/t/,/k/) using the teager energy operator (TEO) for automatic detection of accented English," in *Proc. 6th Nordic Signal Process. Symp.*, , 2004, pp. 344–347.

[42] D. Hosseinzadeh and S. Krishnan, "Combining vocal source and MFCC features for enhanced speaker recognition performance using GMMs," in *Proc. IEEE 9th Workshop Multimedia Signal Process.*, 2007, pp. 365–368.

[43] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Process. Lett.*, vol. 13, no. 1, pp. 52–55, Jan. 2006.

[44] R. P. Ramachandran, M. S. Zilovic, and R. J. Mammone, "A comparative study of robust linear predictive analysis methods with applications to speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 2, pp. 117–125, Mar. 1995.

[45] P. Li, F. Hu, Y. Li, and Y. Xu, "Speaker identification using linear predictive cepstral coefficients and general regression neural network," in *Proc. 33rd Chin. Control Conf.*, 2014, pp. 4952–4956.

[46] I. V. McLoughlin , "Line spectral pairs," *Signal Process.*, vol. 88, no. 3, pp. 448–467, 2008.

[47] S. M. Saqlain *et al.*, "Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines," *Knowl. Inf. Syst.*, vol. 58, no. 1, pp. 139–167, 2019.

[48] J. E. Kelley, Jr,, "The cutting-plane method for solving convex programs," *J. Soc. Ind. Appl. Math.*, vol. 8, no. 4, pp. 703–712, 1960.

[49] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "More efficiency in multiple Kernel learning," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 775–782.

[50] S. Nakagawa, K. Asakawa, and L. Wang, "Speaker recognition by combining MFCC and phase information," in *Proc. 8th Annu. Conf. Int. Speech Commun. Assoc.*, 2007, pp. 4502–4505.

[51] J. McLaughlin, D. A. Reynolds, and T. Gleason, "A study of computation speed-ups of the GMM-UBM speaker recognition system," in *Proc. 6th Eur. Conf. Speech Commun. Technol.*, 1999, pp. 1215–1218.

[52] A. F. Martin and M. A. Przybocki, "The NIST speaker recognition evaluations: 1996–2001," in *Proc. Speaker Odyssey-Speaker Recognit. Workshop*, 2001, pp. 1–5.

[53] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[54] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Comput. Speech Lang.*, vol. 20, no. 2–3, pp. 210–229, 2006.

[55] M. S. Deshpande and R. S. Holambe, "Text-independent speaker identification using hidden Markov models," in *Proc. 1st Int. Conf. Emerging Trends Eng. Technol.*, 2008, pp. 641–644.

[56] L. E. Baum and J. A. Eagon, "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology," *Bull. Amer. Math. Soc.*, vol. 73, no. 3, pp. 360–363, 1967.

[57] T. Das and K. Nahar, "A voice identification system using hidden Markov model," *Indian J. Sci. Technol.*, vol. 9, no. 4, pp. 1–6, 2016.

[58] N. Brümmer and E. De Villiers , "The Bosaris toolkit: Theory, algorithms and code for surviving the new DCF," 2013, *arXiv:1304.2865*.

[59] What is the mmWave transceiver system. Accessed: 2020. [Online]. Available: http://www.ni.com/sdr/mmwave/

[60] S32r27 reference design kit for high-performance automotive radar. Accessed: 2020. [Online]. Available: https://www.nxp.com/products/power-management/system-basis-chips/functi onal-safety-sbcs/s32r27-reference-design-kit-for-high-performance-automotive-r adar:RDK-S32R274

[61] J. Hasch, E. Topak, R. Schnabel, T. Zwick, R. Weigel, and C. Waldschmidt, "Millimeter-wave technology for automotive radar sensors in the 77 GHz frequency band," *IEEE Trans. Microw. Theory Techn.*, vol. 60, no. 3, pp. 845–860, Mar. 2012.

[62] J. Singh, B. Ginsburg, S. Rao, and K. Ramasubramanian, "AWR1642 mmWave sensor: 76–81-GHz radar-on-chip for short-range radar applications," *Texas Instrum.*, pp. 1–7, 2017. [Online]. Available: https://www.ti.com/lit/wp/spyy006/spyy006.pdf

[63] R. Liu, C. Cornelius, R. Rawassizadeh, R. Peterson, and D. Kotz, "Vocal resonance: Using internal body voice for wearable authentication," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 1, p. 19, 2018.

[64] B. Zhou, J. Lohokare, R. Gao, and F. Ye, "EchoPrint: Two-factor authentication using acoustics and vision on smartphones," in *Proc. 24th Annu. Int. Conf. Mobile Comput. Netw.*, 2018, pp. 321–336.

[65] M. Zhadobov, N. Chahat, R. Sauleau, C. Le Quement , and Y. Le Drean , "Millimeter-wave interactions with the human body: State of knowledge and recent advances," *Int. J. Microw. Wirel. Technol.*, vol. 3, no. 2, p. 237–247, 2011.

[66] T. Chugh, K. Cao, and A. K. Jain, "Fingerprint spoof buster: Use of minutiae-centered patches," *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 9, pp. 2190–2202, Sep. 2018.

[67] Z. Li, Z. Yang, C. Song, C. Li, Z. Peng, and W. Xu, "E-Eye: Hidden electronics recognition through mmWave nonlinear effects," in *Proc. 16th ACM Conf. Embedded Netw. Sensor Syst.*, 2018, pp. 68–81.

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

[69] Q. Wang *et al.*, "VoicePop: A pop noise based anti-spoofing system for voice authentication on smartphones," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, 2019, pp. 2062–2070.

[70] 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.

[71] L. Blue, H. Abdullah, L. Vargas, and P. Traynor, "2MA: Verifying voice commands via two microphone authentication," in *Proc. Asia Conf. Comput. Commun. Secur.*, 2018, pp. 89–100.

[72] J. Qian, H. Du, J. Hou, L. Chen, T. Jung, and X.-Y. Li, "Hidebehind: Enjoy voice input with voiceprint unclonability and anonymity," in *Proc. 16th ACM Conf. Embedded Netw. Sensor Syst.*, 2018, pp. 82–94.

[73] A. Janicki, F. Alegre, and N. Evans, "An assessment of automatic speaker verification vulnerabilities to replay spoofing attacks," *Secur. Commun. Netw.*, vol. 9, no. 15, pp. 3030–3044, 2016.

[74] J. Shang, S. Chen, and J. Wu, "Defending against voice spoofing: A robust software-based liveness detection system," in *Proc. IEEE 15th Int. Conf. Mobile Ad Hoc Sensor Syst.*, 2018, pp. 28–36.

[75] J. Shang, S. Chen, and J. Wut, "SRVoice: A robust sparse representation-based liveness detection system," in *Proc. IEEE 24th Int. Conf. Parallel Distrib. Syst.*, 2018, pp. 291–298.

[76] F. Adib, H. Mao, Z. Kabelac, D. Katabi, and R. C. Miller, "Smart homes that monitor breathing and heart rate," in *Proc. 33rd Annu. ACM Conf. Hum. Factors Comput. Syst.*, 2015, pp. 837–846.

[77] O. J. Kaltiokallio, H. Yigitler, R. Jäntti, and N. Patwari, "Non-invasive respiration rate monitoring using a single COTS TX-RX pair," in *Proc. 13th Int. Symp. Inf. Process. Sensor Netw.*, 2014, pp. 59–70.

[78] J. Lien *et al.*, "Soli: Ubiquitous gesture sensing with millimeter wave radar," *ACM Trans. Graph.*, vol. 35, no. 4, p. 142, 2016.

[79] W. Jiang *et al.*, "Towards environment independent device free human activity recognition," in *Proc. 24th Annu. Int. Conf. Mobile Comput. Netw.*, 2018, pp. 289–304.

[80] M. Zhao, S. Yue, D. Katabi, T. S. Jaakkola, and M. T. Bianchi, "Learning sleep stages from radio signals: A conditional adversarial architecture," in *Proc. 34th Int. Conf. Mach. Learn..*, 2017, pp. 4100–4109.

[81] M. Zhao, F. Adib, and D. Katabi, "Emotion recognition using wireless signals," in *Proc. 22nd Annu. Int. Conf. Mobile Comput. Netw.*, 2016, pp. 95–108.

[82] Z. Yang, P. H. Pathak, Y. Zeng, X. Liran, and P. Mohapatra, "Monitoring vital signs using millimeter wave," in *Proc. 17th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, 2016, pp. 211–220.

[83] D. T. Petkie, E. Bryan, C. Benton, and B. D. Rigling, "Millimeter-wave radar systems for biometric applications," in *Proc. SPIE Millimetre Wave Terahertz Sensors Technol. II*, 2009, p. 748502.

[84] J.-H. Choi, J.-H. Jang, and J.-E. Roh, "Design of an FMCW radar altimeter for wide-range and low measurement error," *IEEE Trans. Instrum. Measure.*, vol. 64, no. 12, pp. 3517–3525, Dec. 2015.

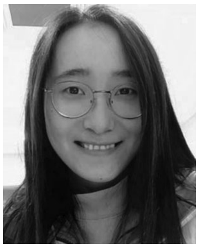
[85] F. Chen *et al.*, "A novel method for speech acquisition and enhancement by 94 GHz millimeter-wave sensor," *Sensors*, vol. 16, no. 1, p. 50, 2016.

[86] A. Bud, "Facing the future: The impact of Apple faceID," *Biometric Technol. Today*, vol. 2018, no. 1, pp. 5–7, 2018.

[87] F. Al-Turjman , "5G-enabled devices and smart-spaces in social-IoT: An overview," *Future Gener. Comput. Syst.*, vol. 92, pp. 732–744, 2019.

[88] "High-performance DSP and control processing for complex 5G requirements." Accessed: 2020. [Online]. Available: https://www.synopsys.com/designware-ip/technical-bulletin/high-performa nce-dsp-for-5g-dwtb-q418.html

[89] Texas Instruments. "AWR1642 datasheet," 2021. [Online]. Available: https://www.ti.com/lit/ds/symlink/awr1642.pdf?ts=1619028562635



Huining Li received the BS degree from Nanjing University of Information Science and Technology in 2016. She is currently working toward the PhD degree at the Department of Computer Science and Engineering, State University of New York at Buffalo. Her research interests include wireless sensing, human-computer Interaction, as well as smart health.

Chenhan Xu is currently working toward the third year of PhD degree at the Department of Computer Science and Engineering, State University of New York at Buffalo. His research interests include Internet of Things, mobile computing, human-computer interaction, and mobile health.

Aditya Singh Rathore received the BS degree in computer engineering in 2017 from the State University of New York at Buffalo, where he is currently working toward the fourth year of PhD degree with the Department of Computer Science and Engineering. His research interests include mobile security, biometrics, Internet of Things, and human-computer interaction. During his academic term, he has been the recipient of multiple honors including UB International Freshman Scholarship, SEAS Senior Scholar Research Scholarship and Presidential Fellowship from the University at Buffalo, and the ACM APSys Student Travel award. Most recently, he was the recipient of the Best Paper Award in ACM MobiSys'20.

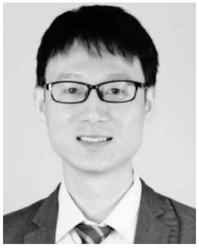
Zhengxiong Li (Member, IEEE) received the BS and MS degrees in electronic engineering from Hangzhou Dianzi University. He is currently working toward the PhD degree at the Department of Computer Science and Engineering, State University of New York at Buffalo. His research interests include Internet-of-Things (IoT), cyber-physical security, emerging technologies, and applications (e.g., smart health).

Hanbin Zhang (Member, IEEE) is currently working toward the PhD degree with the Department of Computer Science and Engineering, State University of New York at Buffalo. His life goal is to become a scientist in smart health and he would love to partner with doctors, nurses, and other healthcare professionals to help improve the care patients receive. Throughout his PhD career, he collaborated with medical professionals and leveraged AI and mobile sensing to assist in diagnosing Parkinson's disease, predicting health outcomes of patients with Alzheimer's disease, preventing stroke, and managing stroke rehabilitation. His research interests include smart health and AI.

Chen Song (Memeber, IEEE) received the BS degree from Fudan University, Shanghai, China, in 2010, the MS degree in electrical engineering, and the PhD degree in computer science and engineering from the State University of New York at Buffalo, New York, NY, USA, in 2012 and 2019, respectively. He is currently an assistant professor of computer science at San Diego State University. He has actively participated in more than 40 authored or coauthored research papers and coauthored one book. His research interests include cognitive security and reliable interaction.

Kun Wang (Senior Member, IEEE) received two PhD degrees, one in computer science from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2009, and the other from the University of Aizu, Aizuwakamatsu, Japan, in 2018. He was a postdoctoral fellow with the University of California, LA, USA, from 2013 to 2015, where he is currently a senior research professor. He was a research fellow with the Hong Kong Polytechnic University, Hong Kong, from 2017 to 2018, and a professor with the Nanjing University of Posts and Telecommunications. His research interests include the area of Artificial Intelligence and Internet of Things, AI hardware acceleration, and blockchain. He was the recipient of the ACM CHI 2020 Hourable Mention Award, ACM FPGA 2020 Best Paper Award Candidate, ACM MobiSys 2020 Best Paper Award, ACM SenSys 2019 Best Paper Award, IEEE GLOBECOM 2016 Best Paper Award, IEEE ISJ Best Paper Award 2019, IEEE TCGCC Best Paper Award 2018, IEEE TCBD Best Paper Award 2019, and CBD 2019 Best Student Paper Award. He is/was the symposium chair/co-chair of the IEEE GLOBECOM 2021, IEEE CNCC 2017, IEEE WCSP 2016, etc. He was an associate editor for the *IEEE Access*, an editor of *Journal of Network and Computer Applications*, and a guest editor of the *IEEE Network*, *Future Generation Computer Systems*, *Journal of Systems Architecture*, *Peer-to-Peer Networking and Applications*, the *IEICE Transactions on Communications*, *IEEE Access*, *Journal of Internet Technology*, and *Future Internet*.

Lu Su (Senior Member, IEEE) received the PhD degree in computer science, and the MS degree in statistics, both from the University of Illinois at Urbana-Champaign, in 2013 and 2012, respectively. He is currently an associate professor at the School of Electrical and Computer Engineering, Purdue University. He has also worked with IBM T. J. Watson Research Center and National Center for Supercomputing Applications. He has authored or coauthored more than 100 papers in referred journals and conferences. His research interests include the general areas of Internet of Things and cyber-physical systems, with a current focus on wireless, mobile, and crowd sensing systems. He is an associate editor of the *ACM Transactions on Sensor Networks*. He was the recipient of NSF CAREER Award, University at Buffalo Young Investigator Award, ICCPS'17 Best Paper Award, and the ICDCS'17 Best Student Paper Award. He is a member of the ACM.



Feng Lin (Senior Member, IEEE) received the PhD degree from the Department of Electrical and Computer Engineering, Tennessee Technological University, USA, in 2015. He is currently a professor at the School of Cyber Science and Technology, College of Computer Science and Technology, Zhejiang University, China. He was an assistant professor at the University of Colorado Denver, USA, a research scientist with the State University of New York at Buffalo, USA, and an engineer with Alcatel-Lucent (currently, Nokia). His research interests include mobile sensing, Internet of Things security, biometrics, AI security, and IoT applications. He was a recipient of the Best Paper Awards from the ACM MobiSys'20, IEEE Globecom'19, IEEE BHI'17, and the Best Demo Award from ACM HotMobile'18, and the First Prize Design Award from the 2016 International 3-D printing competition.

Kui Ren (Fellow, IEEE) received the PhD degree in electrical and computer engineering from Worcester Polytechnic Institute. He is currently a professor and an associate dean of the College of Computer Science and Technology, Zhejiang University, where he also directs the Institute of Cyber Science and Technology. Before that, he was with the State University of New York at Buffalo. He has authored or coauthored extensively in peer-reviewed journals and conferences. His h-index is 74, and his total publication citation exceeds 32,000 according to Google Scholar. His research interests include data security, IoT security, AI security, and privacy. He was the recipient of Guohua Distinguished Scholar Award from ZJU in 2020, the IEEE CISTC Technical Recognition Award in 2017, SUNY Chancellor's Research Excellence Award in 2017, Sigma Xi Research Excellence Award in 2012, and NSF CAREER Award in 2011. He was also the recipient of the Test-of-time Paper Award from IEEE INFOCOM and many Best Paper Awards from the IEEE and ACM including MobiSys'20, ICDCS'20, Globecom'19, ASIACCS'18, and ICDCS'17. He is a distinguished member of the ACM and a Clarivate highly-cited researcher. He is a frequent reviewer for funding agencies internationally and is on the editorial boards of many IEEE and ACM journals. He is currently the chair of SIGSAC of ACM China.

Wenyao Xu (Senior Member, IEEE) received the PhD degree from the University of California at Los Angeles, LA, USA, and both thebachelor's and master's degrees from Zhejiang University, China. He is currently an associate professor at tenure of the Computer Science and Engineering Department, State University of New York at Buffalo. The results of his research interest have been published in peer reviewed top research venues across multiple disciplines, including Computer Science conferences (e.g., ACM MobiCom, SenSys, MobiSys, UbiComp, ASPLOS, ISCA, HPCA, Oakland, NDSS, and CCS), Biomedical Engineering journals (e.g., IEEE TBME, TBioCAS, and JBHI), and Medicine journals (e.g., LANCET). To date, his group has authored or coauthored more than 180 peer-reviewed papers. His research interests include exploring novel sensing and computing technologies to build up innovative Internet-of-Things (IoT) systems for high-impact human-technology applications in the fields of smart health and cyber-security. To date, his group has been the recipient of nine best paper awards, two best paper nominations, and three international best design awards. His inventions have been filed within U.S. and internationally as patents, and have been licensed to industrial players. His research has been reported in high-impact media outlets, including the Discovery Channel, CNN, NPR, and the Wall Street Journal. He is currently an associate editor for the *IEEE Transactions on Biomedical Circuits and Systems*, the technical program committee of numerous conferences in the field of smart health and Internet of Things, and has been a TPC co-chair of the IEEE Body Sensor Networks in 2018.