WavoID: Robust and Secure Multi-modal User Identification via mmWave-voice Mechanism

Tiantian Liu
Zhejiang University
Zhejiang, China
tiantian@zju.edu.cn

Feng Lin*
Zhejiang University
Zhejiang, China
flin@zju.edu.cn

Chao Wang
Zhejiang University
Zhejiang, China
wangchao5001@zju.edu.cn

Chenhan Xu
University at Buffalo
New York, USA
chenhanx@buffalo.edu

Xiaoyu Zhang
University at Buffalo
New York, USA
zhang376@buffalo.edu

Zhengxiong Li
University of Colorado Denver
Colorado, USA
zhengxiong.li@ucdenver.edu

Wenyao Xu
University at Buffalo
New York, USA
wenyaoxu@buffalo.edu

Ming-Chun Huang
Duke Kunshan University
Jiangsu, China
mingchun.huang@dukekunshan.edu.cn

Kui Ren
Zhejiang University
Zhejiang, China
kuiren@zju.edu.cn

ABSTRACT

With the increasing deployment of voice-controlled devices in homes and enterprises, there is an urgent demand for voice identification to prevent unauthorized access to sensitive information and property loss. However, due to the broadcast nature of sound wave, a voice-only system is vulnerable to adverse conditions and malicious attacks. We observe that the cooperation of millimeter waves (mmWave) and voice signals can significantly improve the effectiveness and security of user identification. Based on the properties, we propose a multi-modal user identification system (named WavoID) by fusing the uniqueness of mmWave-sensed vocal vibration and mic-recorded voice of users. To estimate fine-grained waveforms, WavoID splits signals and adaptively combines useful decomposed signals according to correlative contents in both mmWave and voice. An elaborated anti-spoofing module in WavoID comprising biometric bimodal information defend against attacks. WavoID produces and fuses the response maps of mmWave and voice to improve the representation power of fused features, benefiting accurate identification, even facing adverse circumstances. We evaluate WavoID using commercial sensors on extensive experiments. WavoID has significant performance on user identification with over 98% accuracy on 100 user datasets.

CCS CONCEPTS

• **Human-centered computing** → *Interactive systems and tools.*

KEYWORDS

User authentication, voice identification, mmWave sensing, multi-modal fusion

ACM Reference Format:
Tiantian Liu, Feng Lin, Chao Wang, Chenhan Xu, Xiaoyu Zhang, Zhengxiong Li, Wenyao Xu, Ming-Chun Huang, and Kui Ren. 2023. WavoID: Robust and Secure Multi-modal User Identification via mmWave-voice Mechanism. In *The 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23), October 29–November 01, 2023, San Francisco, CA, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3586183.3606775

*Corresponding author

UIST '23, October 29–November 01, 2023, San Francisco, CA, USA

ACM ISBN 979-8-4007-0132-0/23/10…$15.00
https://doi.org/10.1145/3586183.3606775

1 INTRODUCTION

Recently, voice identification has become a boosted topic as a more "natural" means for recognizing users [10, 65]. Voice identification liberates users from remembering passwords and auxiliary gesturing due to the unity of command and verification once receiving utterances. With its great convenience, voice identification is expanding throughout social and individual applications, including smart speakers, financial systems [57], and entrance guards [20].

Whereas voice-controlled systems are rapidly expanding into companies and public facilities, a voice system is desired to achieve goals of robustness and security. However, due to the inherent broadcast nature of sound waves, voice-only identification is susceptible to malicious attacks. Except for replay attacks [76, 77] and mimicry attacks [76], nowadays, attackers can launch adversarial attacks [11, 12] to make voice-identification systems misclassify an adversarial token as an enrolled speaker. Additionally, such unimodal systems, i.e., voice-only systems, suffer from accuracy degradation when exposed to noise and reverberation.

To alleviate those risks and deficiencies, researchers have proposed multi-modal systems. Multi-modal identification integrates complementary modalities, e.g., gesture [44, 47], WIFI [43, 48], image [1, 30], and ECG [9], to discriminate multiple traits unique to users. These approaches only take into account defending against unimodal attacks targeting voice, overlooking multi-modal attacks aiming at both voice and additional modalities. Furthermore, these

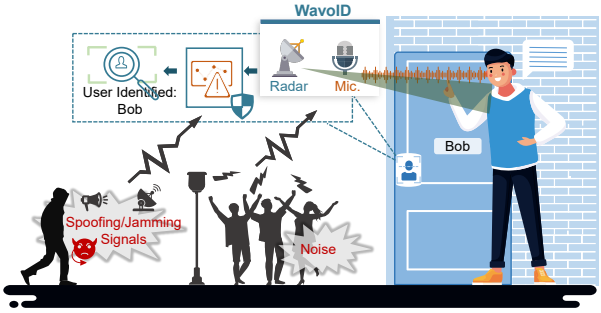


Figure 1: An illustration of WavoID to identify a user and reject malicious signals in noisy scenes, by fusing mmWave signals and voice signals.

combined modalities, like WIFI and image, are independently exploited, leading to insufficient discriminative capabilities when faced with interference like motion interference or light changes. Some multi-modal methods, like ECG-voice and gesture-voice, require users to execute specific actions, which constrains the user experience of voice-controlled systems.

After thoroughly understanding the need and challenge of identification systems in voice-controlled devices, the system should satisfy the following requirements: (i) no auxiliary operation: it carries on identification tasks once a user speaks commands without any extra operations. (ii) resilient to adverse conditions: it should be robust to the adverse conditions such as noise, user motion interference, and environmental changes, out of the need of deployed in open areas. (iii) resistant to multiple attacks: it should guarantee the security and privacy of users no matter what multi-modal attacks or unimodal attacks.

We focus on millimeter waves(mmWaves), referring to radio-frequency signals whose frequencies start at 24GHz and above. As its high resolution in tiny vibration measurement [28], mmWave signals have huge potential in applications such as wireless sensing [33], imaging [4], and communications [45]. Prior work [36] leverages the vocal displacement uniqueness captured by mmWave signals to achieve noise-resistant user identification. Despite the advantage of noise resistance and user-friendliness, mmWave-only identification is vulnerable to motion interference of users. In contrast, voice signals are robust to motion interference, which can compensate for the demerit of mmWave signals to some extent. Therefore, we consider a multi-modal identification system fusing mmWave and voice modality, as shown in Figure 1, exploring their strengths to enhance the identification capability of the system. With the aid of their correlative cooperation, the multi-modal system promises user identification in spite of interference and malicious attacks.

To realize a practical and secure mmWave-voice identification system, we need to cope with the following challenges. (i) How to fuse two intrinsic features from different modalities to maximize the system's effectiveness. (ii) How to estimate fine-grained mmWave and voice signals for subsequent identification in spite of interference on both signals. (iii) How to distinguish genuine users from adversaries when facing multiple attacks against mmWave and voice modality.

In this paper, we introduce the inherent correlation between the mmWave modality and voice modality. Based on their correlation, we propose WavoID, a multi-modal identification system fusing mmWave signals and voice signals for user identification. The proposed WavoID firstly extracts the fine-grained signals based on the inherent correlation between mmWave and voice modality. Then, WavoID introduces a bimodal liveness detection by using biometric information from these two modalities to resist multi-modal attacks. To enable the fusion of two different modalities, WavoID generates response maps like thermodynamic diagrams weighing the importance of elements in mmWave or voice domain. Using the sectioned convolution [59] among these produced response maps can enhance bimodal characteristics and still preserve individual features. Finally, the fused feature is fed into the identification network to identify users. To prove the built-in superiority of WavoID, this paper gives a theoretical analysis from a statistical perspective.

In conclusion, the contributions of our work are as follows.

- We design a multi-modal identification called WavoID, leveraging the uniqueness of vocal vibrations and voice biometrics sensed by mmWave radars and microphones. We also theoretically analyze its superiority on identification from statistics.
- We utilize cross-modal knowledges between mmWave and voice domain to refine the fine-grained signals and then selectively fuse features for robust identification. The cross-modal mechanism can boost the discrimination ability of the system under adverse conditions.
- We evaluate the security of WavoID in resisting comprehensive attacks, including counterfeit, jamming, replay, mimicry, and adversarial attacks. WavoID can reject above 99% malicious samples through the designed bimodal liveness detection.
- We demonstrate the effectiveness and robustness of WavoID among the 100 user dataset. Experimental results show that WavoID maintains a balanced accuracy over 98% and an average equal error rate of 1.24%.

2 RELATED WORK

This work builds and extends on prior work mainly in three fields: voice identification, mmWave-based identification, and multi-modal identification. We clarify the position of our study after summarizing these fields.

2.1 Voice identification.

Voice identification, as a non-intrusive biometric identification [7], uses innate biometric characteristics of users' voices to distinguish the identity of input voices and protect voice-controlled devices from non-user voices like Siri and Alexa. Prior research on voice identification focuses more on exploiting acoustic features unique to users, such as linear predictive cepstral coefficients and the mel-frequency cepstral coefficients, which will be fed into machine learning models to identify persons. Recently, motivated by the advanced deep learning networks (DNNs), mainstream voice identification applies the encoder-decoder framework for feature modeling and speaker matching. However, studies have demonstrated that voice identification is vulnerable to spoofing attacks, i.e., replay attacks, mimicry attacks, and adversarial attacks. Kinnunen et al. [32] have shown that replaying recorded user samples can deceive

voice identification systems into permitting unauthenticated audio. Nowadays, researchers consider adding a third phase which aims to detect malicious voice input. A few studies [3, 71, 80] characterize spectro-temporal distortions between genuine and spoofing voices to defend against attacks and identify users. Rajaratnam et al. [50] propose to detect adversarial examples by flooding particular frequency bands of audio signals with random noise. Except for the utilization of unimodal voice modality, there is a visible trend that auxiliary modalities are also used in voice identification systems. VoiceGesture [85] uses built-in loudspeaker-microphone pairs of smartphones to emit ultrasound to sense articulatory gestures for liveness detection during voice authentication. Pradhan et al. [49] propose combining acoustic and WiFi channels to detect spoofing attacks by exploiting the synchronized changes in voice features and breathing sensed by WiFi.

2.2 mmWave-based sensing.

The mmWave sensing technology has aroused enthusiasm among different domains ranging from human activity recognition, vital signs monitoring, and wireless communication, owing to its high resolution and short wavelength[37, 39, 69, 70]. Towards privacy-preserving tracking manner, Kong et al. [33] extract shape features from mmWave sensing signals and reconstruct 3D human posture through a designed deep learning model. Xu et al. [78] present CardiacWave to achieve non-contact heart monitoring based on the frequency response of the cardiac electromagnetic field under mmWave interrogation. Previous studies have demonstrated the feasibility of mmWave in user identification due to its exceptional ability to sense the slightest vibrations. Yang et al. [82] designed a multi-person detection and identification system by using mmWave to sense multi-person's distinct gait patterns. Together with the increasing popularity of voice user interfaces, there is an emerging trend in the development of speech recovery for millimeter wave radar. Recent works have proven that the phase and amplitude of mmWave can be modulated by sensed vibrations of human vocal activities, which can be exploited to extract speech information. Xu et al. [79] collect reflected mmWave signals from users' throats and transform their spectrum into speech spectrum through a deep neural network. Hu et al. [26] use the generative adversary network to reconstruct audio directly from the captured mmWave spectrogram without prior knowledge. Despite the advantage of noise resistance and user-friendliness, mmWave-only sensing is vulnerable to the motion interference of users. Fortunately, voice signals are robust to motion interference, which can compensate for the demerit of mmWave signals. Therefore, we consider a multi-modal identification system fusing mmWave and voice modality, exploring their strengths to enhance the identification capability of the system. With the aid of their correlative cooperation, the multi-modal system promises user identification despite interference and malicious attacks.

2.3 Multi-modal identification.

Multi-modal identification fuses multiple information sources from different sensors for identification, outperforming unimodal identification in stability and security [48, 86]. SpeechXRays, devised by Adami et al. [2], is an audio-visual identification system that provides anti-spoofing capabilities to enable secure access to eHealth. Moreover, Gupta et al. [21] utilize three biometric modalities, i.e., swipe, voice, and face, to construct multi-modal identification to guarantee riders' security. Recently, deploying multi-modal identification has become a trend on mobile devices. Khan et al. [29] design Itus as a framework for implicit smartphone identification. Itus allows researchers to incorporate multiple sensor data or other identification mechanisms into the platform. Similarly, CORMORANT proposed by Hintze et al. [25], is a multi-modal identification system that can deploy continuous cross-sensor identification. However, these multi-modal methods overlook the importance of defense against multi-modal attacks and independently exploit each modality, which cannot promise systems' significant effectiveness and security at the utmost. Unlike all existing works, WavoID completely fuses multiple modalities based on their intrinsic correlation to defend against multi-modal attacks significantly and accurately identify users in complex scenes.

2.4 Position of our study

To the best of our knowledge, WavoID is the first to integrate mmWave and voice modality as a security guard of voice interaction. The mmWave-based sensing captures throat vibration caused by users' voice activity to identify users, which does not require extra actions and passwords. Despite its advantages, mmWave sensing is sensitive to multipath noise, especially when users move or when environmental conditions change. Inversely, voice-only identification is resistant to multipath interference but is easily impaired by ambient noise. In this study, we overcome their defects to elaborate an identification framework named WavoID by fusing these seemingly different modalities according to their correlation content, benefiting stable and secure voice interactions. Facing traditional problems in voice identification, such as noise reduction and feature fusion, WavoID has the key insight of using information from one modality to guide the system to search and enhance relevant information from another. Regarding resilient security against various attacks, WavoID extracts biometrics characteristics from mmWave-voice modalities and reconstructs bimodal feature maps to reject agressire signal inputs.

WavoID allows users to undergo secure identification procedures without the need to wear or carry any specialized equipment while interacting with voice intelligence products. This user-friendly approach provides a convenient and efficient way for users to access WavoID's identification services.

3 ANALYSIS ON MMWAVE AND VOICE

3.1 Apparatus

Figure 2 shows the setup of the mmWave-microphone combination that captures mmWave signals and voice signals synchronously when participants speak commands. We achieve the deployment of WavoID by using a COTS microphone and a COTS mmWave radar: a IWR1642BOOST radar [63] paired with a data collection board DCA1000EVM [62]. The radar has a sensing field of view of 120° in the azimuth and 30° in the elevation. The configuration of the radar is set as follows: the frequency slope is 15MHz/μs; the frequency

band ranges from 77GHz to 81GHz; the cycle of chirps is 260μs containing 190 data samples, the cycle of frames is 50μs; and the ADC sampling rate is 5000k samples per second. The above configuration guarantees that the radar has a 15m maximum detectable range with 310μm resolution. The radar and microphone are connected and controlled by a laptop, actuated by mmWaveStudio GUI [64] and MATLAB. The laptop synchronously collects mmWave data with above configurations and microphone data with 48k sampling rate. The default experiments are measured in a laboratory with a background noise of 40-60dB. During the data collection and processing, all participants are relaxed and speak at a natural volume.

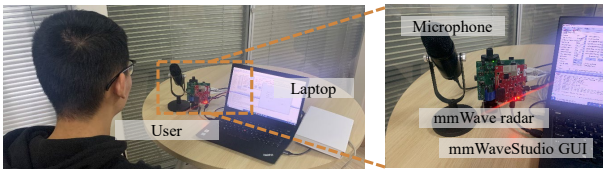


Figure 2: The evaluation setup. A microphone and a mmWave radar collect samples at a distance of 50cm away from the subject.

3.2 Correlation between mmWave and voice

This subsection begins with the mathematical model of voice production and mmWave-sensed vocal vibration. Afterward, we introduce the correlation between voice and mmWave signals based on their models.

Voice production. The production of voice signals is a fluid-structure-acoustic interaction process that depends on the geometry and material properties of the larynx, the vocal tract, and the vocal fold [56, 87]. The airflow induced by the lung propagates through the vocal tract and collides the vocal fold. The fluctuation of the vocal folds modulates the airflow to generate voice phonemes. For unvoiced phonemes, the vocal fold releases or blocks the airflow without constrictions. The vocal fold vibration can be mathematically formulated as a one-degree-of-freedom damping system. The human voice is the cause-effect production when the airflow is modulated by the vocal fold that acts as a filter and modulator. The relationship [15] between the vocal fold vibration and voice can be simplified as:

$$m\ddot{x}(t) + r\dot{x}(t) + kx(t) = e^{j(2\pi f_F t + \phi_F)}, \tag{1}$$

$$v(t) = \mathcal{H}(\dot{x}(t)), \tag{2}$$

where m, r, and k are the parameters depending on physiological properties of the vocal fold. The $\dot{x}(t)$ is the displacement change of vocal fold and $\ddot{x}(t)$ is the second derivative of vocal fold displacement x. The mathematical model of vocal fold vibration $x(t)$ is a one-degree-of-freedom damping system. The $e^{j(2\pi f_F t + \phi_F)}$ is the negative Coulomb force with the frequency f_F and the initial phase ϕ_F, which would vary on the degree of vocal fold tightness. The $\mathcal{H}(\cdot)$ is the transfer function from the vocal fold vibration velocity $\dot{x}(t)$ to the human voice. That is, human voice $v(t)$ is determined by the vocal fold vibration.

mmWave sensing vocal vibration. We leverage a frequency modulated continuous wave (FMCW) radar to measure the vocal vibration. The radar emits electromagnetic chirp signals and then receives the echoes reflected by the subject's throat [28]. The transmitted signal $m_T(t)$ and received signal $m_R(t)$ are given by [60]:

$$m_T(t) = A_T cos 2\pi \left(f_c t + \int_0^t \frac{B}{T_c} \tau d\tau \right), \tag{3}$$

$$m_R(t) = A_R cos 2\pi \left(f_c\,(t - \Delta t) + \int_0^t \frac{B}{T_c} (\tau - \Delta t) d\tau \right), \tag{4}$$

where A_T and A_R are the amplitude of the transmitted signal and the received signal, respectively, f_c is the chirp start frequency, B denotes the bandwidth, T_c denotes the duration, and Δt represents the two round time delay between the radar and the target. Once the radar receives the echo, it immediately uses an integrated mixer and a low-pass filter to produce an intermediate frequency (IF) signal. The IF signal $m_{IF}(t)$ can be mathematically represented as:

$$m_{IF}(t) = \frac{1}{2} A_T A_R cos 2\pi \left(f_c + \frac{B}{T_c} t \right) \Delta t. \tag{5}$$

Thus, we can obtain the phase $\varphi(t) = 2\pi \left(f_c + \frac{B}{T_c} t \right) \Delta t$. Prior research has already revealed that the phase difference contains the displacement of the vocal organ. By differing the phase, we have:

$$\Delta\varphi(t) = \frac{4\pi}{c} \left(f_c + \frac{B}{T_c} t \right) \Delta x(t), \tag{6}$$

where c denotes the speed of light.

The underlying correlation between mmWave and voice. Based on the Eq.6 and Eq.2, it is observed that both phase changes of mmWave signals and voice signals basically rely on the vocal fold displacement, i.e., x. To prove it, we ask a participant to speak to a mmWave radar and microphone while collecting corresponding signal pairs. We calculate the average Pearson correlation coefficient from ten pairs of voice signals and phase difference of mmWave signals. The average Pearson correlation coefficient is 0.4906, which indicates a fairly correlation between the human voice and mmWave signals when sensing human throats. The interrelated content existing in both modalities can benefit the performance improvement of a mmWave-voice system, even when one modality suffers from information loss caused by any interference and attack. With the aid of exchanging each characteristic belonging to mmWave or voice, the signal pattern related to voice activity tends to be enhanced, enabling the signal denoising and fusion.

4 SYSTEM DESIGN

In this section, we describe the workflow of WavoID. Figure 3 provides an overall design of the system to recognize the user's identity by exploiting a mmWave radar and a microphone.

WavoID employs a mmWave radar to sense the vocal vibration when the enrolled user utters a word, and the microphone receives the corresponding voice signal at the same time. The system estimates the fine-grained waveform by decomposing a pair of collected signals and selecting useful decomposed signals in *Fine-grained Waveform Estimation*. After that, WavoID aggregates biometric information from mmWave and voice modality to judge whether the input samples are derived from malicious adversaries. Afterward, in *Cross-modality Fusion*, WavoID fuses two different types of features by utilizing discriminative correlative filters (DCF) to boost the representation of the fused features. The fused features are fed into the pre-trained *Identification Network* to identify the

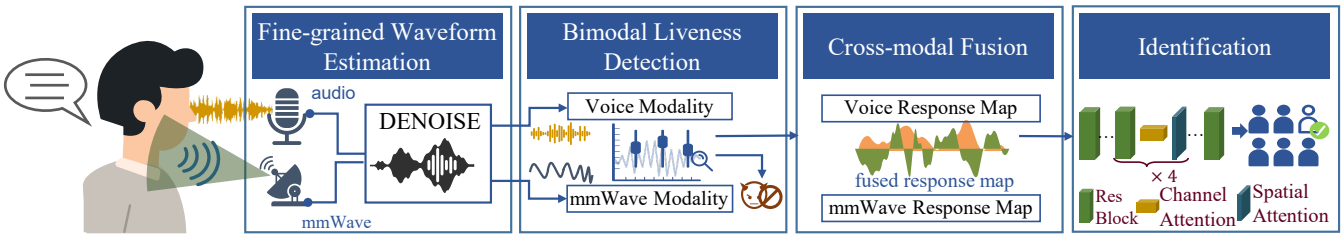
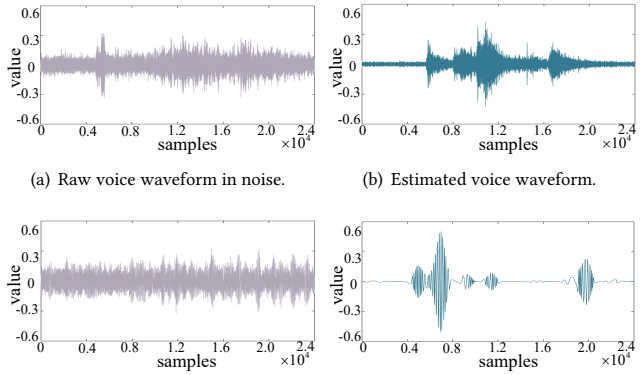


Figure 3: The system overview of WavoID that integrates a mmWave radar and a microphone to achieve user identification against attacks and interference.

legitimate user. We discuss the details for each module of WavoID as follows.



(a) Raw voice waveform in noise. (b) Estimated voice waveform.

(c) Raw mmWave waveform under motion. (d) Estimated mmWave waveform.

Figure 4: Compared with raw noisy signals, the estimated signals including mmWave signals and voice signals via Fine-grained Waveform Estimation preserve informative characteristics regardless of interference.

4.1 Fine-grained Waveform Estimation

When users face sensors of system and speak voice commands, the radar will receive echo signals modulated by vocal vibration and the microphone will synchronously record audio signals. The input mmWave signal of WavoID is the phase difference determined by vocal vibration. In realistic scenes, the obtained audio signals and mmWave signals are prone to be contaminated by ambient noise and motion artifacts, respectively. Given the requirement of convenient identification in various complex applications, it is necessary for the system to estimate the fine-grained waveform. A straightforward approach is to design a band-pass filter to remove the artifacts in frequency domain. However, the filter-based methods rely heavily on abundant reference signals responsible for filter parameters, which is undesirable due to the unpredictable ambient noise and users' body waggling. Therefore, we develop cross-modality waveform estimation to extract fine-grained mmWave and audio signals, described as Algorithm 1.

The workflow of the algorithm is summarized as follows. First, the system performs decomposition methods on the input mmWave and audio signal by using fast independent component analysis (FastICA) [27] and dual-tree complex wavelet transform (DTCWT) [54], respectively. We can obtain a series of decomposed mmWave sub-signals and audio sub-signals. Then, a correlation matrix is constructed by computing Pearson correlation coefficients [53] between a pair of mmWave and audio sub-signals. Hence, the correlation matrix contains the cross-modality information where a higher coefficient indicates a stronger correlation between the corresponding pair of sub-signals. The stronger correlation means that the mmWave sub-signal contains the vocal vibration while the audio sub-signal records voice activity rather than irrelevant noise. Herein, the irrelevant interference such as ambient noise and body motion could be adaptively removed since their coefficients are relatively low and steady, as line (6)~(13) indicate. In (6)~(13), we choose the sub-signal satisfying that the average energy exceeds the average variance by a predetermined ratio. The ratio is empirically set as 5 in our 100 user datasets. Afterwards, the system applies FastICA to each sub-signals and sums up all sub-signals after principal component analysis (PCA). Finally, we can obtain the reconstructed mmWave and audio signal, in which the covering interference is basically removed. The throat vibration and lip movements belong to vocal vibration, which benefits semantic information of mmWave modality. The throat vibration is more informative than lip movements due to its complex high-frequency motion. In fusion processing, the rich motion will be applied more weights since it is more relevant to voice modality. Thus, the throat vibration will receive more attention and utilization in multi-modal fusion and final identification. To verify the algorithm's effectiveness, we compare the waveform processed by the algorithm with the raw waveform. The data, in comparison, is collected from a mmWave radar and a microphone when a user speaks voice commands. The comparison result in Figure 4 shows that the processed voice signal perfectly preserves fine-grained waveform patterns regardless of ambient noise. The mmWave waveform containing the displacement of vocal vibration appears explicitly, which is previously covered by massive noise.

4.2 Bimodal Liveness Detection

Such an identification system based on biometric traits may suffer from artificial attacks such as replay attack, jamming attack, and adversarial attack via some well-designed sensors. Existing defense mechanisms aim to defend against arbitrary attack either on voice modality or mmWave modality. Nevertheless, an attacker can bypass the multi-modal system by simultaneously launching counterfeit voice signals and mmWave signals. To cope with such attacks especially multi-modal attack, we propose a bimodal liveness detection mechanism that can continuously distinguish a genuine one and a malicious one. The key insight of the anti-spoofing method

Algorithm 1 Cross-modality Waveform Estimation

Input:
The sample of mmWave signals m , The sample of audio signals s ;
Output:
The fine-grained mmWave signals $\tilde{m}$, The fine-grained audio signals $\tilde{s}$,

Extract the set of decomposed mmWave samples $m_i, i = 1, 2, 3, \cdots, p$ from m by usng FastICA;
Extract the set of decomposed audio samples $s_j, j = 1, 2, 3, \cdots, q$ from s by running DTCWT;
Compute the correlation matrix: $Coeff(i, j) = Pearson(m_i, s_j)$;
Calculate the mean and variance of the correlation matrix in column and row: $col_m(j) = mean(Coeff(:, j)), col_v(j) = var(Coeff(:, j)), row_m(i) = mean(Coeff(i, :)), row_v(i) = var(Coeff(i, :))$;
// Remove the interference signal
for $i \le p \vee j \le q$ **do**
  if $row_m(i) < \left(\frac{1}{p}\sum_{k=1}^{p} row_v(k)\right) \times row_m(i) \times 5$ **then**
    Remove m_i;
  end if
  if $col_m(j) < \left(\frac{1}{q}\sum_{k=1}^{q} col_v(k)\right) \times col_m(j) \times 5$ **then**
    Remove s_j;
  end if
end for
$\tilde{m}$=$\sum_i$PCA(FastICA(m_i)),$\tilde{s}$=$\sum_j$PCA(FastICA(s_j));
return $\tilde{m}, \tilde{s}$;

is to extract inherent liveness traits subject to voice and mmWave separately, and then aggregate similarity comparison results.

mmWave modality. The reflected mmWave signal can simply be seen as a lossy version of the transmitted signal. When the mmWave signal reaches the human body, the absorption of electromagnetic radiation is mostly restricted to the skin because of the submillimeter depth of penetration [5, 67, 75]. The amplitude of the signal suffers from attenuation because of heterogeneous permittivity determined by the biological tissue of the human skin. Recalling the modal of mmWave signals as shown in Eq.3 and Eq.4, we have the relationship between A_T and A_R as: $A_R = \sqrt{R}e^{-2\alpha L}A_T$, where R is the reflective coefficient [6] depending on biomaterials (e.g., muscle, fat, skin) and α denotes the attenuation constant caused by the distance L between the human body and the radar. We replace the Eq.5 into the Eq.4. Herein, the received signal can also be formulated as:

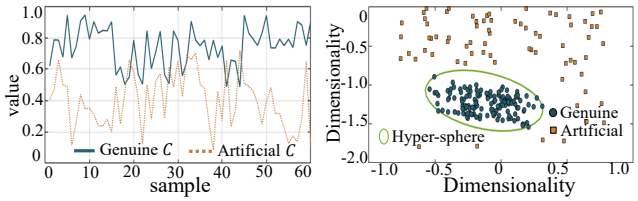
$$m_{IF}(t) = \frac{1}{2}\sqrt{R}e^{-2\alpha L}A_T^2 cos2\pi\left(f_c + \frac{B}{T_c}t\right)\Delta t. \tag{7}$$

The reflective coefficient R is determined by the permittivity of the human skin [6] while other parameters are constant. The amplitude of the IF signal preserves biomaterial properties. To obtain the amplitude containing biomaterial properties to detect liveness, we square the IF signal:

$$m_{IF}^2 = \frac{R}{8}e^{-4\alpha L}A_T^4\left(1 + cos4\pi\left(f_c + \frac{B}{T_c}t\right)\Delta t\right). \tag{8}$$

Then, we extract the DC component $\frac{R}{8}e^{-4\alpha L}A_T^4$ from m_{IF}^2 as liveness traits from the mmWave modality.

Voice modality. As for the voice modality, we also extract the liveness trait that is unique to the individual. Due to the inherent nonlinearity of loudspeakers and microphones, the record-and-play attack is prone to induce nonlinear distortion into the spectral of victim signals. Likewise, the synthetic signals produces unnatural distortion compared to the human voice. It is reasonable to obtain the frequency characteristic to differentiate the original voices and artificial ones. Furthermore, prior research demonstrates that characteristics stemming from acoustic organs are distributed over the frequency band of voice signals. To obtain liveness traits, we first perform the short-time Fourier transform on voice input to acquire a voice spectrum. Then, we segment the low-frequency and high-frequency components of the spectrum rather than choose the audible frequency range. The low-frequency and high-frequency spectrum are rarely affected by the content and are more likely to be observed natural energy peaks differing from loudspeakers and microphones [35]. It costs less time and energy compared with processing on the complete spectrum. The system refines the constant Q cepstral coefficients (CQCC) [66] from the segmented spectrum as voice liveness traits. It has been demonstrated that CQCCs preserve the trait of voice biometrics.



(a) The criterion C. (b) The visualization result.

Figure 5: The comparison result between genuine users and artificial ones.

Bimodal Feature. The system combines the above liveness traits sourced from mmWave and voice modality to defend against adversaries regardless of the mode of attacks. Once the system estimates mmWave and voice signals, these two signals are processed by the two mechanisms above to acquire mmWave and voice liveness features. We assume that M and V individually represent the mmWave and voice features. The system performs normalization on both features with a min-max feature scaling. Next, we match each feature with the corresponding template from the user profile. The system calculates the mmWave similarity sequence S_M between the input mmWave feature and the template M_T using the equation:

$$S_M = \frac{M \times M_T}{\sqrt{M^2 \times M_T^2}}. \tag{9}$$

Similarly, for the voice feature, we compute the voice similarity sequence S_V. The system aggregates these two similarity coefficients to obtain a comprehensive criterion C for assessing the input

source, defined as:

$$C = \min(|\log \frac{S_M}{\sqrt{S_M}}|, |\log \frac{S_V}{\sqrt{S_V}}|). \tag{10}$$

To show the difference between genuine user and artificial signals, we calculate the criterion C when a user speaks commands, or the malicious radar and loudspeaker concurrently transmit artificial signals, respectively. Figure 5(a) displays that the criterion C of a genuine user is relatively higher than artificial ones, since the pair of genuine mmWave and voice signals has higher similarity indexes, leading to the criterion C rising. The obtained criterion from pre-collected samples, also called registration data, is input to the support vector data description (SVDD) [61] to train hyper-sphere, which can accept genuine samples from users and then reject anomalies. The needed number of registration data is no less than 100 pieces (50 samples per piece), which is no longer than 10 seconds. The visualized detection results of refined hyper-sphere is shown in Figure 5(b).

4.3 Cross-modality Fusion

Following the selection of genuine mmWave and voice signals via liveness detection, it is necessary to extract user-specific features for the subsequent identification network. However, due to heterogeneity between the mmWave and voice modality, conventional deep learning methods disregard their intra-correlation and waste much time on invalid information, resulting in feature fusion and enhancement failures. To address it, the system fuses knowledge across the mmWave and voice modality based on discriminative correlation filter (DCF) [40, 41].

The DCF plays an essential role in visual object tracking. It tracks and learns a valid appearance model of the target over frame flow with the advantage of fast processing. The DCF models the region-of-interest (ROI), e.g., people in pictures or even energy trace in spectra, and produces a response map to search for targets considering the temporal and spatial variation in tracking process. An important factor in DCF is the response map, where useful information in ROI is adaptively emphasized with more weights while less informative ones are suppressed. Based on DCFs, WavoID can optimize two response maps individually from the mmWave and voice modality and then significantly fuse them.

For ease of illustration, we only take the mmWave signal as an example to explain how to get the response map. Firstly, the system calculates the spectrum flow of mmWave by using Fast Fourier transform (FFT) on successive 20ms frames of mmWave signals. We assume that $x \in R^2$ is the location of the valuable energy trace. The problem of tracking characteristic traces is formulated by computing a response map r that measures the target location likelihood:

$$\begin{aligned} r(x) &= P(x|\Theta) \\ &= \sum\nolimits_{y \in \Omega_x} P(x, r(y)|\Theta) \\ &= \sum\nolimits_{y \in \Omega_x} P(x|r(y), \Theta) P(r(y)|\Theta), \end{aligned} \tag{11}$$

where Θ denotes that targets, i.e., energy traces, present in the frame and Ω_x is defined as the neighborhood of the location x. The conditional probability $P(x|r(y), \Theta)$ models the spatial relationship between the target and its ambient information. It helps more accurate estimation by utilizing surrounding texture information, even if the x suffers ambiguities under obstructions. According to the spatial context model [84], $P(x|r(y), \Theta)$ can be modeled as $F^{-1}(\frac{F(be^{-|\frac{y-x}{\alpha}|^\beta})}{F(E(y)w(x-y))})$, where b is a normalization constant, α is a scale parameter, β is a hyper-parameter, $E(y)$ is the energy power at the location y in the spectrum, F^{-1} denotes the inverse FFT function, and F denotes the FFT function. The $w(\cdot)$ is a weighted function defined by: $w(y) = ae^{-\frac{|y|^2}{\sigma^2}}$, where a is a normalization constant and σ is a scale parameter. Theoretically, $P(r(y)|\Theta)$ is modeled by $E(y)w(y-x)$. These models take into account different spatial relationships between the target and its ambient textures. It tracks the target like human eyes to focus more on the center of delicate regions requiring more detailed analysis. To obtain the response map and the target location, we optimize the following function:

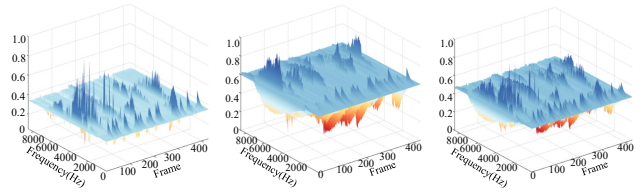
$$\tilde{x} = \text{argmax} F^{-1}\left(F(P(x|r(y), \Theta)) * F(P(x|r(y), \Theta))\right). \tag{12}$$

We use augmented lagrange method (ALM) [8] to update locations and parameters for mmWave response maps. We initially set the trace location x coarsely by using Canny edge detection [52]. Then, the response map is generated when refined locations are tracked. The system can acquire a voice response map from voice spectrum in the same way. The system then performs sectioned convolution between the voice response map and the mmWave response map to fulfill the cross-modality fusion. The sectioned convolution weighted measures and overlays the related response value with fast processing speed, meanwhile weakening the unrelated values corresponding to interferences.

Insight. According to the fusion results in Figure 6, the energy peaks in low-frequency response maps unique to vocal vibration tend to be enhanced, while the characteristic feature individually sourced from medium-frequency mmWave modality and high-frequency voice modality is still retained in the final fusion. Compared with unimodal systems, i.e, mmWave-only or voice-only systems [36, 83], multi-modal systems provide a more comprehensive and vigorous feature representation that can combat arbitrary interference. Specifically, the system can absorb supplementary voice signals from microphones to compensate for damaged vibration information when the radar sensing users in motion. As for the sensing orientation, the multi-modal system integrating a omnidirectional-sensing microphone and a limited-field but long-distance sensing radar will expand his field of perception, outperforming traditional unimodal systems to some extent. On the other hand, vocal vibration features sourced from radar can mitigate the effect of acoustic noise on the recorded voice signals.

4.4 Identification Network

The system adopts CBANet [73] to absorb the fused feature through cross-modality fusion for identification. The CBANet utilizes a residual neural network (ResNet) [24] as the backbone model and recruits two attention-based modules, including channel and spatial attention modules. The insight of the CBANet is channel and spatial attention modules. The nature of attention modules has been extensively validated to learn "what " and "where" to attend in the channel and spatial of feature maps. Specifically, channel and



(a) mmWave. (b) Voice. (c) Fusion.

Figure 6: The produced response maps through Cross-modality Fusion: (a) mmWave: a mmWave response map; (b) Voice: a voice response map; (c) Fusion: fusion production of (a) and (b).

spatial attention modules guide the ResNet to select and enhance meaningful knowledge in mixed feature maps. In channel module in CBANet, given the input feature map as $\mathrm{F} \in \mathbb{R}^{H\times W\times C}$ with the height H, width W, and C channels, the channel module aggregates the channel-wise spatial information of the feature map by using average-pooling $\mathrm{Pool}_{\mathrm{Avg}}$ and max-pooling $\mathrm{Pool}_{\mathrm{Max}}$ operations as the following function:

$$\mathrm{F_c} = \delta(\mathrm{Conv}(\mathrm{Pool}_{\mathrm{Avg}}(\mathrm{F})) + \mathrm{Conv}(\mathrm{Pool}_{\mathrm{Max}}(\mathrm{F}))), \tag{13}$$

where $\delta(\cdot)$ denotes a ReLu function, Conv is one dimension convolution with 1×1 kernel size, and $\mathrm{F_c}$ is the channel attention map. Different from the channel module, the spatial module generates a spatial attention map $\mathrm{F_s}$ that encodes where to emphasize or suppress as described below:

$$\mathrm{F_s} = \delta(\mathrm{Conv}_{3\times3}([\mathrm{Pool}_{\mathrm{Avg}}(\mathrm{F}); \mathrm{Pool}_{\mathrm{Max}}(\mathrm{F})])). \tag{14}$$

The average-pooled and max-pooled features across channels are concatenated and convolved by convolution layers, producing a 2D spatial attention map.

The channel and spatial modules are placed behind each residual block in a sequential manner. The CBANet is formed of such four residual blocks that are embedded with one channel module and one spatial module. Besides, the network configuration of residual blocks is consistent with the official article [24]. Specifically,as a user-specific identification network, the last layer of the network is a linear layer with two output units, connected after an average pooling layer with a kernel size of 7 and 512 channels. The network is optimized by Adam [31] with an initial learning rate of 0.01.

5 THEORETICAL ANALYSIS

Section 2 clarifies the underlying correlation between the human voice and mmWave signal. Based on their correlation, we design a multi-modal identification system named WavoID to carry on identification tasks in Section 4. In the following, we explain why the multi-modal system outperforms unimodal systems from a statistical point of view [16, 74].

The general biometric identification problem can be simplified to predict the probability of the user's identity based on intrinsic characteristics [34]. For example, the audio-only identification system extracts the feature from the unlabeled utterance token. It then predicts the probability of the unlabeled utterance belonging to one of the registered identifications. Multi-modal identification systems also have the same idea but upon multiple features. We assume that X^V and X^M are voice feature vector and mmWave feature vector, respectively. An identification system is designed to recognize enrolled users $I_n, n = 1, 2, \cdots, N$. The system would partition the input feature space into N disjoint decision regions $R_n, n = 1, 2, \cdots, N$. The probability of correct identification P for the joint feature vector X^V and X^M is defined as:

$$\begin{aligned} \mathrm{P} &= \sum_{n=1}^{N} P(X^V \in R_n^V, X^M \in R_n^M, I_n) \\ &= \sum_{n=1}^{N} \iint \rho(x^m|I_n, x^v)\rho(x_n^v|I_n)dx^v dx^m P(I_n), \end{aligned} \tag{15}$$

where R_i^V and R_j^M are the partitioned decision rejoins in the voice feature space and mmWave feature space, respectively. They determine that the input data belongs to the n-th person index. The $P(X^V \in R_n^V, X^M \in R_n^M|I_n)$ is the conditional probability of correct recognition for the n-th class and $P(I_n)$ is the prior probability of the n-th enrolled user. Since the conditional probability must not be less than the probability without conditional knowledge, we have the lower bound:

$$\begin{aligned} \mathrm{P} &\geq \sum_{n=1}^{N} \int_{R_n^M} \rho(x^m|I_n)dx^m \int_{R_n^V} \rho(x_n^v|I_n)dx^v P(I_n) \\ &= \sum_{n=1}^{N} P_{X_n^V} P_{X_n^M} P(I_n) \end{aligned} \tag{16}$$

where $P_{X_n^V}$ and $P_{X_n^M}$ individually denote the correct recognition probabilities of the voice input and the mmWave input for n-th person identity index. Since any probability is upper bounded by one, we calculate the corresponding upper bound of P:

$$\mathrm{P} \leq \sum_{n=1}^{N} \max[P^*_{X_n^V}, P^*_{X_n^M}]P(I_n). \tag{17}$$

The superscript $*$ means the maximum probability. The maximum probability will increase when prior correlated knowledge from a correlated mmWave or voice domain is offered into another domain.

Summary. Based on the above statistical analysis, the lower bound can be reached if the multiple modalities are completely independent. when the information in multi-modal modalities has a strong correlation, the multi-modal identification system performs at its upper bound. Accordingly, the multi-modal system comprising mmWave and voice features is not destined to reach the lower bound whereby the inherent correlation between the mmWave and voice. As for the upper bound, a mmWave-voice system correlated to vocal vibration tends to keep user identification ability at utmost probability. This property guarantees that the mmWave-voice system has higher accuracy in adverse scenes, or even in larger user data libraries than unimodal systems. Moreover, each feature space has an enhanced prediction ability for identification when integrating supplements from complementary modalities.

6 EVALUATION

Dataset. In the experiment, we recruit 100 participants (67 males and 33 females) aging from 16 to 47 (mean=31, sd=7.95) to speak commands in natural speech speed and volume. The command

corpus includes ok-google.io [19] and Google speech commands [72] which are commonly used to interact with voice-controlled machines. Figure 2 shows the experimental setup. All participants are asked to utter 40 speech commands in a laboratory environment when facing the setup of WavoID. Each command is required to be said 40 times, each time about 0.8 seconds long. Thus, we collect the corresponding 160,000 pairs of mmWave and voice signals. Each pair of signals are framed into 4000 sampled points per segment, equivalent to 0.08 seconds per piece of data. In total, there are about 1500,000 pieces of samples in evaluation after removing invalid data. In evaluation, we will divide all the collected data into two parts based on the specific time node. One part before the time node is regarded as training data, and the other part after the time node is regarded as test data. The order of the data in the training set and test set will be disordered, respectively, but the time of the training set will not be earlier than the time of the test set. During the procedure of data collection, the radar and microphone are placed at a distance of 50cm from the participant, directly facing the user. For each user identification, we divide 300 pieces of samples into the enrollment dataset and the rest into the verification dataset. The experiment conditions involve noise, motion, location, clothing, and realistic scenes. Note that all participants are informed and approved of the purpose of our experiments. Our research is approved by the IRB: anonymous university.

Metrics. We adopt following metrics to evaluate WavoID's performance: False positive rate (FPR): it is the probability where the system fallaciously accepts the spoofer. False rejection rate (FRR): it is the probability where the system refuses access to the user. True positive rate (TPR): it is the probability in where the system correctly identifies the user. True negative rate (TNR): it is the probability in where the system correctly rejects the spoofer. Receiver operating characteristic (ROC) curve: it represents the relationship between FPR and TPR under different identification threshold. Equal error rate (EER): the rate in which FPR equals FRR. Balanced accuracy (BAC): BAC is a better metric to use with imbalanced data, which can be calculated from $BAC = (TPR + TNR)/2$. F1-score: it is defined as the mean of positive predictive value and TPR.

6.1 Performance Analysis

Performance on large-scale dataset. We first evaluate the overall performance of WavoID across all 100 subjects. For user identification, we train the user modal based on the enrollment dataset (i.e., 80% of the collected dataset), and the rest samples are considered as positive samples in the verification. Meanwhile, others act as imposters to log into the trained user model. With the increasing number of enrolled users, WavoID shows solid performance on user identification with high accuracy above 98% and averaging EERs of 1.24%, as shown in Figure 7. It means that WavoID has a significant identification ability by fusing information from different modalities, as proved in Section 5. We show the averaging ROC curve and confusion matrix from ten random users in datasets. The results reveal that the system is highly effective in distinguishing users regardless of the size of datasets. We also study how many required enrolled samples are needed in a large-scale dataset. According to Figure 7(a), WavoID achieves BAC larger than 99% with only 300 training samples, corresponding to 25 seconds of collecting data. The data collection is equivalent to the fingerprint registration time, meaning the high user-friendliness.

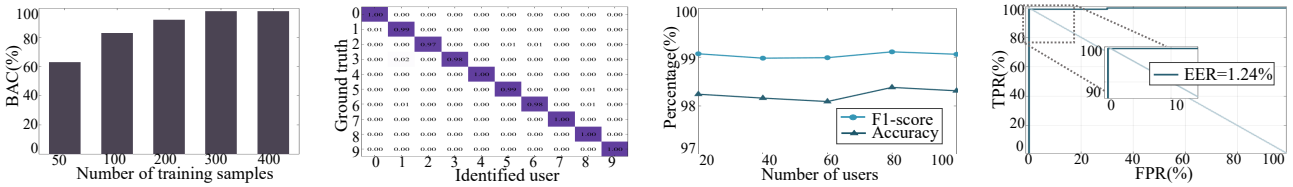
Performance on ambient noise. Since the identification system has to face noisy scenes, it is required to evaluate its robustness to noise. We place two loudspeakers at a distance of 70 cm away from the microphone. Two loudspeakers play common noise with 70 dB sound pressure level (SPL), e.g., speech, traffic, music, and white noise. Forty of the 100 participants are required to speak ten commands ten times under four kinds of ambient noise, respectively. The corresponding collected data is input to test WavoID. Note that the system is pre-trained with 300 pieces of the dataset in Section 6.1. The results shown in Figure 8 represent the performance of WavoID under noise interference. The average BACs under these four noise conditions are 97.9%, 98.3%, 98.0%, and 99.0%, respectively. Compared with baseline in a relatively quiet environment, the performance of WavoID seems to be unaffected by ambient noise. It is speculated that valid fusions of bimodal modality contributes to the identification ability of the system.

Performance under motion interference. Users' body motion could bring multipath interference to the radar, and further impact the performance of WavoID. To quantify the impact, we request 20 participants to speak ten commands ten times under different body movements, i.e., sitting down, standing up, walking, and telephoning, respectively. The testing data is fed into the system pre-trained 300 pieces of the dataset in Section 6.1. As depicted in Figure 9, WavoID steadily maintains high significant identification with BACs above 97% regardless of the user's actions. Though the motion will cause interference in the collected mmWave signal, acoustic information from the voice modality can compensate for missing features in mmWave.

Performance comparison. To further prove the superiority of the proposed system, we carry out a comparison study to quantify the role of each component in the proposed system. Moreover, we introduce a voice-only method and a mmWave-only method as baselines for comparison:

- **VGGVox (voice-only)** [14], is one of the mainstream speaker recognition systems, commonly applied in academic and industrial areas.
- **VoicePop (voice-only)** [71], a voice-only identification system, captures spectrographic features of frequency components in voice signals.
- **VocalPrint (mmWave-only)** [36], a mmWave-only identification system, extracts vocal fold features when mmWave perceives voice activity.
- **W/O Fine-grained Waveform Estimation**, where no proposed fine-grained waveform estimation module is performed. The signals of two modalities are directly fed into subsequent operations for identification.
- **W/O Cross-modality Fusion**, where no proposed cross-modality fusion module is performed. The spectrum of two modalities are concatenated and fed into the network for identification.

All of the above models are fairly and fully pre-trained on identical datasets, and then verified on the same testing samples. Specifically, we choose two dataset collected from 100 people under the same noise conditions and motion interference. The performance results of the above different methods are shown in Table 1. In Noise+Motion Condition, 40 participants speak and sit



(a) The BAC of different training size.

(b) The confusion matrix of randomly selected users.

(c) The performance on the different size of datasets.

(d) The ROC curve of large-scale datasets.

Figure 7: Performance on large scale datasets.

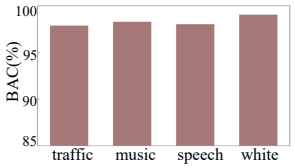


Figure 8: Performance under noise.

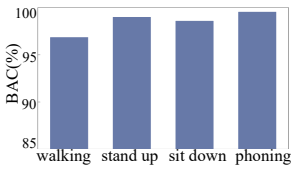


Figure 9: Performance under motion.

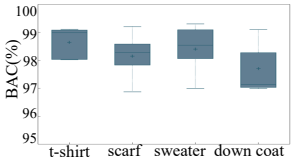


Figure 10: Impact of wearable accessories.

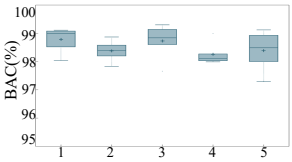


Figure 11: Impact of people around.

down/stand up/walk/ in noisy environments where two loudspeakers play speech noise. The number of training sets and testing sets are 500 and 10,000, respectively.

Table 1: Performance comparison among identification methods under different conditions. (No.+Mo.: Noise+Motion; FWE: Fine-grained Waveform Estimation; CF: Cross-modality Fusion; W/O: Without)

Method	Noise(%)		Motion(%)		No.+Mo.(%)	
	BAC	F1	BAC	F1	BAC	F1
VGGVox [14]	64.1	59.0	76.5	72.9	53.6	55.2
VoicePop [71]	66.7	77.5	87.7	93.1	58.7	59.7
VocalPrint [36]	96.9	98.3	63.3	76.3	70.2	71.8
W/O FWE	73.5	82.7	68.7	78.1	61.4	62.1
W/O CF	83.4	89.9	76.0	84.6	66.8	73.8
WavoID	99.0	99.4	99.2	99.5	99.3	99.0

According to Table 1, the voice-only method is worse than the mmWave-only and mmWave-voice methods with low accuracy in terms of distinguishing noise-polluted samples. Concerning the performance under motion interference, WavoID outperforms unimodal system with 99.21% BAC and 99.57% F1-score, which exceeds VGGVox by 22.65 %in BAC, VoicePop by 11.47% in BAC and VocalPrint by 23.26% in F1-score. This experimental comparison reveals that unimodal methods like VocalPrint will suffer from performance degradation when facing strong motion disturbances or complex conditions, especially combinations of noise and motion interference. Note VocalPrint, VoicePop, and VGGVox trained with 300 samples, the same as WavoID, we speculate that WavoID as a multimodal system owns a notable feature presentation and extraction capability within limited training samples. This phenomenon further confirms that a unimodal feature is less distinguishable than a multi-modal feature regarding identifying people, particularly in complex conditions. By comparing W/O FWE, W/O CF, and WavoID, we can find significant improvement on both noisy and motion datasets. This is because that noise and motion interference can be regarded as additive noise, while our waveform estimation and correlation filter-based fusion rely on correlation algorithms to calculate and select sub-signals among diverse signals that own maximum relevance between mmWave and voice modality, i.e., voice activity. Thus, these non-relevant additive noises will decay with iteration optimization. Moreover, these information losses rarely happen in both two modalities at the same time. In most cases, the system can compensate for the loss in one modality by using relevant features in another modality through cross-modality fusion algorithms.

7 ROBUSTNESS ANALYSIS

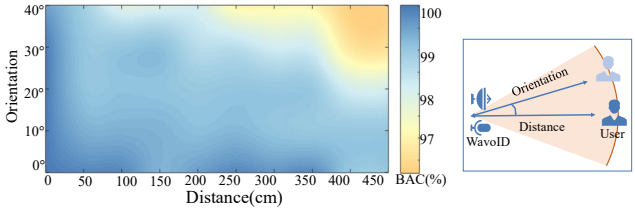
In this section, we further analyze the robustness of WavoID under the impact of wearable accessories, people around, distance and orientation. Ten of the 100 participants are asked to repeat ten commands ten times under different conditions. Such collected data is fed into the user-specific system pre-trained by 300 pieces of datasets in Section 6.1, to test the performance.

Impact of wearable accessories. Since wearable accessories in users influence the radar's perception, we further evaluate the performance of WavoID when users in different wearable accessories like t-shirts, scarfs, sweaters, and down coats. We require users to wear different clothing and speak commands at the same time. As shown in Figure 10, WavoID achieves an averaging 98.2% BAC, while BACs in down coats are slightly low but are mostly larger than 97%. We speculate that in the proposed fusion method, the characteristic of voice modality can make up for the mmWave information loss caused by clothes-covering. Generally, wearable accessories have a minimal impact on the capability of the system. The results prove that WavoID can execute identification tasks in spite of wearings.

Impact of people around. It is common to see that the user is surrounded by several passersby who utter a word and move arbitrarily. However, as WavoID exploits correlated features unique to the user, WavoID ought to identify the enrolled user adequately. To investigate it, we ask different numbers of volunteers to walk and talk freely around the target user during the identification. Identification results in Figure 11 show that all BACs fluctuate between 97% and 99%, even when the number of people around increases to 5.

We envision that the proposed cross-modality mechanism enables the system to obtain the informative feature and distinguish the genuine user.

Impact of distance and orientation. We study the effect of distance and orientation between the user and sensors, i.e., the radar and microphone. We place sensors 0.5 to 4.5 m with a 0.5 m step away from the user and 10 to 40° with a 10° step to the user. Figure 12 displays the distribution of BACs for sensors placement. It is observed that BACs still reach over 97% when the distance is less than 3 m. As the distance reaches 4.5 m, the identification ability of WavoID is slightly reduced by approximately 2%. In particular, within 30°, WavoID can still achieve approximately 99% as sensing distance expands to 3 m. The identification performance illustrates that the fusion of these two modalities guarantees the system in a large field to identify users accurately.



(a) The distribution of BAC.

(b) Illustration of location.

Figure 12: The performance of WavoID at different orientations and distances.

8 SECURITY ANALYSIS

In this section, we conduct a security analysis to explore the anti-attacking ability of WavoID. Although a onefold attack targeting voice or mmWave modality is likely to fail on multi-modal systems, due to its missing information from other modality, it is unsure what effects they cause on WavoID when they simultaneously attack WavoID. Malicious attacks for voice identification can be classified into three categories: mimicry attack, replay attack, and adversarial attack. Likewise, attacks for mmWave identification are mainly classified into two types: counterfeit attack and jamming attack.

- Replay Attack. The adversary pre-records the utterance from the target and then replays the recorded utterance through loudspeakers to speaker recognition systems. Numerous research has pointed out that the replay attack can effectively spoof most speaker recognition system [77].
- Mimicry Attack. The adversary imitates the target's pronunciation to utter passphrases without helps of any devices. Existing study indicates that mimicry attacks are less effective in spoofing modern speaker recognition systems compared to replay and adversarial attacks [76]. Nevertheless, it is necessary to conduct impersonation trials with the aim of comprehensive security evaluation.
- Adversarial Attack. The latest research results manifest that the adversarial attack has become the inherent threat to the security of state-of-the-art speaker recognition systems. The computation core of speaker recognition systems is machine learning models, including state-of-the-art neural networks. Consequently, the adversary utilizes adversarial training to craft adversarial samples that are embedded with imperceptible perturbation. The adversary launches the generated adversarial sample over-the-air to deceive the trained model in spite of the black-box setting.
- Counterfeit Attack. The adversary knows that the characteristics of vocal vibration are fused into the feature template in the system. Hence, the adversary detailedly observes and stimulates the vocal vibration from the target through professional audio transducers and throat bionic models, thus, deceiving the radar in WavoID.
- Jamming Attack. The adversary continuously transmits jamming signals with high power by utilizing a mmWave radar, bringing about the radar failing to receive legitimate reflective signals.

To investigate the security of the proposed system against multi-modal attacks, we set up several combinations of attack modes mentioned above to compare the attack detection rate and identification results. Those multi-modal attack combinations are shown in Table 2.

Table 2: The combinations of attacks. (Rep: Replay Attack; Mim: Mimicry Attack; Adv: Adversarial Attack; Cou: Counterfeit Attack; Jam: Jamming Attack.)

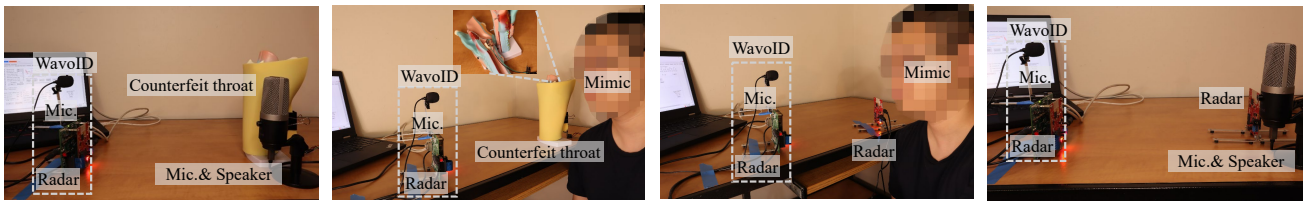
Multi-modal Attack	Multi-modal Attack
Cou + Rep	Jam + Rep
Cou + Mim	Jam + Mim
Cou + Adv	Jam + Adv

8.1 Attack Setup

We analyze the resilience of our system to six types of multi-modal attacks listed in Table 2. The experimental setup for each attack is the same in each combination. For ease of illustration, we describe the experimental setup for each attack rather than attack combinations. The setup for each attack is shown as follows (Figure 13). The data of five participants with three males and two females (mean=24.4, sd=2.07) as identification users is to pre-train user-specific models that will be attacked by the following attacks. The training number of identification is still 300 pieces. Once combinations of mmWave and acoustic attacks transmit malicious signals to the setup of WavoID, the system will receive corresponding malicious mmWave-voice signals that will be processed and examined by the algorithm of WavoID.

Replay Attack Setup. It is easy to deploy replay attacks only by using a loudspeaker and microphone. The adversary firstly employs a professional recorder to pre-record five utterances from users. Then, the adversary replays the corresponding recording 20 trials for each command through a professional loudspeaker.

Mimicry Attack Setup. The adversary imitates the legitimate user's voice and pronunciation habits, after completely watching and listening to the speaking process of users. To execute the mimicry attack, we recruit 10 participants (mean= 25.3, sd=2.26, female=4, male=6) to observe the users' corresponding recordings and then imitate five users for 20 trials for each five command. During the attack, all participants are facing the system in the same posture as the user.



(a) counterfeit attack and replay/adversarial attack. (b) counterfeit attack and mimicry attack. (c) jamming attack and mimicry attack. (d) jamming attack and replay/adversarial attack.

Figure 13: The setup of different attack scenes. The counterfeited throat, a mmWave radar, a professional microphone, and subjects are put to attack WavoID. The microphone records the genuine voice of users in replay attacks and then is connected with a loudspeaker to replay. Note that replay attack and adversarial attack share the same device but have different malicious acoustic sources.

Adversarial Attack Setup. The adversary crafts adversarial samples, resorting to state-of-art adversarial algorithms, i.e., FAKEBOB [12], Carlini-Wagner [11], FGSM [18], and PGD [42]. Assuming that the adversary has complete knowledge of the network of WavoID, the adversary performs white-box attacks to force the model to classify adversarial samples as a arbitrary enrolled speaker. For the sake of clarity, we only illustrated the training and attack process of FAKEBOB, whereby other three attack methods have the same attacking preparation. One widely used dataset: LibriSpeech [46], is regarded as training sets and testing sets. Concretely, the train-clean-100 in LibriSpeech containing 28539 utterances from 251 speakers (female=125, male=126) is used to train the speaker recognition model, i.e., the recognition network in WavoID. The dev-clean containing 2703 utterances from 40 speakers (female=20, male=20) is chosen as original waves that will be added with perturbation during the iterations of adversarial optimizations. For every attack trial, the adversary will replay 20 pieces of crafted audio samples at a distance of 30cm away from the victim. In all, the adversary carries on five trials with the sum of 100 attacks per adversarial algorithms.

Counterfeit Attack Setup. The adversary has known that the system also leverages vocal vibration features modulated on mmWave signals. Thus, the adversary elaborates a bionic throat model and then puts an audio transducer into the throat. The eavesdropped voice stimulates the transducer so that the transducer vibrates the throat to counterfeit trails of vocal organs. The internal structure of throat model is shown in Figure 15(b). The throat model is placed at a distance of 30cm away from the system to attack the radar.

Jamming Attack Setup. Given that the adversary has access to the configuration of the victim radar, the adversary could use a similar mmWave radar to transmit high-energy FMCW signals with the same parameters as the victim. Moreover, The adversary's mmWave radar is directly placed 50 cm away from the victim. The victim's radar transmission power is 12.5dBm. Meanwhile, the attack radar transmission power is set as 15dBm.

8.2 Detection Results

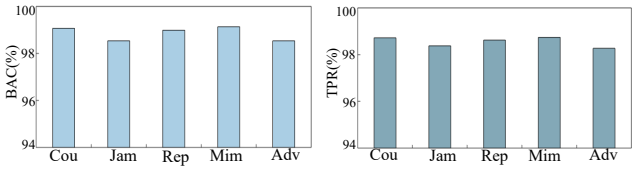
The defense results of our proposed mmWave-voice method, i.e., bimodal liveness detection, are reported in Table 3. The detection accuracy is close to 100% for all multi-modal attacks, which means that our bimodal liveness detection method is reliable against multifarious attacks. Comparing the detection result of jamming attacks and counterfeit attacks, we can observe that detecting jamming attacks is slightly more difficult than detecting counterfeit attacks, but still keeps 99% rejection rates. We can further speculate that the injected mmWave signals through jamming attacks tend to provoke abnormal outliers, interpreting the detection performance. To better defend against jamming attacks, WavoID is encouraged to randomly alter chirp periods and frequency bandwidths by using frequency hopping. Overall detection results present that all kinds of multimodal attacks almost can not bypass our bimodal liveness detection. This is expected given that our proposed mmWave-voice approach can capture biometric information different from machine-caused or non-user imitated signals.

Table 3: The detection result of WavoID under various multimodal attacks. (Rep: Replay Attack; Mim: Mimicry Attack; Adv: Adversarial Attack; Cou: Counterfeit Attack; Jam: Jamming Attack.)

Attack	Accuracy(%)	EER(%)	FPR(%)
Cou+Rep	99.11	0.74	0.72
Cou+Mim	99.21	0.83	0.79
Cou+Adv	99.25	0.79	0.76
Jam+Rep	99.09	0.91	0.85
Jam+Mim	99.19	0.81	0.78
Jam+Adv	99.32	0.89	0.81

We also add an experiment to evaluate the impact of the unimodal attack on the identification accuracy of the system. In this case, we consider a situation where attackers bypass the liveness detection to directly inject their malicious signals into our identification framework. Thus, we feed malicious signals produced by the attacks mentioned earlier into our proposed system without a liveness detection module. Figure 14 illustrates the BAC and TPR of WavoID on each attack in this case. It is observed that the overall BAC is above 98.5%, with no degradation. In summary, WavoID is robust to various spoofing attacks.

Insight. With the trend of multi-modal authentication, multi-modal attacks emerge as new threats. For example, a presentation attack can be easily deployed to hinder audio-visual identification by using a low-cost electronic display device[51]. Some multi-modal defenses experience a degradation in performance when they are exposed to multi-modal attacks [13]. It is critical to design a defensive system to defend against these multi-modal attacks [38]. Those acoustic motion-based identification has limited sensing distance, low range measurement, and hardly penetrate barrier, because the utilized ultrasound has high attenuation, low bandwidth, and low penetration. Motion-based methods cannot support fine-grained voice activity detection, especially under adverse conditions. WavoID can detect millimeter-scale mouth and throat vibrations regardless of long distance and wearable accessories, which can compensate for defects in speech modality and achieve robust identification.



(a) The BAC of WavoID. (b) The TPR of WavoID.

Figure 14: The performance of WavoID under unimodal attack.

9 SYSTEM EFFICIENCY

As a biometric system, it is necessary to examine the system effiency of WavoID. In the last experiment, we investigate the runtime latency of WavoID on Linux servers equipped with an NVIDIA 2080 RTX. The deep learning-based system is implemented on Pytorch. The runtime latency is defined as the interval from the beginning time when the system collects signals to the ending time when the system outputs the result. The overall latency includes the time for signals collection, signals processing, liveness detection, and identification. We execute the entire workflow 20 times to calculate the runtime. The average runtime is 2.175s. Considering the significant computing ability of cloud servers, such a time overhead is acceptable for human-machine interactions.

10 DISCUSSION

Software Overhead. Considering the practicability of migrating WavoID to mobile devices, e.g., smart speaker, the occupied size of the system software should be as small as possible. The system software of WavoID is merely 43MB, which is suitable for edge devices and cloud platforms. Furthermore, we can use pruning and quantization techniques [22, 88] to compress the model size. The proposed system has the potential to be applied to different edge devices for ubiquitous real-time identification through online or offline pipelines, relying on the technology of edge computing [55] and distributed computation [81].

Hardware Support. To enable the deployment of WavoID, we need a commercial mmWave radar and a microphone, which summing cost is less than 40 dollars. This cost overhead is acceptable for security and assurance systems that focus more on precision and security, concerning its properties of anti-spoofing and high accuracy under complex circumstances. In terms of power consumption, the sum of energy power is below 20mW, relatively lower than WIFI [17]. The radar of system has a range coverage of 15 meters with 120° field-of-view. It can cope with most practical applications where users face sensors within permissible deviation. Besides, it is possible to employ three mmWave radars into WavoID for fully open space with 360°coverage. Although it is hard to deploy mmWave radar into mobile devices due to its cost and complexity, considering the fact that indoor mmWave-based fall detection devices[23, 68] exist in homes, hospitals, firms, etc., it is feasible to combine the audio collection of smart speakers and mmWave sensing of detection devices for user identification in the future deployment. In many cases, the fall detector is mounted on top of the ceiling to get the widest view of the detection, which also facilitates cooperation with smart speakers for real-time indoor authentication.

Application Scenarios. The first application scenario of WavoID is to protect voice-controlled devices from malicious injection attacks, which aims to verify the legitimate identity of received voice commands. For example, when a user speaks voice commands to his smartphones or smart speakers, the microphone and radar of devices simultaneously receive the signal related to voice activities. The built-in system first determines whether the signal source is from genuine users, and then recognizes and executes commands. Another interesting application scenario we envision is to integrate the vehicle-borne radar [58] and microphone to enable car authentication and straightforward manipulation, which requires long-distance sensing from mmWave modality and complementary acoustic information from voice modality to resist user motion interference.

11 CONCLUSION

In this paper, we propose a mmWave-voice identification system called WavoID for efficiently identifying users in a robust and secure way. This paper provides a theoretical analysis of correlated multi-modal identification systems beneficial to accurately predicting users' identities. WavoID aggregates biometric information derived from mmWave and voice modality to guarantee the security of identification systems. To thoroughly characterize the unique feature and efficiently fuse individual modalities, WavoID utilizes response maps to measure the quality of each extracted component. We evaluate WavoID among 100 users and thoroughly test its defense against multi-modal attacks. The result shows that WavoID maintains over 98% identification accuracy and prevents about 99% of malicious attacks.

ACKNOWLEDGMENTS

This paper is partially supported by the National Key R&D Program of China (2020AAA0107700), National Natural Science Foundation of China (62032021, 61772236, 61972348).

REFERENCES

[1] Anter Abozaid, Ayman Haggag, Hany Kasban, and Mostafa Eltokhy. 2019. Multimodal biometric scheme for human authentication technique based on voice and face recognition fusion. *Multimedia Tools and Applications* 78, 12 (2019), 16345–16361.

[2] Ilia Adami, Margherita Antona, and Emmanouil G Spanakis. 2017. Multi-modal user interface design for a face and voice recognition biometric authentication system. In *International Conference on Wireless Mobile Communication and Healthcare*.

[3] Muhammad Ejaz Ahmed, Il-Youp Kwak, Jun Ho Huh, Iljoo Kim, Taekkyung Oh, and Hyoungshick Kim. 2020. Void: A fast and light voice liveness detection system. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*. 2685–2702.

[4] Mohammed Aladsani, Ahmed Alkhateeb, and Georgios C Trichopoulos. 2019. Leveraging mmWave imaging and communications for simultaneous localization and mapping. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4539–4543.

[5] SI Alekseev, AA Radzievsky, MK Logani, and MC Ziskin. 2008. Millimeter wave dosimetry of human skin. *Bioelectromagnetics: Journal of the Bioelectromagnetics Society, The Society for Physical Regulation in Biology and Medicine, The European Bioelectromagnetics Association* 29, 1 (2008), 65–70.

[6] SI Alekseev and MC Ziskin. 2007. Human skin permittivity determined by millimeter wave reflection measurements. *Bioelectromagnetics: Journal of the Bioelectromagnetics Society, The Society for Physical Regulation in Biology and Medicine, The European Bioelectromagnetics Association* 28, 5 (2007), 331–339.

[7] Zhongxin Bai and Xiao-Lei Zhang. 2021. Speaker recognition based on deep learning: An overview. *Neural Networks* 140 (2021), 65–99.

[8] Stephen Boyd, Neal Parikh, and Eric Chu. 2011. *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc.

[9] Marcin D Bugdol and Andrzej W Mitas. 2014. Multimodal biometric system combining ECG and sound signals. *Pattern Recognition Letters* 38 (2014), 107–112.

[10] Joseph P Campbell. 1997. Speaker recognition: A tutorial. *Proc. IEEE* 85, 9 (1997), 1437–1462.

[11] Nicholas Carlini and David Wagner. 2018. Audio Adversarial Examples: Targeted Attacks on Speech-to-Text. In *2018 IEEE Security and Privacy Workshops (SPW)*.

[12] Guangke Chen, Sen Chenb, Lingling Fan, Xiaoning Du, Zhe Zhao, Fu Song, and Yang Liu. 2021. Who is real bob? adversarial attacks on speaker recognition systems. In *2021 IEEE Symposium on Security and Privacy (SP)*.

[13] Jinyin Chen, Chengyu Jia, Haibin Zheng, Ruoxi Chen, and Chenbo Fu. 2023. Is multi-modal necessarily better? Robustness evaluation of multi-modal fake news detection. *IEEE Transactions on Network Science and Engineering* (2023).

[14] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622* (2018).

[15] Livija Cveticanin. 2012. Review on mathematical and mechanical models of the vocal cord. *Journal of Applied Mathematics* 2012 (2012).

[16] Engin Erzin, Yücel Yemez, and A Murat Tekalp. 2005. Multimodal speaker identification using an adaptive classifier cascade based on modality reliability. *IEEE Transactions on Multimedia* 7, 5 (2005), 840–852.

[17] Karina Gomez, Roberto Riggio, Tinku Rasheed, and Fabrizio Granelli. 2011. Analysing the energy consumption behaviour of WiFi networks. In *2011 IEEE Online Conference on Green Communications*. 98–104.

[18] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[19] Google. 2021. ok-google.io. https://ok-google.io/

[20] Harsh Gupta, Ville Hautamäki, Tomi Kinnunen, and Pasi Fränti. 2005. Field evaluation of text-dependent speaker recognition in an access control application. *Speech and Image Processing Unit. Department of Computer Science, University of Joensuu, Finland* (2005).

[21] Sandeep Gupta, Attaullah Buriro, and Bruno Crispo. 2019. DriverAuth: A risk-based multi-modal biometric-based driver authentication scheme for ride-sharing platforms. *Computers & Security* 83 (2019), 122–139.

[22] Song Han, Huizi Mao, and William J Dally. 2015. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149* (2015).

[23] Haozee. 2023. mmWave human presence sensor. https://a.co/d/1zkKFL4.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[25] Daniel Hintze, Matthias Füller, Sebastian Scholz, Rainhard D Findling, Muhammad Muaaz, Philipp Kapfer, Eckhard Koch, and René Mayrhofer. 2019. Cormorant: Ubiquitous risk-aware multi-modal biometric authentication across mobile devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–23.

[26] Pengfei Hu, Yifan Ma, Panneer Selvam Santhalingam, Parth H Pathak, and Xiuzhen Cheng. 2022. Milliear: Millimeter-wave acoustic eavesdropping with unconstrained vocabulary. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. 11–20.

[27] Aapo Hyvarinen. 1999. Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks* 10, 3 (1999), 626–634.

[28] Chengkun Jiang, Junchen Guo, Yuan He, Meng Jin, Shuai Li, and Yunhao Liu. 2020. MmVib: Micrometer-Level Vibration Measurement with Mmwave Radar. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*.

[29] Hassan Khan, Aaron Atwater, and Urs Hengartner. 2014. Itus: an implicit authentication framework for android. In *Proceedings of the 20th annual international conference on Mobile computing and networking*.

[30] Dong-Su Kim and Kwang-Seok Hong. 2008. Multimodal biometric authentication using teeth image and voice in mobile environment. *IEEE Transactions on Consumer Electronics* 54, 4 (2008), 1790–1797.

[31] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[32] Tomi Kinnunen, Md Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee. 2017. The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. (2017).

[33] Hao Kong, Xiangyu Xu, Jiadi Yu, Qilin Chen, Chenguang Ma, Yingying Chen, Yi-Chao Chen, and Linghe Kong. 2022. m3track: mmwave-based multi-user 3d posture tracking. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*. 491–503.

[34] Sun Yuan Kung, Man-Wai Mak, Shang-Hung Lin, MW Mak, and S Lin. 2005. *Biometric authentication: a machine learning approach*. Prentice Hall Professional Technical Reference New York.

[35] Howard Lei and Eduardo Lopez. 2009. Mel, linear, and antimel frequency cepstral coefficients in broad phonetic regions for telephone speaker recognition. In *Tenth Annual Conference of the International Speech Communication Association*.

[36] Huining Li, Chenhan Xu, Aditya Singh Rathore, Zhengxiong Li, Hanbin Zhang, Chen Song, Kun Wang, Lu Su, Feng Lin, Kui Ren, et al. 2020. VocalPrint: exploring a resilient and secure voice authentication via mmWave biometric interrogation. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*.

[37] Tiantian Liu, Ming Gao, Feng Lin, Chao Wang, Zhongjie Ba, Jinsong Han, Wenyao Xu, and Kui Ren. 2021. Wavoice: A Noise-resistant Multi-modal Speech Recognition System Fusing mmWave and Audio Signals. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*.

[38] Tiantian Liu, Feng Lin, Zhangsen Wang, Chao Wang, Zhongjie Ba, Li Lu, Wenyao Xu, and Kui Ren. 2023. MagBackdoor: Beware of Your Loudspeaker as a Backdoor for Magnetic Injection Attacks. In *Proceedings of the 44th IEEE Symposium on Security and Privacy (SP)*. 3416–3431.

[39] Tiantian Liu, Chao Wang, Zhengxiong Li, Ming-Chun Huang, Wenyao Xu, and Feng Lin. 2023. Wavoice: A MmWave-Assisted Noise-Resistant Speech Recognition System. *ACM Transactions on Sensor Networks* (may 2023).

[40] Yixiu Liu, Yunzhou Zhang, Meiyu Hu, Pengju Si, and Chongkun Xia. 2017. Fast tracking via spatio-temporal context learning based on multi-color attributes and pca. In *2017 IEEE International Conference on Information and Automation (ICIA)*.

[41] Alan Lukezic, Tomas Vojir, Luka ˇCehovin Zajc, Jiri Matas, and Matej Kristan. 2017. Discriminative correlation filter with channel and spatial reliability. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[42] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).

[43] Yan Meng, Zichang Wang, Wei Zhang, Peilin Wu, Haojin Zhu, Xiaohui Liang, and Yao Liu. 2018. Wivo: Enhancing the security of voice control system via wireless signal in iot environment. In *Proceedings of the Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing*.

[44] Sarah Morrison-Smith, Aishat Aloba, Hangwei Lu, Brett Benda, Shaghayegh Esmaeili, Gianne Flores, Jesse Smith, Nikita Soni, Isaac Wang, Rejin Joy, Damon L. Woodard, Jaime Ruiz, and Lisa Anthony. 2020. MMGatorAuth: A Novel Multimodal Dataset for Authentication Interactions in Gesture and Voice. In *Proceedings of the 2020 International Conference on Multimodal Interaction*.

[45] Yong Niu, Yong Li, Depeng Jin, Li Su, and Athanasios V Vasilakos. 2015. A survey of millimeter wave communications (mmWave) for 5G: opportunities and challenges. *Wireless networks* 21 (2015), 2657–2676.

[46] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. 5206–5210.

[47] Ge Peng, Gang Zhou, David T. Nguyen, Xin Qi, Qing Yang, and Shuangquan Wang. 2017. Continuous Authentication With Touch Behavioral Biometrics and Voice on Wearable Glasses. *IEEE Transactions on Human-Machine Systems* 47, 3 (2017), 404–416.

[48] Swadhin Pradhan, Wei Sun, Ghufran Baig, and Lili Qiu. 2019. Combating Replay Attacks Against Voice Assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019).

[49] Swadhin Pradhan, Wei Sun, Ghufran Baig, and Lili Qiu. 2019. Combating replay attacks against voice assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–26.

[50] Krishan Rajaratnam and Jugal Kalita. 2018. Noise flooding for detecting audio adversarial examples against automatic speech recognition. In *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE, 197–201.

[51] Raghavendra Ramachandra, Martin Stokkenes, Amir Mohammadi, Sushma Venkatesh, Kiran Raja, Pankaj Wasnik, Eric Poiret, Sébastien Marcel, and Christoph Busch. 2019. Smartphone multi-modal biometric authentication: Database and evaluation. *arXiv preprint arXiv:1912.02487* (2019).

[52] Weibin Rong, Zhanjing Li, Wei Zhang, and Lining Sun. 2014. An improved CANNY edge detection algorithm. In *2014 IEEE international conference on mechatronics and automation*. 577–582.

[53] Patrick Schober, Christa Boer, and Lothar A Schwarte. 2018. Correlation coefficients: appropriate use and interpretation. *Anesthesia & Analgesia* 126, 5 (2018), 1763–1768.

[54] Ivan W Selesnick, Richard G Baraniuk, and Nick C Kingsbury. 2005. The dual-tree complex wavelet transform. *IEEE signal processing magazine* 22, 6 (2005), 123–151.

[55] Weisong Shi, Jie Cao, Quan Zhang, Youhuizi Li, and Lanyu Xu. 2016. Edge computing: Vision and challenges. *IEEE internet of things journal* (2016), 637–646.

[56] Petr Šidlof, Stefan Zörner, and Andreas Hüppe. 2013. Numerical simulation of flow-induced sound in human voice production. *Procedia Engineering* 61 (2013), 333–340.

[57] Amrita Singh and Amit M Joshi. 2020. Speaker identification through natural and whisper speech signal. In *Optical and Wireless Technologies*. Springer, 223–231.

[58] Lihao Song, Xiaoping Li, and Yanming Liu. 2021. Effect of time-varying plasma sheath on hypersonic vehicle-borne radar target detection. *IEEE Sensors Journal* 21, 15 (2021), 16880–16893.

[59] Thomas G. Stockham. 1966. High-Speed Convolution and Correlation. In *Proceedings of the April 26-28, 1966, Spring Joint Computer Conference (AFIPS '66 (Spring))*. Association for Computing Machinery, 229–233.

[60] Andrew G Stove. 1992. Linear FMCW radar techniques. In *IEE Proceedings F-Radar and Signal Processing*.

[61] David MJ Tax and Robert PW Duin. 2004. Support vector data description. *Machine learning* 54, 1 (2004), 45–66.

[62] TI. 2021. DCA1000EVM. https://www.ti.com/tool/DCA1000EVM.

[63] TI. 2021. IWR1642. https://www.ti.com/tool/IWR1642BOOST.

[64] TI. 2021. mmWave Studio. https://www.ti.com/tool/MMWAVE-STUDIO.

[65] Sreenivas Sremath Tirumala, Seyed Reza Shahamiri, Abhimanyu Singh Garhwal, and Ruili Wang. 2017. Speaker identification features extraction methods: A systematic review. *Expert Systems with Applications* 90 (2017), 250–271.

[66] Massimiliano Todisco, Héctor Delgado, and Nicholas Evans. 2017. Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Computer Speech & Language* 45 (2017), 516–535.

[67] Deepak Vasisht, Guo Zhang, Omid Abari, Hsiao-Ming Lu, Jacob Flanz, and Dina Katabi. 2018. In-body backscatter communication and localization. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*.

[68] VAYYAR. 2023. Touchless Fall Detection for The Home. https://a.co/d/5UW0HWW.

[69] Chao Wang, Feng Lin, Tiantian Liu, Ziwei Liu, Yijie Shen, Zhongjie Ba, Li Lu, Wenyao Xu, and Kui Ren. 2022. mmPhone: Acoustic Eavesdropping on Loudspeakers via mmWave-characterized Piezoelectric Effect. In *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*. 820–829.

[70] Chao Wang, Feng Lin, Tiantian Liu, Kaidi Zheng, Zhibo Wang, Zhengxiong Li, Ming-Chun Huang, Wenyao Xu, and Kui Ren. 2022. MmEve: Eavesdropping on Smartphone's Earpiece via COTS MmWave Device. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*. 338–351.

[71] Qian Wang, Xiu Lin, Man Zhou, Yanjiao Chen, Cong Wang, Qi Li, and Xiangyang Luo. 2019. Voicepop: A pop noise based anti-spoofing system for voice authentication on smartphones. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*.

[72] Pete Warden. 2018. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209* (2018).

[73] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*. 3–19.

[74] Lizhong Wu, Sharon L. Oviatt, and Philip R. Cohen. 1999. Multimodal integration-a statistical view. *IEEE Transactions on Multimedia* 1, 4 (1999), 334–341.

[75] Ting Wu, Theodore S Rappaport, and Christopher M Collins. 2015. The human body and millimeter-wave wireless communication systems: Interactions and implications. In *2015 IEEE International Conference on Communications (ICC)*.

[76] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li. 2015. Spoofing and countermeasures for speaker verification: A survey. *speech communication* 66 (2015), 130–153.

[77] Zhizheng Wu, Sheng Gao, Eng Siong Cling, and Haizhou Li. 2014. A study on replay attack and anti-spoofing for text-dependent speaker verification. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*.

[78] Chenhan Xu, Huining Li, Zhengxiong Li, Hanbin Zhang, Aditya Singh Rathore, Xingyu Chen, Kun Wang, Ming-chun Huang, and Wenyao Xu. 2021. Cardiacwave: A mmwave-based scheme of non-contact and high-definition heart activity computing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–26.

[79] Chenhan Xu, Zhengxiong Li, Hanbin Zhang, Aditya Singh Rathore, Huining Li, Chen Song, Kun Wang, and Wenyao Xu. 2019. Waveear: Exploring a mmwave-based noise-resistant speech sensing for voice-user interface. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*.

[80] Chen Yan, Yan Long, Xiaoyu Ji, and Wenyuan Xu. 2019. The catcher in the field: A fieldprint based spoofing detection for text-independent speaker verification. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*.

[81] Lichao Yang, Heli Zhang, Xi Li, Hong Ji, and Victor CM Leung. 2018. A distributed computation offloading strategy in small-cell networks integrated with mobile edge computing. *IEEE/ACM Transactions on Networking* (2018), 2762–2773.

[82] Xin Yang, Jian Liu, Yingying Chen, Xiaonan Guo, and Yucheng Xie. 2020. MU-ID: Multi-user identification through gaits using millimeter wave radios. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*.

[83] Xin Yang, Jian Liu, Yingying Chen, Xiaonan Guo, and Yucheng Xie. 2020. MU-ID: Multi-user Identification Through Gaits Using Millimeter Wave Radios. In *IEEE Conference on Computer Communications*.

[84] Kaihua Zhang, Lei Zhang, Qingshan Liu, David Zhang, and Ming Hsuan Yang. 2014. Fast visual tracking via dense spatio-temporal context learning. *Lecture Notes in Computer Science* 8693, 5 (2014), 127–141.

[85] Linghan Zhang, Sheng Tan, and Jie Yang. 2017. Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*.

[86] Xiang Zhang, Lina Yao, Salil S. Kanhere, Yunhao Liu, Tao Gu, and Kaixuan Chen. 2018. MindID: Person Identification from Brain Waves through Attention-Based Recurrent Neural Network. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018).

[87] Zhaoyan Zhang. 2016. Mechanics of human voice production and control. *The journal of the acoustical society of america* 140, 4 (2016), 2614–2635.

[88] Michael Zhu and Suyog Gupta. 2017. To prune, or not to prune: exploring the efficacy of pruning for model compression. *arXiv preprint arXiv:1710.01878* (2017).