AFace: Range-flexible Anti-spoofing Face Authentication via Smartphone Acoustic Sensing

ZHAOPENG XU, Ocean University of China, China
TONG LIU, Shanghai University, China
RUOBING JIANG, Ocean University of China, China
PENGFEI HU, Shandong University, China
ZHONGWEN GUO, Ocean University of China, China
CHAO LIU*, Ocean University of China, China

User authentication on smartphones needs to balance both security and convenience. Many image-based face authentication methods are vulnerable to spoofing and are plagued by privacy breaches, so models based on acoustic sensing have emerged to achieve reliable user authentication. However, they can only achieve reasonable performance under specific conditions (i.e., a fixed range), and they can not resist 3D printing attacks. To address these limitations, we present a novel user authentication system, referred to as AFace. The system mainly consists of two parts: an iso-depth model and a range-adaptive (RA) algorithm. The iso-depth model establishes a connection between acoustic echoes and facial structures, while taking into account the influence of biological materials on echo energy, making it resistant to 3D printing attacks (as it's difficult to replicate material information in 3D printing). RA algorithm can adaptively compensate for the distance between the user and the smartphone, enabling flexible authentication modes. Results from experiments with 40 volunteers demonstrate that AFace achieves an average accuracy of 96.9 % and an F1 score of 96.9 %, and no image/video-based attack is observed to succeed in spoofing.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**; • **Security and privacy → Privacy protections**.

Additional Key Words and Phrases: anti-spoofing, authentication, acoustic, privacy protect

ACM Reference Format:
Zhaopeng Xu, Tong Liu, Ruobing Jiang, Pengfei Hu, Zhongwen Guo, and Chao Liu. 2024. AFace: Range-flexible Anti-spoofing Face Authentication via Smartphone Acoustic Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 1, Article 26 (March 2024), 33 pages. https://doi.org/10.1145/3643510

1 INTRODUCTION

*Chao Liu is the corresponding author.
Zhaopeng Xu and Tong Liu contributed equally to this research.

Authors' addresses: Zhaopeng Xu, xzp@stu.ouc.edu.cn, Ocean University of China, Qingdao, China, 266500; Tong Liu, tong_liu@shu.edu.cn, Shanghai University, Shanghai, China, 200444; Ruobing Jiang, Ocean University of China, Qingdao, China, jrb@ouc.edu.cn; Pengfei Hu, Shandong University, Qingdao, China, phu@sdu.edu.cn; Zhongwen Guo, Ocean University of China, Qingdao, China, guozhw@ouc.edu.cn; Chao Liu, Ocean University of China, Qingdao, China, liuchao@ouc.edu.cn.

ACM 2474-9567/2024/3-ART26
https://doi.org/10.1145/3643510

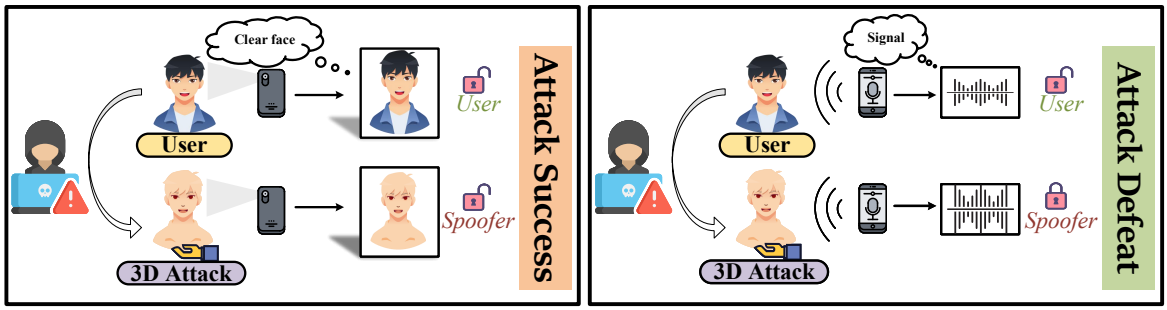


Fig. 1. 3D attack model can spoof image-based authentication, but cannot spoof signal-based authentication.

Traditional identity authentication methods such as PIN, fingerprint, and iris are gradually being replaced by facial authentication methods due to their limitations. For instance, PINs can easily be forgotten by users and are susceptible to shoulder-surfing attacks [37]. Fingerprint sensors can be compromised through the creation of fake fingerprints using fingerprint information left by the user [23, 40]. While iris sensors are highly secure, they are also expensive and not widely available on most smartphones [30]. In contrast, image face authentication technology is gaining popularity due to its convenience and exceptional performance in accurate identification.

Despite its widespread use on smartphones, face authentication technology has raised concerns due to its potential risks. The technology relies on optical cameras to capture facial features, which can lead to performance degradation if images are not captured in good lighting conditions or are unclear [1]. Additionally, face authentication is not necessarily secure, as attackers can use various methods to cheat the system, depending on the level of security [12, 13]. These attack methods can be categorized into three types: 2D attacks, where a single photo can fool traditional photo-based face authentication systems, video replay attacks, where higher security systems can be fooled by simple movements such as nodding or blinking, and 3D attacks, where face models with depth information can easily trick identity authentication systems that rely primarily on facial structure.

Several traditional face anti-spoofing methods have been proposed in the past. Many of these methods [4, 28] rely on extracting traditional features such as gradients and textures, while others [5, 15, 48] use deep learning techniques to learn the differences between real and spoofed faces. Despite some progress, there are still issues with poor generalization, which leads to a decline in performance when encountering unseen samples [29, 38]. Moreover, these models necessitate excessively deep neural network architectures, giving rise to a heightened model complexity that places substantial demands on computational resources and storage capacity. Deploying these models on mobile devices requires a significant amount of computational resources, resulting in slow system response times (average response time is around 4 seconds), making them unsuitable for real-time applications [17, 26, 36]. Lastly, the utilization of neural networks for facial recognition entails the processing of extensive personal data, thereby presenting privacy risks [33]. The latest depth cameras/dot projectors technology, such as Apple's Face ID, involves installing a dot projector and an infrared depth sensor within a small area to perceive the 3D structure of the face. However, this technology comes at an additional cost of 5 % to the bill of materials [31], making it expensive and not widely adoptable in most smartphones. Furthermore, its functioning relies on facial structural information for identity authentication, which is susceptible to deception through 3D-printed models [11]. Meanwhile, we have noticed that signal-based anti-fraud methods have been widely proposed. Wang et al. [41], Zhang et al. [50], and Xu et al. [46] use gait features for identity authentication, Xu et al. [44] employ RFID for facial recognition, and Yang et al. [47] capture user behavior patterns through WiFi signals for identity authentication. The success of these methods has inspired us to use acoustic

signals to achieve anti-3D deception identity authentication on mobile phones due to their advantages, such as low interference between frames and high resolution [3, 21, 39].

We introduce a cutting-edge face authentication method that utilizes acoustic signals, dubbed AFace[1]. It can serve as an auxiliary system for traditional identity authentication to provide higher accuracy and security, or it can work independently. Now, we only focus on its independent performance. Capturing facial 3D structure using acoustic signals inherently provides resistance against 2D attacks involving images or videos, so our primary focus is on addressing 3D deception issues. Illustrated in Fig. 1, the left side portrays the conventional visually-based identity authentication method, wherein images of a genuine user taken by a smartphone and images of a 3D printed model captured by the same device share identical structural attributes. Traditional image-based authentication relies on identifying facial features and their positional arrangement for verification, rendering it susceptible to manipulation by 3D models. Concurrently, the use of clear facial images heightens the risk of privacy breaches. Conversely, the right side illustrates the acoustically-based identity authentication method. Despite genuine users and 3D printed models displaying similar physical structures (resulting in similar echo waveform patterns), our research reveals that distinct biological materials exhibit variations in signal absorption, reflection, and related factors [2]. Consequently, energy features can be harnessed to distinguish between various biological materials, thereby enhancing resilience against 3D attacks. Furthermore, the echo information serves to safeguard users' facial privacy.

In order to achieve robust, secure, and user-friendly acoustic-based authentication, we must overcome several challenges. Firstly, acoustic signals are less intuitive compared to images and establishing a connection between the original acoustic signal and facial features can be challenging. Additionally, current signal-based authentication systems require a fixed range between the user and the signal transceiver, limiting the system's flexibility and reducing its performance when the user's position changes. To improve ease-of-use, we address these issues in two ways: (1) We propose a facial iso-depth model that connects acoustic signals to facial features, demonstrating that information related to the facial structure can be obtained from echo signals. (2) We introduce a RA algorithm that compensates for energy and phase variations to eliminate the requirement for a fixed smartphone-to-face distance.

Thorough evaluations have been carried out on the proposed AFace system, which employs a commercial off-the-shelf (COTS) speaker and microphone as its acoustic transceiver. The system exhibits exceptional face authentication performance, with an average accuracy of 96.9 % and robustness against various real-world impacts. We upload the dataset and code to GitHub[2] to provide more transparency and convenience for others to verify and reproduce our experimental results. The major contributions of AFace are as follows:

- We establish a 3D spoofing-resistant identity authentication system and introduce the first iso-depth model that links acoustic echo signals to human facial features.
- We present a unique RA algorithm to enhance the system's versatility, ensuring high accuracy at various distances between the smartphone and the user's face.
- Through extensive experiments with 40 participants, the response time of AFace for the registration and authentication are around 10 seconds and 1 second, respectively. It demonstrates an average accuracy of 96.9 %, the precision of 96.9 % and recall rate of 96.91 %, and no image/video-based attack successfully compromises the system.

The remainder of this paper is organized as follows: In Section 2, we review related work in authentication and smartphone acoustic perception. In Section 3, we provide a brief overview of the system. The details of acoustic sensing are presented in Section 4. The design of the neural network and authentication module is explained in Section 5. The performance of AFace is evaluated in different scenarios in Section 6. Next, we discussed the

[1]Demo URL: https://youtu.be/hT2mLl0do6g
[2]https://github.com/ouc-chao-liu/AFace.git

limitations of the system in Section 7. Finally, the paper analyzes future research directions and concludes in Section 8.

2 RELATED WORK

In this section, we will discuss the evolution of various authentication methods used in smartphones, including traditional methods, the growth of signal-based authentication, and acoustic-based authentication.

2.1 Traditional authentication on smartphones

Traditional smartphone authentication methods include passwords, fingerprints, face authentication, and iris.

The traditional method of using passwords, such as a personal identification number (PIN) or a text/graphical password, is the earliest and most widely used method of smartphone authentication. However, this method can be easily forgotten by users and is vulnerable to shoulder-surfing attacks [37]. To improve security, some have proposed using a combination of gestures and passwords, which has shown to be more secure [9, 34]. However, this method is still not foolproof, as wifi and acoustic signals can still potentially uncover the unlock password or pattern [49, 53].

Fingerprint authentication on smartphones provides both security and convenience due to the unique and consistent nature of fingerprints [6]. However, the fingerprint information left behind by users can be used to create fake gelatin fingerprints to deceive the system [23]. Additionally, the trend towards larger screens on smartphones is leaving less space for fingerprint sensors, and the use of phone cameras to capture fingerprints [35] solves some problems but also raises new privacy concerns [40].

Vision-based face authentication methods are vulnerable to spoofing, as images or videos can easily mimic a user's face [19, 27]. Farrukh et al. [14] try to address this issue by requiring users to take photos of their faces from different angles for 3D authentication, but this approach can still be easily fooled by 3D-printed models [12, 13]. To increase resistance to spoofing, the use of both front and rear cameras on smartphones to acquire both face and finger information, and then calculate photoplethysmograms (PPG), has been suggested as a better approach [10]. However, this approach uses dual cameras which increases overhead and creates the possibility of privacy leaks.

More advanced iris authentication technologies utilize infrared light for imaging and patterned feature point authentication in the human eye [20, 30]. Despite their improved accuracy, these sensors are still expensive and not widely available on most smartphones.

2.2 Signal-based authentication

The signal-based authentication approach involves utilizing signals to recognize an individual through sensing various features such as the structure of the gait information and biological behaviors.

The general idea behind gait-based authentication methods [41, 46, 50] is to use sensors to monitor the spatial changes caused by a user's movement and extract their physical characteristics and gait cycle features from the reflected signals. These features are then used to train a neural network for authentication purposes. However, there are two main drawbacks to this approach. Firstly, the features extracted from the user's gait are often unstable and can be affected by factors such as their clothing and walking speed or direction. Secondly, the channel state information is not intuitive, which makes it difficult to label data and requires a significant amount of efforts.

There are various authentication methods based on biological habits. RFID is employed to extract user heartbeat characteristics for identity verification [25]. The concept of motion effect and reflection effect is introduced, where the motion effect pertains to the periodic movement of the RFID tag caused by the user's chest motion during breathing, and the reflection effect involves variations in signal multipath reflection due to the user's heart

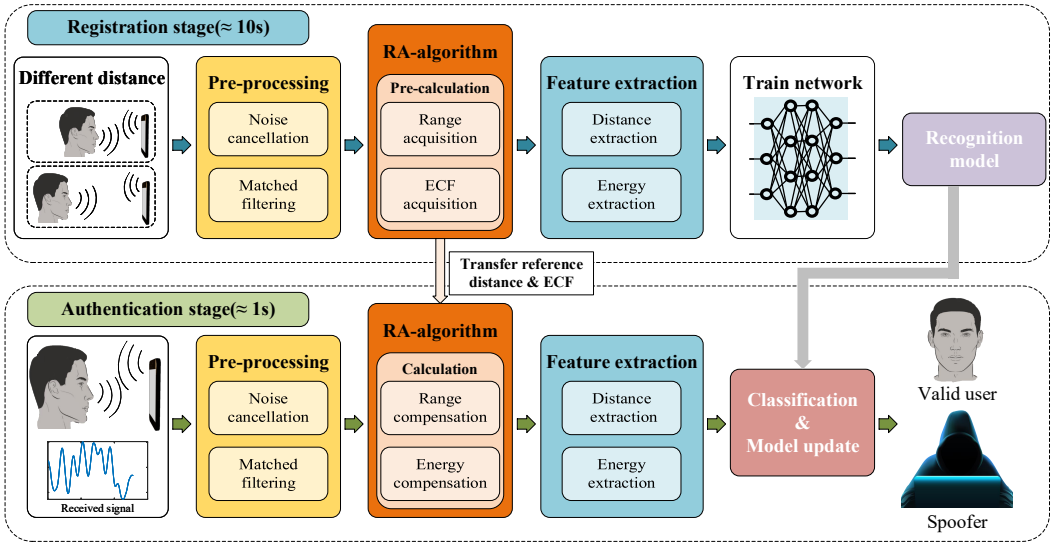


Fig. 2. System Overview.

activity. However, this method necessitates users to wear specialized tags, which adds to the user's burden. Yang et al. [47] present an authentication method that utilizes Wi-Fi signals to detect users' habitual gestures during input, irrespective of the specific input content, focusing on capturing user behavioral patterns. This approach requires the use of two additional NIC 5300 network cards as transmitting and receiving devices, thereby limiting its applicability and increasing costs.

2.3 Acoustic-based authentication on smartphones

The acoustic-based authentication approach involves utilizing signals to recognize an individual through sensing various features such as the structure of the ear canal and face structure.

Wang et al. [42, 43] propose a authentication method that captures changes in the ear canal caused by the user speaking specific words using in-ear devices, and they propose to combine both static and dynamic information from the ear canal for improved reliability in authentication. However, these approaches require the user to wear an in-ear device for an extended period, which is not user-friendly. Chauhan et al. [7] discover that the microphone on the in-ear device can detect the user's breathing sounds, which can change significantly during intense physical activity.

The face authentication methods based on signals capture facial echoes using acoustic signals, leveraging their slow wave speed. Studies have shown that ultrasonic waves emitted through sonar can be used to classify faces, but this method requires specially-made equipment [24]. Researchers have found that the microphones and speakers that come with smartphones can replace sonar for simple and reliable authentication [8, 18, 45, 51, 52]. However, these methods rely on recognizing the 3D structure of the user's face, which can be easily fooled by a 3D printed model.

However, our approach differs from these in that we aim to achieve resistance to 3D spoofing by extracting the structural and biomaterial features of the user's face from the acoustic signal.

3 SYSTEM OVERVIEW

Our proposed system utilizes acoustic sensing technology, comprising speaker&microphone, to capture facial depth and energy information from the echo signal. This data is subsequently analyzed using deep learning techniques to differentiate between users based on extracted acoustic signal features. Fig. 2 illustrates the two stages of our system: registration and authentication.

Registration Stage (≈ 10s). The AFace process is comprised of four stages. Firstly, users employ a smartphone to backscatter acoustic waves from their facial features at two distinct distances in order to acquire energy measurements. By studying the correlation between distance and energy, we derive a function capable of computing energy data for arbitrary distances. This function is referred to as the ECF (Energy Compensation Function). Next, the received echoes undergo preprocessing, which includes applying a noise cancellation algorithm to eliminate background reflections and direct signals, and match filtering to obtain distance dimension information. The third stage involves precomputing the RA (Range Adaptive) algorithm to determine the echoes' reference range and ECF. The final step involves feature extraction and training of the classifier.

Recognition Stage (≈ 1s). In the authentication stage, the user positions their smartphone at a convenient range and perceives their face with a pre-modulated signal emitted from the speaker. The microphone captures the reflected signal, which then undergoes the same preprocessing steps as in the registration stage, including applying a noise cancellation algorithm and match filtering. The system utilizes the reference range and ECF obtained in the registration stage to compensate for the echo signal using the RA algorithm. Finally, the extracted features are compared to the trained classifier to determine the user's identity. If the user is recognized as a legitimate user, the model is updated accordingly.

4 ACOUSTIC SENSING

Next, we introduce the details of our system, including signal model, signal preprocessing, RA algorithm, and feature extraction.

4.1 Signal Model

In this subsection, the send and receive signals are modeled to facilitate the subsequent processing of the signals.

4.1.1 Signal Design. Our signal design takes the following factors into account: 1) The signal's frequency range should be distinguishable from everyday background noise (typically below 8 KHz) and be compatible with most commercial smartphones. 2) The signal's bandwidth needs to be sufficiently wide to achieve high distance resolution and detect subtle variations in the user's facial features. 3) The signal's duration and volume should be kept to a minimum to prevent interference with the user.

Taking into account the aforementioned criteria, we select the FMCW signal due to its well-established use in ranging and high resilience to multipath effects. The signal frequency ranges from 10 KHz to 22 KHz, making it easily distinguishable from ambient noise by employing a high-pass filter. Although most smartphones can operate at frequencies up to 24 KHz, we observe significant signal attenuation beyond 20 KHz, prompting us to set our signal's upper limit at 22 KHz. The 12 KHz bandwidth provides a resolution of 1.4 cm, enough to distinguish subtle differences in the face. The signal duration is 10 ms, and a Hanning window is applied to concentrate the signal energy at around 16 KHz while preventing spectrum leakage. In our experiments, the volume of the phone is set to 70 % of the maximum volume, users can barely hear any sound, and we add a 40 ms interval between consecutive signals to prevent the echoes from overlapping.

As mentioned above, we modulate the signal in the following form:

$$s(\hat{t}, t_m) = \text{rect}(\frac{\hat{t}}{T_p}) e^{j2\pi(f_c t + \frac{1}{2}k\hat{t}^2)}, \tag{1}$$

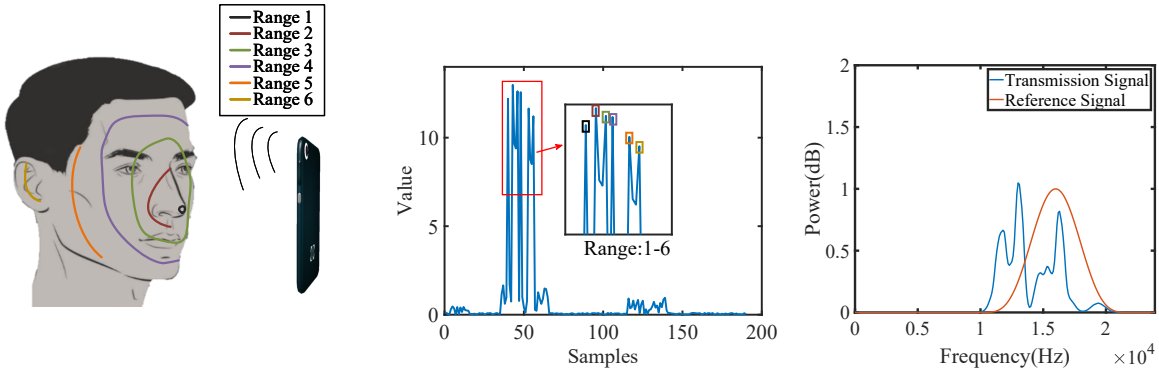


Fig. 3. (a) is the user's facial range model, which is roughly divided into six regions, (b) is the correspondence between the processed user face echo and the model, (c) is the frequency of transmission signal and reference signal.

where $\text{rect}(u) = \begin{cases} 1 & |u| \le \frac{1}{2} \\ 0 & |u| > \frac{1}{2} \end{cases}$, t is the time from the send of the signal, f_c is the carrier frequency, k is the modulation frequency, $\hat{t}$ and t_m are fast time and slow time, respectively, and T_p is the signal width. The slow time $t_m = mT(m = 0, 1, 2, \cdots)$, and the fast time $\hat{t}$ and t are related as $\hat{t} = t - t_m$, which represents the time in one period.

4.1.2 Iso-depth Model. We first model the echo, and then introduce the iso-depth model.

To aid comprehension, we categorize the microphone's received signals into four distinct types. The first type comprises environmental sounds that originate from sources other than the speaker. The second type includes signals sent from the speaker that propagate directly to the microphone, either via the smartphone or via a line-of-sight path. The third type involves signals reflected from the user's face, and the fourth type comprises signals reflected from other objects in the vicinity. We use a high-pass filter to remove the first type of noise and subsequently employ a multi-path channel model to describe the microphone's received signal:

$$s_r(\hat{t}, t_m) = \sum_i \text{rect}(\frac{\hat{t} - \tau_i}{T_p}) e^{j2\pi(f_c(t-\tau_i)+\frac{1}{2}k(\hat{t}-\tau_i)^2)}, \tag{2}$$

where τ_i is the delay time from the i-th reflected path. To make it easier to understand, we rewrite the signal in the following form:

$$s_r(\hat{t}, t_m) = \sum_i s_i + \sum_j s_j, \tag{3}$$

where s_i denotes reflections from the target (i.e., the third type) and s_j denotes reflections from the others. In section 4.2, we describe how to remove the s_j.

To further understand the connection between echo signals and facial features, we match filtering the received signals. We transform the 3D structural features of the face into relative smartphone distances and the unique skin material information into energy features. Meanwhile, reflected signals at the same distance produce overlapping peaks after match filtering, which leads to the accumulation of energy at the corresponding distance

taps. The form is as follows,

$$E = \sum_i s_i \odot s_{ref}, \tag{4}$$

where $\odot$ is the signal convolution, s_i is echo signal of the i-th distance tap, and s_{ref} denotes the reference signal. The peak appears at a range that exactly matches the reference signal, and the width of the peak increases when the distance between the two peaks is less than the distance resolution, which provides the basis for building the iso-depth model.

Next, we established an iso-depth model, as shown in Fig. 3a, in which we divided the face into six regions based on their proximity to the smartphone. The nearest reflection point to the smartphone is the tip of the nose, followed by a circle around the nose. Next, the eye sockets, lips, and lower forehead are close to the line of the phone. The cheeks, being large, are divided into several different areas. The front half is close to the jaw and upper forehead, while the back half is a separate reflection area. Finally, the ear provides the farthest reflection. It is important to note that the energy of the received signal is affected by various factors, including the biomaterial of the reflective surface, and the distance, area, and orientation of the reflector. For example, although the tip of the nose is closest to the phone, the size of the reflective area is small, resulting in a relatively low energy signal. The eye sockets and forehead, on the other hand, have a large reflective area facing the phone, producing the highest reflected energy. Despite the cheek being larger in size, it does not face the phone head-on, and the microphone only picks up a small portion of the reflected signal. Finally, the ear is farthest away from the phone, and its complex structure absorbs most of the energy, thus producing the lowest reflected energy. As shown in Fig. 3b, the peak energy corresponds to our model. The change in energy occurring in the experiments with some areas of the face masked in Section 4.4 verifies the correctness of the iso-depth model.

In summary, we use match filtering to process the echo signal and then associate it with facial features based on their energy.

4.2 Signal Preprocessing

Firstly, the gain of smartphones for different frequency bands varies, so we cannot directly use the transmitted signal as a reference standard. As shown in Fig. 3c, the red line represents the frequency of the reference signal after the addition of the Hanning window, which is smoothly distributed from 10 KHz to 22 KHz. The blue line represents the frequency of the signal after the influence of the loudspeaker and microphone gain, and the energy in the band after 18 KHz tends to 0. Therefore, we need to obtain the actual signal sent by the speaker. Secondly, due to the limitation of mobile hardware itself, the microphone and speaker cannot work simultaneously, and we cannot determine the true send time of the received signal. As a result, we use cross-correlation to align the signals correctly. Thirdly, the received echo contains significant noise. The energy of the direct signal is greater than that of the reflected signal, and the target signal is buried in the noise. Therefore, we perform noise cancellation to remove unwanted noise. Finally, we apply a match filtering algorithm to obtain facial distance information, which is essential in identifying the user.

Obtain Reference Signal. To eliminate hardware effects and obtain the true transmitted signal, we measure their combined frequency response. Specifically, we collect signals twice in the same environment. For the second collection, we place a 1 cm × 1 cm cardboard surface at a distance of 15 cm in front of the smartphone's speaker. By taking the difference between the signals collected on both occasions, we remove environmental reflections. Finally, we use the single reflection from the cardboard as the reference signal. This process has no specific requirements regarding the device and environment, and users can easily replicate it. Mao's experiments in an anechoic chamber have demonstrated the effectiveness of this method [22].

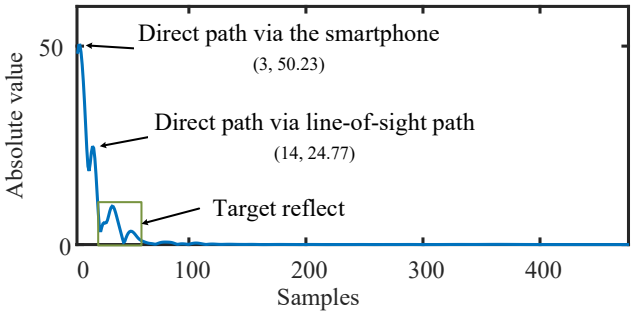


Fig. 4. Cross-correlation results.

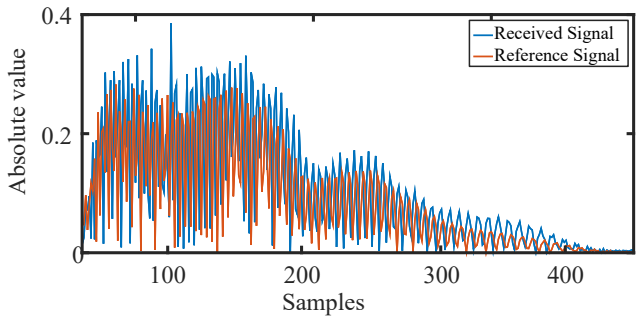


Fig. 5. Synchronize signal.

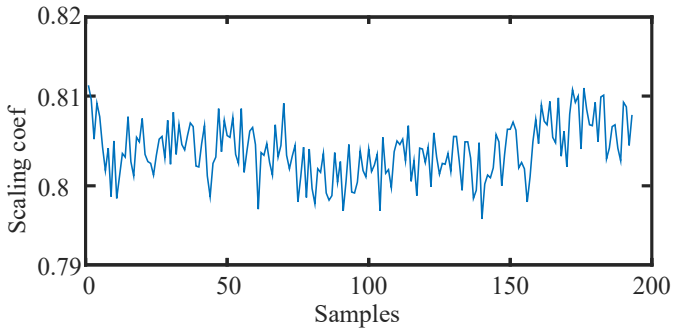


Fig. 6. Scaling coefficients.

Synchronize. In order to obtain the true send time of the signal, we use cross-correlation to locate the arrival time of the solid transmission signal. Sound travels faster in solids, and can reach the microphone from the speaker in a short time. The distance from the top speaker to the bottom microphone on a smartphone is only a centimeters, which takes about 0.04 ms. Meanwhile, the energy attenuation of solid propagation is smaller than

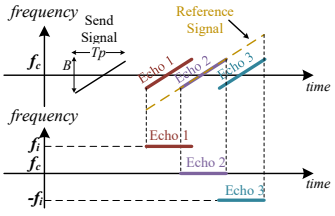
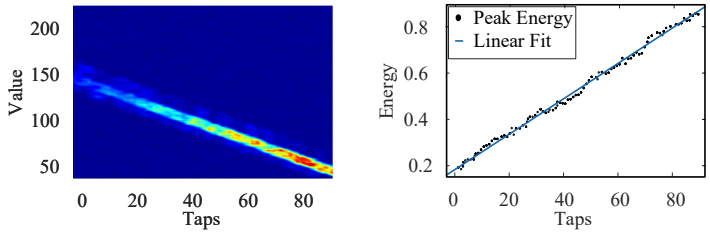


Fig. 7. The match filtering results of the reference signal and the echoes of three targets in different ranges.



(a) (b)

Fig. 8. (a) is the energy change as the phone approaches the target at a constant speed, and (b) is the fitting results of linear correlation between the energy of the received signal from a smartphone and the range of the reflected object.

that of air propagation, therefore, it is easy to locate the signal of solid propagation. Fig. 4 shows the correlation result of the synchronized signal, and we can easily distinguish the direct signal from the target. Fig. 5 shows the waveform of the synchronized signal. Although this method introduces an error of 1 to 2 sample points. This has no effect on the relative distance features of the user's face (i.e., difference of any two distances) and only causes an error of 0.7 cm on the absolute distance, which can be eliminated by the RA algorithm.

Noise Cancellation. We use different methods to eliminate environmental noise and signal noise. For environmental noise, most of it is below 10 KHz, so it can be removed using a high-pass filter. As for signal noise, we locate the noise start position through cross-correlation and subtract a reference signal to eliminate it. We notice that direct subtraction cannot completely eliminate the noise due to the different energy decay of reflected signals at various distances; they have similar waveforms but different amplitudes. Therefore, we introduce a scaling factor, ε. When the noise is completely canceled out, the value of $S_r - \varepsilon \times S_c$ is minimized, where S_r is the received signal and S_c is the pre-recorded clean signal. Hence, we can calculate the value of ε using the least squares method. Fig. 6 shows the variation of the scaling coefficients for a 200-frame signal. Once ε is determined, we subtract $\varepsilon \times S_c$ from S_r to remove the noise signals.

Matched Filtering Algorithm. After the alignment and noise cancellation in the previous two steps, the echo contains only s_i in Equation (3). Then match filtering is performed on the signal (i.e., conjugate convolution with the reference signal),

$$y_t = s_i \odot s_{ref} = \frac{\sin(\pi k T_p t)}{\pi k t} \text{rect}(\frac{t}{2T_p}) e^{j2\pi f_c t}. \tag{5}$$

The result can be approximated as a sinc function, so there will be peaks at ranges where there are reflectors. Fig. 7 illustrates the process of match filtering, the bandwidth of the send signal is B, the pulse width is T_p and the frequency modulation is k. After match filtering, the echo becomes a single-frequency pulse signal, and its frequency is proportional to the difference (ΔR) between the distance of the echo and the reference signal, i.e., $f_i = -k \times 2\Delta R/c$. Therefore, by performing a Fourier transform on the demodulated signal, sinusoidal narrow pulses corresponding to each echo can be obtained in the frequency domain. And then, the range where the reflected object is located can be calculated. We obtained different depth information about the user's face by match filtering.

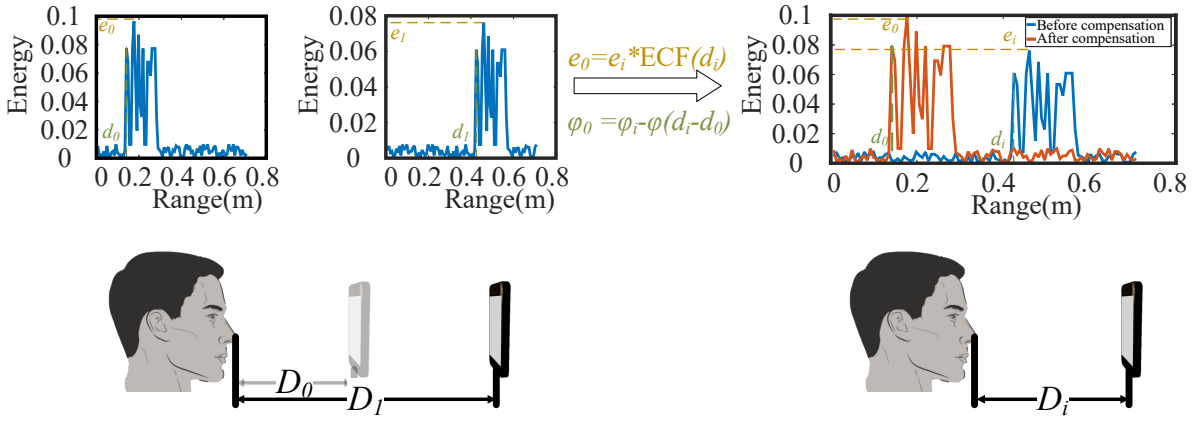


Fig. 9. Compensate the range from d_i to d_0.

Algorithm 1 Range Adaptive Algorithm

Input : signal after noise cancellation S_{nc}, reference signal S_{ref}, reference range D_{ref}, linear fitting function f
Output : signal after compensation S_c

1: $Relevance$ = XCORR(S_{nc}, S_{ref})
2: $range, maxEnergy$ = FINDPEAK($Relevance$)
3: $coef = ECF(range)/maxEnergy$
4: $\Delta d = D_{ref} - range$
5: $S_c = coef * S_{nc} * e^{2\pi(fc\frac{2\Delta d}{c} - kt\frac{2\Delta d}{c} + k\tau\frac{2\Delta d}{c} - \frac{1}{2}k(\frac{2\Delta d}{c})^2)}$
6: **return** S_c

4.3 Range Adaptive Algorithm

After completing the noise cancellation process, the signal only contains reflections from the user's face. However, the distance between the user's face and the smartphone is not constant for each authentication attempt. To address this issue, we propose a RA algorithm as depicted in Fig. 9. During the registration stage, the user places the smartphone at two different ranges, D_0 and D_1, to obtain facial echoes. AFace then extracts the closest ranges (i.e., d_0 and d_1) and the maximum energies (i.e., e_0 and e_1) from the echoes. The system then records d_0 and the ECF calculated from d_0, d_1, e_0, and e_1 as the reference range and energy compensation, respectively.

During the authentication stage, it is difficult to keep the distance from the user to the smartphone the same as during the registration stage. AFace extracts the closest range d_i and maximum energy e_i from the echoes and compensates the range to d_0 using the phase and ECF. Algorithm 1 gives the specific details of the algorithm. Firstly, cross-correlation is performed between the preprocessed signal and the reference signal to locate the range bin of the facial region. Then, the peak is located to determine the facial fine-grained range and maximum energy. Next, the signal scaling factor $coef$ and distance difference Δd is determined based on the ECF and reference distance. Finally, range and energy compensation are performed to obtain the final signal. Specific range and energy compensation methods are introduced in Section 4.3.1 and Section 4.3.2.

(a) The experimenter does not wear anything for the experiment.

(b) The experimenter wears a mask for the experiment.

(c) The experimenter covers lip for the experiment.

(d) The experimenter covers forehead for the experiment.

(e) The experimenter covers face for the experiment.

(f) The experimenter covers nose for the experiment.

Fig. 10. Experiment to verify the influence of the user wearing masks.

4.3.1 Range Compensation. The different range will be reflected in the phase deflection, thus the phase can be compensated to change the range. As shown in Fig. 9, when the user performs identity authentication, the range from smartphone to face is d_i. The reference range preset by AFace is d_0. The phase φ_1 of the real echo signal is

$$\varphi_1 = 2\pi f_c(t - \frac{2d_i}{c}) + \pi k(t - \frac{2d_i}{c})^2, \tag{6}$$

where the range from smartphone to the face is reference range d_0, the phase φ_0 is

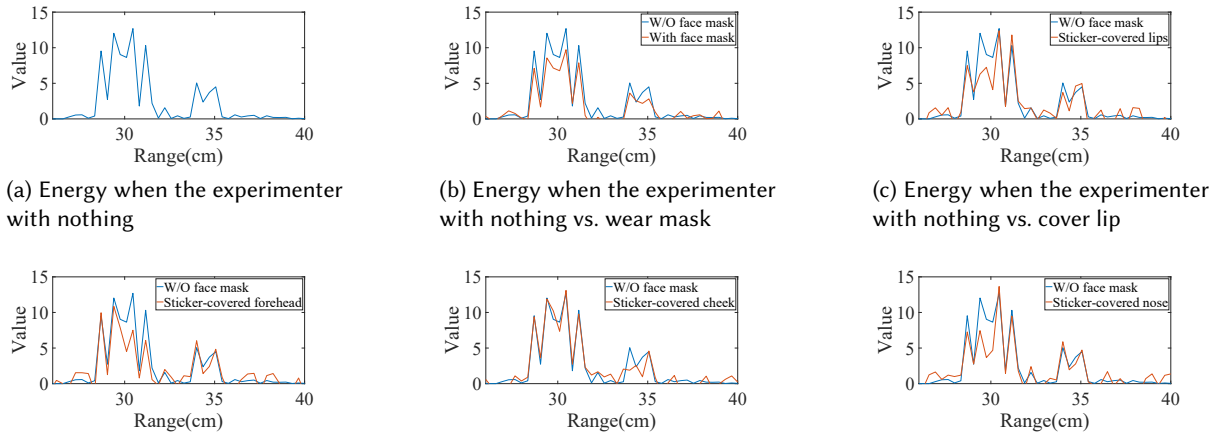
$$\varphi_0 = 2\pi f_c(t - \frac{2d_0}{c}) + \pi k(t - \frac{2d_0}{c})^2, \tag{7}$$

and the phase difference $\Delta\varphi$ between them is

$$\Delta\varphi = -4\pi\frac{f_c}{c}(d_i - d_0) - 4\pi k\frac{t}{c}(d_i - d_0) + \pi k\frac{4}{c^2}(d_i^2 - d_0^2), \tag{8}$$

therefore, we perform $e^{-j\varphi_1}/e^{-j\Delta\varphi}$ to compensate for the range to d_0.

4.3.2 Energy Compensation. We conducted experiments to confirm that there is a significant correlation between the range and energy of the signal. The test items we select include cardboard, wood, and iron plates. These items are moved toward a smartphone at a constant speed while playing 10 s of audio. We move the items toward the phone instead of the other way around to ensure that any reflections outside the test object are static. We extracted the range and peak energy changes from the resulting echoes. To visualize these results, we



(a) Energy when the experimenter with nothing

(b) Energy when the experimenter with nothing vs. wear mask

(c) Energy when the experimenter with nothing vs. cover lip

(d) Energy when the experimenter with nothing vs. cover forehead

(e) Energy when the experimenter with nothing vs. cover face

(f) Energy when the experimenter with nothing vs. cover nose

Fig. 11. Result of the influence of the user wearing masks.

first used a noise cancellation method to eliminate direct and background reflection signals. We then performed match filtering to determine the range variation of the target echo, with the brightness of the graph indicating the energy intensity at that range. Finally, we identified the maximum value for each column in the resulting graphs to represent the energy reflected by the object, as shown in Fig. 8a.

We then performed a linear fit to the peak energy data, as illustrated in Fig. 8b. The results revealed a strong linear correlation between power and range, with an R-square value of 0.9536 and an RMSE (Root Mean Square Error) value of 0.03198. Based on these findings, we can conclude that energy and range are highly correlated.

Therefore, when we know the reflected energy values from the user at two different positions, we can determine the unique relationship between energy and distance, i.e., the energy compensation equation. The form of the ECF obtained during the registration stage is as follows,

$$\mathrm{ECF}(d_i) = \frac{e_1 - e_0}{d_1 - d_0}(d_i - d_0) + e_0. \tag{9}$$

The result is the theoretical energy value of the reflection signal at a distance of d_i. We calculate the real energy value e_i by $e_i \times e_0/\mathrm{ECF}(d_i)$, which compensates the energy of the reflection signal at the distance d_i to d_0. This process does not directly move the energy value to e_0 in order to preserve the true energy characteristics of the signal. For example, when an attacker uses a 3D printing model, the real energy e_i and the theoretical energy $\mathrm{ECF}(d_i)$ differ significantly, so the compensated energy also differs noticeably from e_0, which helps distinguish the deceivers. In the experiment in Section 6.5, we verify the effectiveness of RA algorithm.

4.4 Feature Extraction

This section introduces how to extract the distance and energy features.

Distance Feature. After compensating for the signal, the next step is to extract features. In Section 4.1.2, we use iso-depth models to correlate echo signals with facial features. When extracting information, there are several key features we consider. First, we considered distance features from various facial regions, including the nose tip, nose root, lip and eye socket, cheek to forehead, the side of the face, and ears. We set the distance

of the ears to be 0 and calculated the relative distances among the other five regions. Subsequently, we took into account the pairwise differences among these five relative distances to represent the height of facial features (for instance, the difference in distance between the nose tip and nose root represents the height of the nose). Finally, we considered the area of reflection regions with similar distances, which can be represented using peak width after matched filtering (similar distances result in overlapping peaks in the matched filtering, increasing the width of the peaks). Specifically, we computed the width at 50 % energy of each peak. Since the distance between the nose root and the distance from the lips to the eye sockets are close, we directly calculated the width of these two peaks. In summary, we extracted five facial distance features, ten relative distance pieces of information, and five area-related pieces of information about similar distances.

Energy Feature. After conducting experiments, we have confirmed that the effect of biomaterials on signals, such as absorption rate and reflectivity, can be utilized for authentication. To demonstrate this, we conducted experiments under different facial coverage regions, as shown in Fig. 10, and measured the energy in both cases, as depicted in Fig. 11. The results indicate that the effect of biomaterials on energy is crucial for the resistance of acoustic authentication to 3D model spoofing.

Additionally, we conducted experiments with eight different participants in the same environment and extracted multiple energy features from the echoes, including maximum energy, average peak energy, variance, peak factor, average frequency, and energy density. The maximum energy reflects the comprehensive information of the user's facial skin texture and area, the average frequency, energy density, and average peak energy reflect the overall reflectivity of the user's face. The variance reflects the degree of dispersion of facial distance distribution, and the peak factor reflects whether the face has strong reflections. The anti-spoofing experiments in Section 6.4 have shown that adding energy features is necessary.

The features we used were based on the factors mentioned above and consisted of two parts. The first part is the processed echo signal, where we set a threshold of 30 % of the signal energy maximum, and the first peak above the threshold is considered as a reflection from the nose. We then intercept 40 samples after that as input features, providing a range of 14 cm, which covers the distance from the nose to the ears. The second part is the features extracted from the echoes. In terms of distance, we chose the nose tip, nose root, lip and eye socket, cheek to forehead, and the side of the face, as well as the differences between each of them and the width of the peak. In terms of energy, we extracted the 6 features mentioned earlier. In total, 26 features were extracted.

5 CLASSIFIER ARCHITECTURE

On one hand, the features we extract include not only distance and energy information but also processed signal segments. Therefore, the use of a BiLSTM (Bidirectional Long Short-Term Memory) neural network can better capture the temporal information embedded within these features. On the other hand, users' facial characteristics may change over time. Hence, we employ an iCaRL (Incremental Classifier and Representation Learning) approach to dynamically update the neural network. This method effectively addresses catastrophic forgetting issues and eliminates the need for retraining the neural network [16].

5.1 BiLSTM neural network

The structure of the neural network is depicted in Fig. 12, comprising an input layer, four BiLSTM layers, a fully connected layer, a dropout layer, a softmax layer, and a classifier.

The input to the neural network is a 66×1-dimensional vector, consisting of a 40×1-dimensional processed (denoised and range-compensated) signal segment, 20×1-dimensional distance features, and 6×z-dimensional energy features. The distance features encompass 5 depth measures: nose tip, nose root, lip and eye socket, cheek to forehead, and the side of the face. Then, the additional 10 supplementary features are obtained by taking the difference of any two of them (e.g., the depth difference between the nose tip and nose root can

represent the height of the nose). Finally, we calculate the peak width of each depth feature in the matched filtering. Specifically, for each peak, we compute the width at 50 % of the peak energy. This feature can represent the range of distances of overlapping reflected signals (areas of the face with similar depths). The energy features include maximum energy, average peak energy, variance, peak factor, average frequency, and energy density. The maximum energy and peak factor reflect the presence of strong reflective regions on the face, while average peak energy, average frequency, and energy density reflect the overall reflectance of the face. Variance indicates the degree of dispersion of facial reflectance features. In the appendix, we present the energy features extracted from different users. The input layer is used solely for receiving and passing data, without performing any additional data processing.

The input to the BiLSTM layers is a 66×1 feature vector. We chose the tanh activation function, which scales the values of the hidden states to be between -1 and 1 to facilitate easier training. The gate activation functions use sigmoid, which constrains the output to be in the range of 0 to 1, indicating whether the gates are open. For the initialization of the input layer weights, we used Xavier initialization, which involves randomly sampling weights from a normal distribution with a mean of 0 and scaling the sampled weights to ensure a variance of $2 / (n_{in} + n_{out})$, where n_{in} and n_{out} represent the number of input and output units, respectively. This initialization method ensures that the weights are suitable for the tanh activation function, allowing the network to converge to local minima more quickly, thus improving training efficiency, even on mobile devices with varying computational capabilities. Afterward, we set the BiasInitializer to 'unit-forget-gate', initializing the forget gate's initial value close to 1. This is particularly effective in addressing the gradient vanishing problem in long sequence data.

Following the BiLSTM layers, there is a fully connected layer with 128 units, and the activation function chosen is LeakyReLU. Compared to the traditional ReLU activation function, LeakyReLU allows negative values to pass through, which helps in handling negative inputs and reduces the risk of gradient vanishing. Additionally, it exhibits greater robustness to noise and uncertainty. In our training data, aside from the extracted distance and energy features, we also have a portion of processed signal segments. Therefore, using the LeakyReLU activation function allows for better handling of data that may contain outliers or noise.

Afterward, there is a dropout layer with a rate set to 0.5, randomly dropping neurons to reduce the risk of overfitting. Finally, a Softmax layer is employed to output the features as a probability distribution of 40 classes, which is subsequently fed to the classifier to obtain the final output of 40 categories. This final output represents the prediction of the neural network model for the given input. The proposed neural network architecture provides an effective and efficient approach to the classification of the echo signal segments into different categories.

5.2 iCaRL approach

The incremental learning method [32] involves enabling a pre-trained model to acquire features from new classes (class increment) or updating features of old classes with new data samples (data increment).

Class Incremental Learning. In our identity authentication system, we employ class increment learning when enrolling new users. The system initially processes and extracts features from received reflection signals to obtain training data. Subsequently, a new fully connected layer is added to the neural network to accommodate the increased number of classes, which performs classification on top of the BiLSTM feature representation. Following this, the existing fully connected layers are frozen to ensure the preservation of previously learned features, reducing training time and maintaining model stability. Finally, the model is compiled and trained on the new data.

Data Incremental Learning. When users fail acoustic facial authentication and successfully authenticate using an auxiliary method, the system assumes changes in the user's facial features and stores the modified facial feature data. When the saved new facial feature data exceeds a certain threshold (we set the threshold at

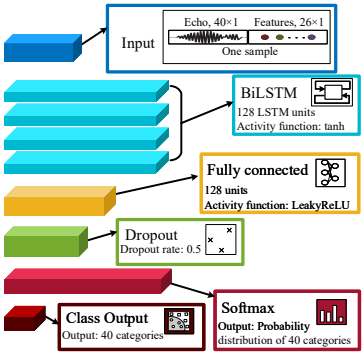


Fig. 12. Incremental learning network model based on BiLSTM.

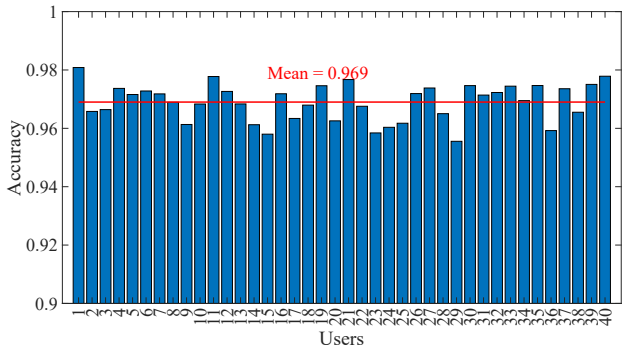


Fig. 13. Overall performance of the 40 users.

5, meaning that after five failures in acoustic facial authentication, with successful auxiliary method authentication), data increment learning is initiated. This means that the original neural network model structure remains unchanged, and the model is fine-tuned with the new data at a low learning rate.

6 PERFORMANCE EVALUATION

In this section, we present the implementation of our system and evaluate its performance.

6.1 Experimental Setup

Environments. We conducted performance tests on various hardware devices, including the Huawei Nova 7, Google Pixel 7 pro, and Redmi K60, and their audio sampling rates are 48 KHz. We developed an Android app using Kotlin that utilized the phone's speaker and microphone for audio transmission and reception. To thoroughly evaluate the performance of our system in various identity authentication scenarios, we selected different locations, such as quiet bedrooms, offices, roadways, and noisy halls. We also allowed experimenters to perform the authentication in different postures, including standing, sitting, and lying down.

Table 1. Range and noise level of different environments

	Bedroom	Office	Hallway	Roadside
Actual scene				
Distance from equipment to surroundings	0.5 m to 1 m	1 m	3 m	Almost no obstacles
Noise level	Below 40 dB	60 dB to 70 dB	70 dB to 90 dB	70 dB to 90 dB

Data Collection. We recruited 40 volunteers, including 20 females and 20 males aged between 20 and 40, to participate in our experiments. For the overall performance, our data is collected in four different environments. As shown in Table 1, including quiet bedrooms, offices, hallways, and roadside. In the bedroom environment, volunteers collect data while standing, sitting, and lying down. In the office and hallway, data is collected while volunteers are standing and sitting. At the roadside, data is collected while volunteers are standing. Volunteers collect data in four scenarios (comprising eight poses) for three sessions (corresponding to 3 devices), each lasting 10 seconds. To mitigate an abundance of similar samples, we randomly select two frames from the reflection signals every second for training or validating the model. In total, we gather 19,200 samples (40 individuals × 8 poses × 3 sessions × 20 samples per session). In Sections 6.3 to 6.8, we test the system's performance in various scenarios. These additional samples are gathered within the office environment. Through these tests, we are able to comprehensively evaluate the performance of AFace and demonstrate its robustness and reliability in different situations.

Metrics. To better evaluate the authentication performance of our system, we have introduced five different metrics: accuracy, recall, precision, defense success rate (DSR), and F1 score.

Accuracy represents the ratio of correctly classified samples to the total number of samples. Recall represents the ratio of the number of samples correctly classified as positive to the actual number of positives. Defense success rate (DSR) represents the ratio of the number of successful defenses to the number of spoofing attempts. Precision represents the ratio of the number of samples correctly classified as positive to the total number of samples classified as positive. The F1 score is used to combine recall and precision.

These metrics will provide a comprehensive evaluation of the system's performance under different scenarios and help us assess its robustness and reliability. They are defined as following:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{10}$$

$$\text{recall} = \frac{TP}{TP + FN}, \tag{11}$$

$$\text{precision} = \frac{TP}{TP + FP}, \tag{12}$$

$$\text{DSR} = \frac{NSD}{NSA}, \tag{13}$$

$$\text{F1score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, \tag{14}$$

where TP is the True Positive, TN is the True Negative, FP is the False Positive, FN is the False Negative, NSD is the number of successful defenses, and NSA is the number of spoofing attempts.

6.2 Overall Performance

In this section, we verify the performance of the system in multi-category mode and single-category mode on different devices, as well as the impact of the environment on the experimental results.

In the multi-category mode, all volunteers collect data in four different environments, including quiet bedrooms, offices, hallways, and roadside. In the bedroom, volunteers gather data while standing, sitting, and lying down. The distance between the device and the nearest wall is approximately 1 meter when standing or sitting, and about 0.5 meters when lying down. In the office, volunteers collect data while standing and sitting, with the device positioned approximately 1 meter away from the nearest wall in both scenarios. In the hallway, volunteers collect data while standing and sitting, and the device is situated over 3 meters away from the surrounding walls. By the roadside, data is collected with volunteers standing. This area is relatively open, with no obstructions, but have a relatively high noise level exceeding 70 dB.

Table 2. Single-Category Performance

		Accuracy	Recall	Precision	F1 Score
Quiet Bedroom	Mean	0.9869	0.9703	0.9872	0.9787
	Midian	0.9862	0.9726	0.9827	0.9776
Office	Mean	0.9709	0.9772	0.9722	0.9747
	Midian	0.9725	0.9757	0.9766	0.9761
Roadway	Mean	0.9657	0.9715	0.9688	0.9701
	Midian	0.9652	0.9689	0.9677	0.9683
Noisy Hall	Mean	0.9618	0.9631	0.9617	0.9624
	Midian	0.9617	0.9613	0.9617	0.9615

To prevent overfitting of the neural network, we use the data collected in the quiet bedroom as our training dataset and employ a 5-fold cross-validation method to evaluate the model's performance on different data partitions. The data collected in other environments served as the validation set to test how the neural network perform on unseen datasets. Additionally, we introduce L2 regularization parameters in the neural network. This regularization technique adds penalty terms to the loss function to constrain the size of the model's parameters and weights. By limiting the parameter values, it encourages the model to generalize better to unknown data, reducing the risk of overfitting. Finally, we implement an early stopping strategy. If the neural network's loss on the validation set do not decrease or accuracy do not improve for five consecutive epochs, we stop training to prevent overfitting. The specific process is as follows: 1) Randomly divide the dataset collected in bedroom into 5 folds, 2) For each fold, train a BiLSTM model using the other 4 folds as training data, evaluate the model's performance on the current fold as the validation set, and record the results, 3) Select the model with the best performance as the final model based on the recorded parameters, 4) Use data collected from other scenarios as a test set to evaluate the performance of the selected final model.

Fig. 13 shows the result, in all scenarios, the mean accuracy is 96.9 %. Fig. 14 shows the performance of the system in different environments. Our system performs exceptionally well in weak noise scenarios (standing and sitting in a bedroom or office), with an average accuracy above 98 %. In intense noise scenarios (standing by the roadway or in a hall), the noise cancellation method removes most of the environmental noise below 10 KHz, and the limited high-frequency noise has minimal impact on system performance due to the strong auto-correlation of the frequency-modulated signal, the average accuracy rate of our system is over 96 %. The performance is poor when users lie down, as the reflections of objects such as pillows and faces mix and are not easily distinguishable. Nevertheless, our noise cancellation method, combined with our selection of distance taps, effectively eliminates the majority of reflected signals in close proximity to the human face, and the accuracy remains above 93 %, and the misclassification rates are all below 3 %. The results prove that our system is able to capture the differences in characteristics of different users and provide reliable authentication.

In single classification mode, the output layer of the neural network provides the probability of a sample belonging to the positive class. To enhance the system's security, we initially set the threshold at 0.8. Subsequently, we adjust the threshold based on changes in system performance, and ultimately, our chosen threshold is set to 0.9. We conduct tests using the Huawei Nova 7. In each test, one volunteer registers in our system, and then all volunteers participate in the authentication testing (i.e., each system corresponds to one legitimate user and 39 illegitimate users). As shown in Table 2, the accuracy is above 96 % in all scenarios, and the average accuracy for all scenarios is 97.13 %. The impact of user pose on performance is similar to that of multicategorization. In real

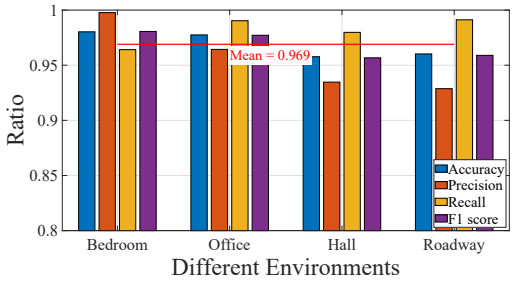


Fig. 14. Experiments at different environments.

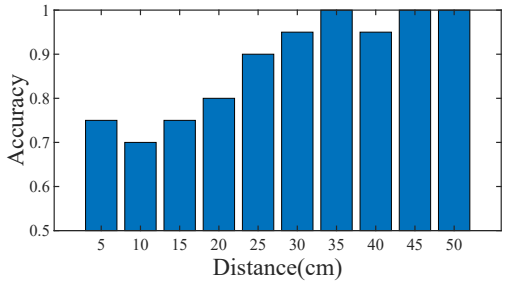


Fig. 15. The influence of obstacles at different distances.

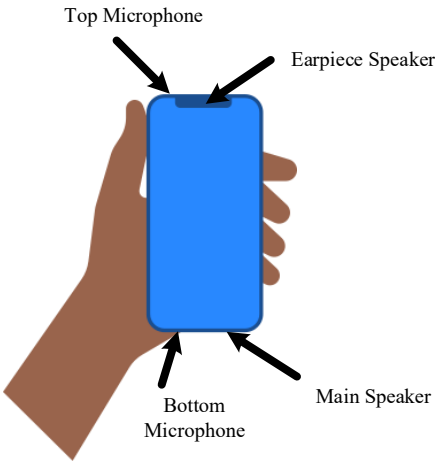


Fig. 16. The location of the microphone and speaker on the smartphone.

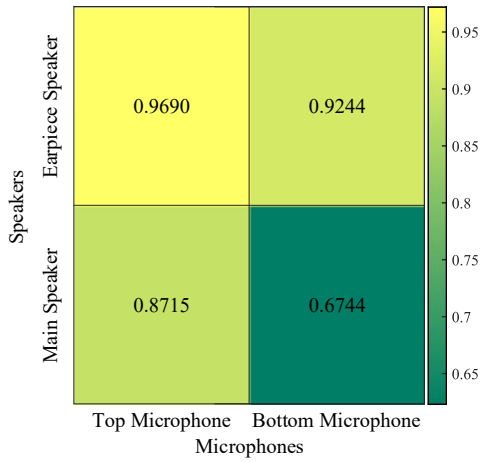


Fig. 17. Experimental results of different microphone&speaker combinations.

life, people's smartphones usually only store the identity information of one or two people. Therefore, this is proof enough that our system can be used for smartphones and provide convenient and reliable authentication.

Finally, we test the impact of environmental reflections at different distances on experimental performance. After registering in the system, volunteers place boards at various distances from their faces and perform identity verification. Due to the directional characteristics of the device's microphone and speaker, and the isolation provide by the specified range bin, the experimental results, as shown in Fig. 15, indicate that distances beyond 30 cm have almost no impact on experimental performance.

6.3 Different Microphone and Speaker Combinations

Next, we verified the effect of different microphone and speaker combinations. There are two speakers and two microphones on commercial smartphones today, as shown in the Fig. 16, located at the top and bottom of the smartphones. We conducted a multiclassification test with 40 volunteers on 4 combination approaches and they performe as shown in the Fig. 17, and it is found that the combination using the top microphone and

speaker performed the best. Analyzing the experimental results, we find that when using the bottom speaker or microphone, the echo signal is mostly reflected from the user's torso, and when the structure and material of the user's clothing changes, the experimental performance drops significantly, so we choose the top speaker and microphone.

6.4 Performance under Attack

Next, we evaluate the performance of our system under attack. To launch such an attack, the attacker steals an image of a legitimate user's face (2D attack) or captures the depth information of a legitimate user's face with an infrared scanner and uses 3D printing technology to obtain a face model (3D attack). For 2D plane attacks, attackers compromise the identity authentication system by surreptitiously obtaining users' photos or videos. Therefore, we use photos or videos of legitimate users to simulate 2D attacks. In the case of 3D attacks, attackers capture the depth information of a legitimate user's face using an infrared scanner or a depth camera. They then utilize 3D printing technology to create a model of the user's face. Considering that there will be small gaps between the 3D printed model and the real user's facial structure, and that it is not easy to simulate print attacks with multiple materials. Therefore, to test the system's resistance against deception with different printing materials, we have legitimate users wear masks made of various materials to mimic their own attackers and conduct 3D attacks.

For the experiments on resisting 2D deception, we use images and videos of real users as attackers. And for the experiments on resisting 3D deception, we have real users wear masks to simulate attackers, and we test three different types of masks. As shown in Table 3, the first type is silicone masks. Silicone is widely used in 3D printing technology due to its high biocompatibility, flexibility, and softness. It finds applications in medical implants, biomedical research, prosthetics, and more. Therefore, the use of silicone masks allows us to recreate real-world scenarios of 3D printing attacks. The second type is hyaluronic acid masks. Hyaluronic acid, naturally present in the human body in a quantity of approximately 15 g, is known for its ability to draw moisture from the skin's surface and enhance skin's long-term hydration capabilities. In the field of medical aesthetics, hyaluronic acid is utilized for skincare, repair, filling, and sculpting. Thus, hyaluronic acid masks closely resemble real human faces. The third type is mud masks. Comprising primarily of sea mud and pearl powder, we selected mud masks for their ability to be applied more closely to the facial surface, replicating the 3D structure of the face and simulating a more authentic 3D printing attack.

We compare two scenarios, with or without energy features, to validate the significance of energy features. Forty users register their personal information in separate systems, each operating as a single-class model with only one legitimate user. They are then subjected to both 2D and 3D deception experiments, where identity

Table 3. Different mask materials

	Silicone mask	Hyaluronic acid	Mud mask
Real picture			

Table 4. Performance under Attack

	Accuracy	Recall	DSR	Precision	F1 Score
2D attack(w/o energy features)	0.9750	0.9719	1	0.9828	0.9773
3D attack(w/o energy features)	0.5000	0.6596	0.0261	0.6679	0.6637
2D attack(with energy features)	0.9746	0.9706	1	0.9834	0.9769
3D attack(with energy features)	0.9666	0.9666	0.9841	0.9656	0.9661

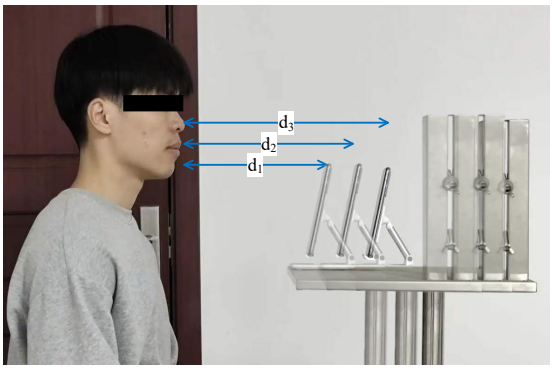


Fig. 18. Experiments at different distances.

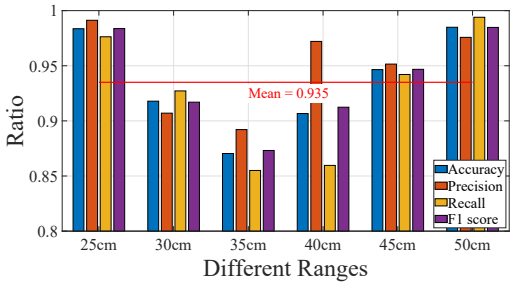


Fig. 19. Recognition performance with different ranges without RA algorithm.

authentication is attempted using photos and masks. This process takes place in the office, with each user repeating data collection 10 times in three different states: as a legitimate user, using a photo, and wearing a mask. As shown in Table 4, the final results reveal that 2D deception attempts are entirely resisted, whether energy features are used or not. This resistance can be attributed to our system's use of acoustic methods to capture facial structures, naturally providing robustness against 2D planar deception techniques. In the 3D deception experiments, without the use of energy features, the resistance success rate is less than 5 %. However, when energy features are employed, the resistance success rate exceeds 98 %. This substantial resistance is due to the fact that the facial structure of legitimate users wearing masks closely matches that of real users, resulting in nearly identical distance features. The varying reflective properties of different mask materials play a key role in these outcomes, resulting in echo signal energy that is different from a real human face. We provide detailed information about the masks in the Appendix A, along with a comparison of reflection signals with and without mask wear, and the neural network can learn the differences from them. Therefore, 3D printing attacks cannot deceive our system.

6.5 Impact of Different Ranges

In this subsection, we examine the impact of the distance between the user and the smartphone on the system and verify the effectiveness of the RA algorithm. We conduct single classification tests on 40 individuals, repeatedly collect data 10 times at 6 different distances (25 cm, 30 cm, 35 cm, 40 cm, 45 cm, and 50 cm), all within the office. Data at 25 cm and 50 cm are specifically used for the execution of the RA algorithm (the group involving the use of the RA algorithm) and model training in the experiments. The specific procedure is illustrated in the Fig.

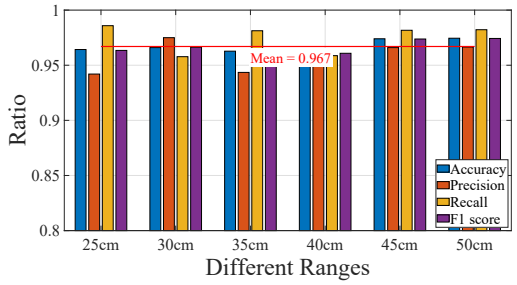


Fig. 20. Recognition performance with different ranges with RA algorithm.

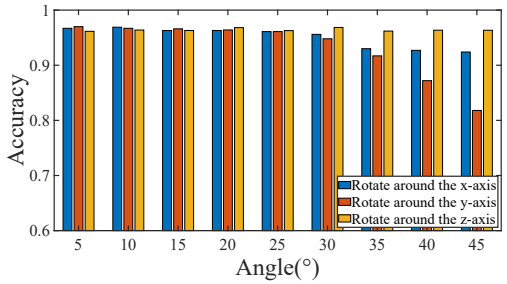


Fig. 21. Recognition performance with different angles.

(a) Rotate along the x-axis.

(b) Rotate along the y-axis.

(c) Rotate along the z-axis.

Fig. 22. Experiments at different angles.

18: First, the participants entered their identity information while sitting 25 cm and 50 cm from the smartphone. Then, they used the authentication system at distances of 25 cm, 30 cm, 35 cm, 40 cm, 45 cm, and 50 cm away from the smartphone. Finally, we compared the system's performance with and without the RA algorithm.

Fig. 19 shows the performance of our system for different ranges without RA algorithm, when the user's face is 25 cm or 50 cm away from the phone, the system's accuracy rate is over 96 %, and when the range is others, the accuracy rate is about 90 %. Fig. 20 shows the performance of our system for different ranges with RA algorithm. The accuracy for authentication at all ranges is above 95 %, and the average accuracy is 96.7 %, which proves that RA can greatly increase the range flexibility of our system.

6.6 Impact of Different Angles

Next, we tested the system's performance when users rotated the phone along the x, y, and z axes by certain angles.

During user authentication, we note three distinct types of phone rotations—around its x, y, and z axes. The rotation around the x-axis essentially signifies alterations in the phone's proximity to the user's face. Then, rotation around the y-axis modifies the plane in which the microphone and speaker are situated. Lastly, rotation around the z-axis adjusts the phone's orientation in relation to the user. Following, we empirically demonstrate the precise effects of these three types of rotations on the system's performance through rigorous experimentation. The experimental data is collected in the office, with performance tests conducted individually for 40 volunteers in a single classification scenario. Each volunteer repeats data collection for different rotation angles under various rotation modes 10 times. The data collected at 0° rotation angle are used for training the model,

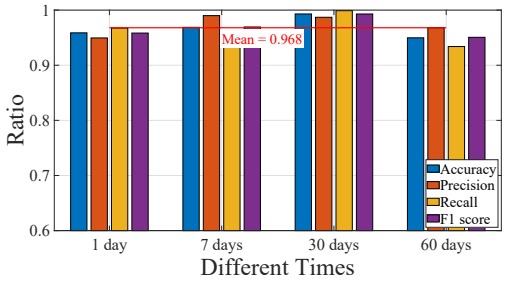


Fig. 23. Authentication performance in the first group over different time of periods.

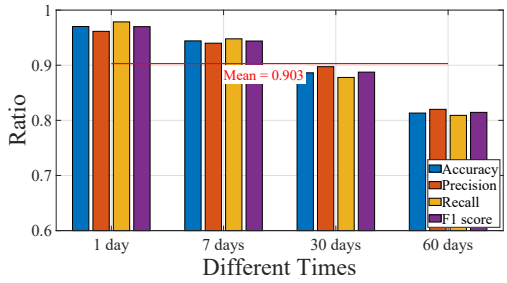


Fig. 24. Authentication performance in the second group over different time of periods.

while the data collected at the other 9 angles are used to test the system's performance under different rotational conditions.

Fig. 22 is a schematic representation of our experiment, where the phone undergoes rotations along the x, y, and z axes at specific angles (5°, 10°, 15°, 20°, 25°, 30°, 35°, 40°, 45°) before identity authentication. The experimental results are depicted in Fig. 21. Rotation along the x-axis primarily alters the horizontal distance from the phone to the user's face. Our system, equipped with the RA algorithm, effectively compensates for this distance change, resulting in minimal impact on performance. When rotating along the y-axis, the microphone and speaker of the phone may face the user's side, affecting the extraction of distance features. Quantitative analysis reveals that when the deflection angle is below 30 degrees, authentication performance remains largely unaffected. Rotation along the z-axis causes a slight change in the orientation of the microphone and speaker, which can mildly impact the extracted energy features, especially when other reflections surround the user's face and merge with the user's facial reflection signals. Considering the practical scenario, users typically position the phone facing their own faces when using the unlock function, our system continues to exhibit reliability.

6.7 Impact of Time

The signal segments contain in the training data may be affected by factors such as hair, glasses, or makeup. These effects can change over time. Therefore, we need to evaluate the robustness of the system over time and the effectiveness of incremental learning, we conduct the following experiments.

We divide all experimenters into two groups of 20 each and conduct tests in multi-classification mode. The first group uses the system with incremental learning, and the second group does not. Specifically, on the first day, we repeat data collection 10 times for each group of 20 users in the office, which is used to train the model. Then, we ask all experimenters to use our system continuously for two months. The time points we consider are the first day, 10 days, 30 days, and 60 days, meaning that at these specified time points, all volunteers repeat collect data 10 times in the office. Fig. 23 shows the results of the first group of experiments, and Fig. 24 shows the second results. We observed a significant decrease in the system's accuracy over time for the second group, and our system can maintain high performance over various time periods for the first group. It is worth noticing that the static geometry of the face will change slightly when the user is aging. Thus, incremental learning is the key to the long-term effectiveness of our system.

6.8 User Study

We next perform a user study based on our dataset on different factors, including gender and age.

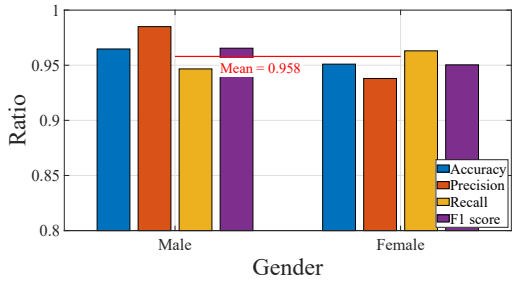


Fig. 25. User study on gender.

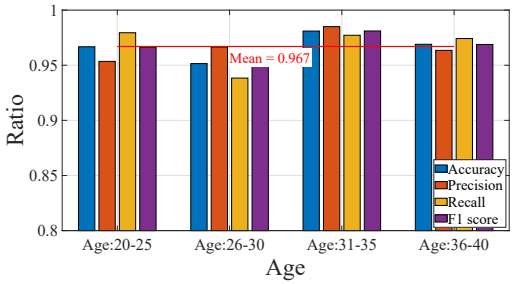


Fig. 26. User study on age.

Gender. First, we conducted a test to examine the influence of user gender on the system's performance. We selected 6 female and 6 male volunteers from our participants, dividing them into two separate groups based on gender. Both groups of users test the system performance in multi-classification mode. Data is collected in the office, with each user repeating the data collection 20 times. Out of these, 10 repetitions are used for cross-validation to train the model, while the remaining 10 repetitions of data collection are used to test the system's performance. To mitigate the influence of other factors, all our experiments were conducted in the same environment on the same day. The experimental results, as shown in Fig. 25, revealed a slightly lower authentication accuracy for female participants compared to male participants. When we analyzed the reflected signals received from female users, we observed that long hair on the front side of the face slightly affected feature extraction. Nevertheless, the recognition accuracy remained above 95 %.

Age. We next study how age could affect our system performance. Our participants are divided into age categories: 20-25, 25-30, 30-35, and 35-40. Each group consists of 5 volunteers, testing the system performance in multi-classification mode. Data is collected in the office, with each user repeating the data collection 20 times. Out of these, 10 repetitions are used for cross-training the model, and the remaining 10 repetitions are used to test the final model performance. Fig. 26 shows the average accuracy, recall, precision, and F1 score for these age groups. We observe that our system provides stable, appropriate performance for all age groups from 20 to 40. Specifically, the authentication accuracies for all age groups are above 95 %, and the variances are less than 5 %.

Overall, our system can provide a reliable method of identity authentication for users of different genders and age groups.

7 DISCUSSION ON LIMITATIONS

In this section, we discuss the limitations of the system, primarily focusing on performance degradation when users are wearing makeup or sweating, wearing accessories, and while walking.

We conduct experiments related to makeup and sweating. In the case of makeup, we have volunteers register in the facial recognition system without makeup and then wear makeup continuously for a week, during which we test the system. We compare the feature changes between makeup and no makeup scenarios. Due to the use of incremental learning, the neural network model is capable of learning slight variations in features, the system's authentication accuracy remain above 93 %. Regarding facial sweating, we discuss two scenarios, (a) a scenario where the user's face is moist but without obvious water droplets, and (b) a scenario with noticeable water droplets on the face. The experimental results show that for the first scenario, there is almost no impact on authentication performance, with the accuracy still maintain at over 95 %. However, for the second scenario, performance significantly decreased. Considering that scenarios involving intense sweating are relatively rare in

everyday life and that alternative authentication methods are available (such as PIN codes), our system remains reliable.

For facial obstructions, we discussed scenarios involving the wearing of glasses and masks. We collect and process the reflection signals when users wear and do not wear glasses. The result shows, when wearing glasses, the distance features remain largely unchanged, but there are significant changes in energy features. We observe that when wearing glasses, the distance from the glasses to the smartphone is similar to the distance from the tip of the nose to the smartphone. As a result, the energy of the first facial reflection signal significantly increased. The sound waves' penetrating characteristics cause relatively strong reflection signals near the eye sockets, but the energy show a slight decrease compared to when not wearing glasses. The authentication system's accuracy remain above 91 %. When wearing a mask, both distance and energy features were significantly affected. Our system cannot function when users wear masks. Combining it with other authentication methods, such as PIN codes, is a more suitable solution in such scenarios.

For the walking scenario, we analyzed situations that could affect performance. When taking the first step or immediately after landing a step, there is slight hand tremor due to a significant change in body posture. This tremor causes the distance from the smartphone to the face to change within one signal time (10 ms) and alters the order in which signals from the same frame reach the face. We verified this through experiments, and the correlation of reflection signals exhibited periodic disruptions, resulting in a significant drop in authentication performance. Fortunately, our authentication duration is around 1 second, allowing users to briefly pause for authentication and then continue walking, thus mitigating the impact of this challenge.

8 CONCLUSION AND FUTURE WORK

In this work, we propose AFace, a passive user authentication system that uses a smartphone to sense the 3D structure and skin texture of a user's face. Our study shows that the 3D structure and skin texture of each user's face is unique. We perceive facial features using an acoustic-based approach that utilizes the microphone and speaker on the smartphone. Extensive experimental results show that AFace is highly accurate in authenticating users. Results also show that the system performs well under different noisy environments with various daily activities. In the future, it is possible to introduce a dual-factor authentication method by having users make specific facial expressions or lip movements. This can further enhance the security and reliability of the system, ultimately enabling the deployment of the authentication system in real-world applications.

ACKNOWLEDGEMENT

This research was funded by the National Natural Science Foundation of China grant No.62272427 and 62072287, Hainan Provincial Joint Project of Sanya Yazhou Bay Science and Technology City No.2021JJLH0060, Youth Innovation Team of Shandong Provincial No.2022KJ043.

REFERENCES

[1] Da Ai, Weixin Fan, Kai Jia, Mingyue Lu, and Ying Liu. 2022. A Method of Dual-Spectrum Feature Fusion for Face Recognition Under Non-Ideal Lighting Conditions. In *Proceedings of the 4th International Symposium on Signal Processing Systems* (Xi'an, China) *(SSPS '22)*. Association for Computing Machinery, New York, NY, USA, 36–41. https://doi.org/10.1145/3532342.3532348

[2] Jorge P Arenas and Malcolm J Crocker. 2010. Recent trends in porous sound-absorbing materials. *Sound & vibration* 44, 7 (2010), 12–18.

[3] Y. Bai, L. Lu, J. Cheng, J. Liu, and J. Yu. 2020. Acoustic-based sensing and applications: A survey. *Computer Networks* 181 (2020), 107447.

[4] Z. Boulkenafet, J. Komulainen, and A. Hadid. 2017. Face Antispoofing Using Speeded-Up Robust Features and Fisher Vector Encoding. *IEEE Signal Processing Letters* 24, 2 (2017), 141–145.

[5] Rizhao Cai, Haoliang Li, Shiqi Wang, Changsheng Chen, and Alex C. Kot. 2021. DRL-FAS: A Novel Framework Based on Deep Reinforcement Learning for Face Anti-Spoofing. *IEEE Trans. Inf. Forensics Secur.* 16 (2021), 937–951. https://doi.org/10.1109/TIFS.2020.3026553

[6] Kai Cao and Anil K Jain. 2018. Automated latent fingerprint recognition. *IEEE transactions on pattern analysis and machine intelligence* 41, 4 (2018), 788–800.

[7] Jagmohan Chauhan, Yining Hu, Suranga Seneviratne, Archan Misra, Aruna Seneviratne, and Youngki Lee. 2017. BreathPrint: Breathing Acoustics-based User Authentication. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services, MobiSys'17, Niagara Falls, NY, USA, June 19-23, 2017*, Tanzeem Choudhury, Steven Y. Ko, Andrew Campbell, and Deepak Ganesan (Eds.). ACM, 278–291. https://doi.org/10.1145/3081333.3081355

[8] Huangxun Chen, Wei Wang, Jin Zhang, and Qian Zhang. 2019. Echoface: Acoustic sensor-based media attack detection for face authentication. *IEEE Internet of Things Journal* 7, 3 (2019), 2152–2159.

[9] Yongliang Chen, Tao Ni, Weitao Xu, and Tao Gu. 2022. SwipePass: Acoustic-based Second-factor User Authentication for Smartphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–25.

[10] Yimin Chen, Jingchao Sun, Xiaocong Jin, Tao Li, Rui Zhang, and Yanchao Zhang. 2017. Your face your heart: Secure mobile face authentication with photoplethysmograms. In *2017 IEEE Conference on Computer Communications, INFOCOM 2017, Atlanta, GA, USA, May 1-4, 2017*. IEEE, 1–9. https://doi.org/10.1109/INFOCOM.2017.8057220

[11] Bkav Corp. 2017. How Bkav tricked iPhone X's Face ID with a mask. https://www.youtube.com/watch?v=i4YQRLQVixM.

[12] Nesli Erdogmus and Sébastien Marcel. 2013. Spoofing 2D Face Recognition Systems with 3D Masks. In *2013 BIOSIG - Proceedings of the 12th International Conference of Biometrics Special Interest Group, Darmstadt, Germany, September 4-6, 2013 (LNI, Vol. P-212)*, Arslan Brömme and Christoph Busch (Eds.). GI, 209–216. https://dl.gi.de/handle/20.500.12116/17670

[13] Nesli Erdogmus and Sebastien Marcel. 2014. Spoofing face recognition with 3D masks. *IEEE transactions on information forensics and security* 9, 7 (2014), 1084–1097.

[14] Habiba Farrukh, Reham Mohamed Aburas, Siyuan Cao, and He Wang. 2020. FaceRevelio: a face liveness detection system for smartphones with a single front camera. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–13.

[15] Y. Jia, J. Zhang, S. Shan, and X. Chen. 2020. Single-Side Domain Generalization for Face Anti-Spoofing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[16] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences* 114, 13 (2017), 3521–3526.

[17] Klaus Kollreider, Hartwig Fronthaler, Maycel Isaac Faraj, and Josef Bigün. 2007. Real-Time Face Detection and Motion Analysis With Application in "Liveness" Assessment. *IEEE Trans. Inf. Forensics Secur.* 2, 3-2 (2007), 548–558. https://doi.org/10.1109/TIFS.2007.902037

[18] Chenqi Kong, Kexin Zheng, Shiqi Wang, Anderson Rocha, and Haoliang Li. 2022. Beyond the Pixel World: A Novel Acoustic-Based Face Anti-Spoofing System for Smartphones. *IEEE Transactions on Information Forensics and Security* 17 (2022), 3238–3253.

[19] Yan Li, Ke Xu, Qiang Yan, Yingjiu Li, and Robert H Deng. 2014. Understanding OSN-based facial disclosure against face authentication systems. In *Proceedings of the 9th ACM symposium on Information, computer and communications security*. 413–424.

[20] Noel B Linsangan, Ayra G Panganiban, Paulo R Flores, Hazel Ann T Poligratis, Angelo S Victa, Jumelyn L Torres, and Jocelyn Villaverde. 2019. Real-time iris recognition system for non-ideal iris images. In *Proceedings of the 2019 11th International Conference on Computer and Automation Engineering*. 32–36.

[21] Chao Liu, Penghao Wang, Ruobing Jiang, and Yanmin Zhu. 2021. AMT: Acoustic Multi-target Tracking with Smartphone MIMO System. In *40th IEEE Conference on Computer Communications, INFOCOM 2021, Vancouver, BC, Canada, May 10-13, 2021*. IEEE, 1–10. https://doi.org/10.1109/INFOCOM42981.2021.9488768

[22] Wenguang Mao, Mei Wang, and Lili Qiu. 2018. Aim: Acoustic imaging on a mobile. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. 468–481.

[23] Ekaterina Maro and Maksim Kovalchuk. 2018. Bypass biometric lock systems with gelatin artificial fingerprint. In *Proceedings of the 11th International Conference on Security of Information and Networks*. 1–2.

[24] Phillip McKerrow and Kok Kai Yoong. 2007. Classifying still faces with ultrasonic sensing. *Robotics and Autonomous Systems* 55, 9 (2007), 702–710.

[25] Jingyi Ning, Lei Xie, Chuyu Wang, Yanling Bu, Fengyuan Xu, Da-Wei Zhou, Sanglu Lu, and Baoliu Ye. 2023. RF-Badge: Vital Sign-Based Authentication via RFID Tag Array on Badges. *IEEE Transactions on Mobile Computing* 22, 2 (2023), 1170–1184. https://doi.org/10.1109/TMC.2021.3097912

[26] Gang Pan, Lin Sun, Zhaohui Wu, and Shihong Lao. 2007. Eyeblink-based Anti-Spoofing in Face Recognition from a Generic Webcamera. In *IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007*. IEEE Computer Society, 1–8. https://doi.org/10.1109/ICCV.2007.4409068

[27] Gang Pan, Lin Sun, Zhaohui Wu, and Shihong Lao. 2007. Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In *2007 IEEE 11th international conference on computer vision*. IEEE, 1–8.

[28] K. Patel, H. Han, and A. K. Jain. 2016. Secure Face Unlock: Spoof Detection on SmartPhones. *IEEE Transactions on Information Forensics & Security* 11, 10 (2016), 2268–2283.

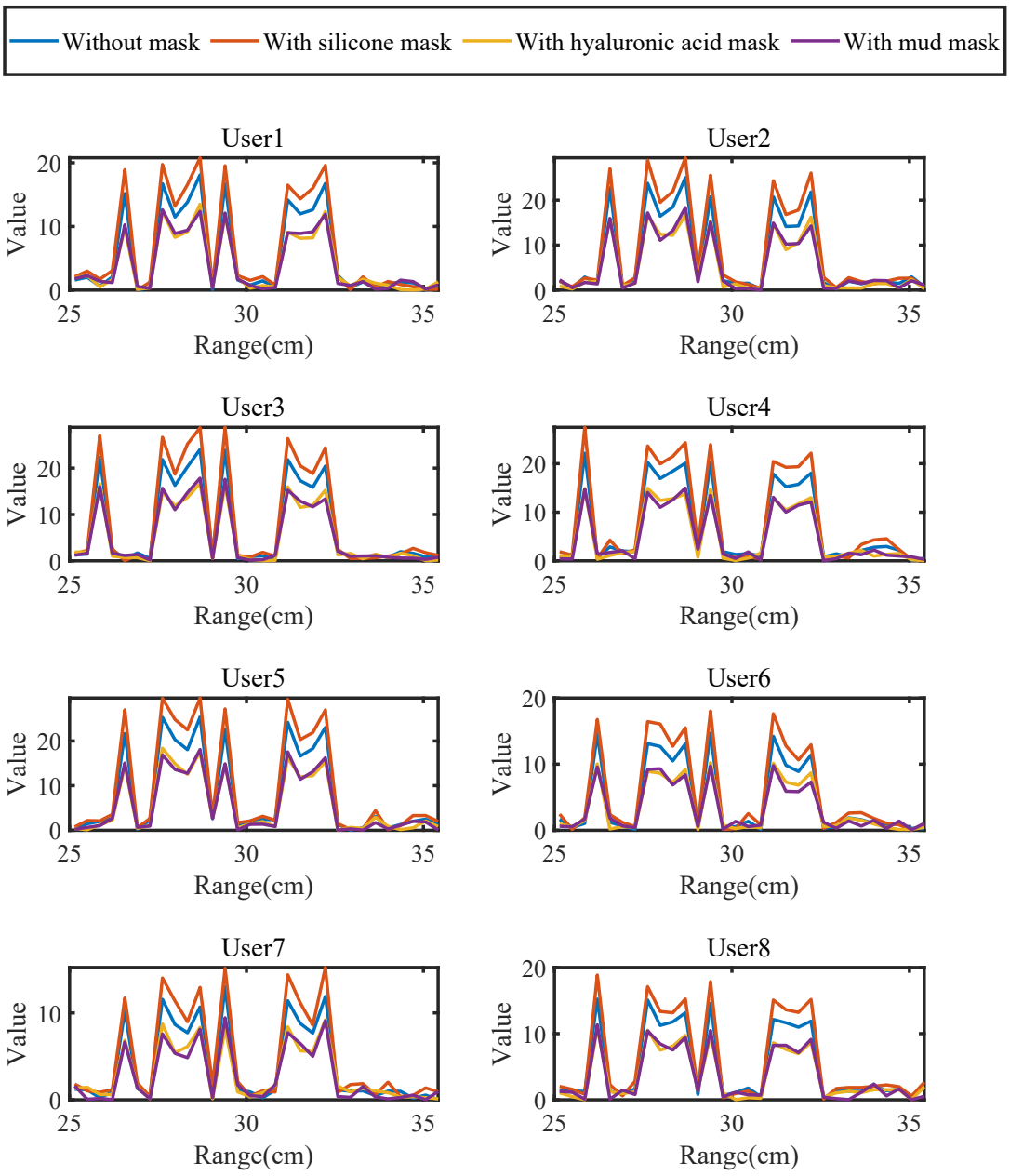
[29] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. 2008. *Dataset Shift in Machine Learning*. The MIT Press. https://doi.org/10.7551/mitpress/9780262170055.001.0001

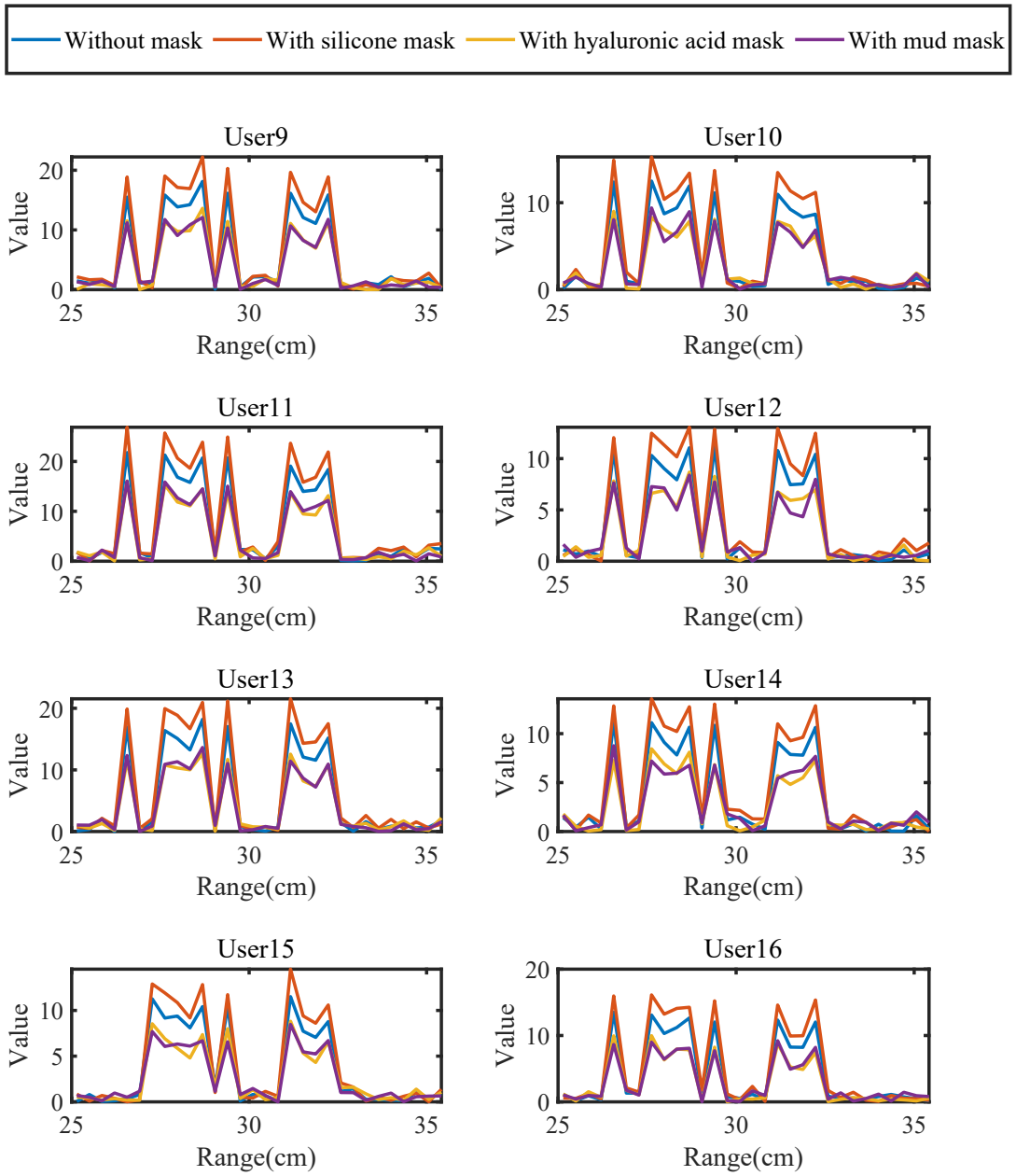
[30] Kiran B Raja, Ramachandra Raghavendra, Vinay Krishna Vemuri, and Christoph Busch. 2015. Smartphone based visible iris recognition using deep sparse filtering. *Pattern Recognition Letters* 57 (2015), 33–42.
[31] Steve Ranger. [n. d.]. iPhone X: This is how much it costs to make one, in components. https://www.zdnet.com/article/iphone-x-this-is-how-much-it-costs-to-make-in-components/.
[32] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. 2017. iCaRL: Incremental Classifier and Representation Learning. (2017). arXiv:1611.07725 [cs.CV]
[33] Antoaneta Roussi. 2020. Resisting the rise of facial recognition. *Nature* 587 (2020), 350 – 353. https://api.semanticscholar.org/CorpusID:227068107
[34] Yunpeng Song, Zhongmin Cai, and Zhi-Li Zhang. 2017. Multi-touch authentication using hand geometry and behavioral information. In *2017 IEEE symposium on security and privacy (SP)*. IEEE, 357–372.
[35] Chris Stein, Claudia Nickel, and Christoph Busch. 2012. Fingerphoto recognition with smartphone cameras. In *2012 BIOSIG-Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG)*. IEEE, 1–12.
[36] Lin Sun, Gang Pan, Zhaohui Wu, and Shihong Lao. 2007. Blinking-Based Live Face Detection Using Conditional Random Fields. In *Advances in Biometrics, International Conference, ICB 2007, Seoul, Korea, August 27-29, 2007, Proceedings (Lecture Notes in Computer Science, Vol. 4642)*, Seong-Whan Lee and Stan Z. Li (Eds.). Springer, 252–260. https://doi.org/10.1007/978-3-540-74549-5_27
[37] Furkan Tari, A Ant Ozok, and Stephen H Holden. 2006. A comparison of perceived and real shoulder-surfing risks between alphanumeric and graphical passwords. In *Proceedings of the second symposium on Usable privacy and security*. 56–66.
[38] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, and Tao Qin. 2021. Generalizing to Unseen Domains: A Survey on Domain Generalization. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, Zhi-Hua Zhou (Ed.). ijcai.org, 4627–4635. https://doi.org/10.24963/IJCAI.2021/628
[39] Penghao Wang, Ruobing Jiang, and Chao Liu. 2022. Amaging: Acoustic Hand Imaging for Self-adaptive Gesture Recognition. In *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications, London, United Kingdom, May 2-5, 2022*. IEEE, 80–89. https://doi.org/10.1109/INFOCOM48880.2022.9796906
[40] Tao Wang, Zhigao Zheng, Ali Kashif Bashir, Alireza Jolfaei, and Yanyan Xu. 2021. FinPrivacy: a privacy-preserving mechanism for fingerprint identification. *ACM Transactions on Internet Technology (TOIT)* 21, 3 (2021), 1–15.
[41] Wei Wang, Alex X Liu, and Muhammad Shahzad. 2016. Gait recognition using wifi signals. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 363–373.
[42] Zi Wang, Sheng Tan, Linghan Zhang, Yili Ren, Zhi Wang, and Jie Yang. 2021. An ear canal deformation based continuous user authentication using earables. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 819–821.
[43] Zi Wang, Sheng Tan, Linghan Zhang, Yili Ren, Zhi Wang, and Jie Yang. 2021. EarDynamic: An Ear Canal Deformation Based Continuous User Authentication Using In-Ear Wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–27.
[44] Weiye Xu, Jianwei Liu, Shimin Zhang, Yuanqing Zheng, Feng Lin, Jinsong Han, Fu Xiao, and Kui Ren. 2021. RFace: anti-spoofing facial authentication using cots rfid. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 1–10.
[45] Weiye Xu, Wenfan Song, Jianwei Liu, Yajie Liu, Xin Cui, Yuanqing Zheng, Jinsong Han, Xinhuai Wang, and Kui Ren. 2022. Mask does not matter: Anti-spoofing face authentication using mmWave without on-site registration. In *Proceedings of the 28th Annual International Conference on Mobile Computing and Networking*. 310–323.
[46] Wei Xu, Zhiwen Yu, Zhu Wang, Bin Guo, and Qi Han. 2019. AcousticID: Gait-based Human Identification Using Acoustic Signal. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3 (2019), 115:1–115:25. https://doi.org/10.1145/3351273
[47] Shulin Yang, Yantong Wang, Xiaoxiao Yu, Yu Gu, and Fuji Ren. 2020. User Authentication leveraging behavioral information using Commodity WiFi devices. In *2020 IEEE/CIC International Conference on Communications in China (ICCC)*. 530–535. https://doi.org/10.1109/ICCC49849.2020.9238889
[48] Z. Yu, X. Li, J. Shi, Z. Xia, and G. Zhao. 2021. Revisiting Pixel-Wise Supervision for Face Anti-Spoofing. *IEEE Transactions on Biometrics Behavior and Identity Science* PP, 99 (2021), 1–1.
[49] Jie Zhang, Xiaolong Zheng, Zhanyong Tang, Tianzhang Xing, Xiaojiang Chen, Dingyi Fang, Rong Li, Xiaoqing Gong, and Feng Chen. 2016. Privacy leakage in mobile sensing: Your unlock passwords can be leaked through wireless hotspot functionality. *Mobile Information Systems* 2016 (2016).
[50] Yi Zhang, Yue Zheng, Guidong Zhang, Kun Qian, Chen Qian, and Zheng Yang. 2021. GaitSense: towards ubiquitous gait-based human identification with Wi-Fi. *ACM Transactions on Sensor Networks (TOSN)* 18, 1 (2021), 1–24.
[51] Bing Zhou, Jay Lohokare, Ruipeng Gao, and Fan Ye. 2018. Echoprint: Two-factor authentication using acoustics and vision on smartphones. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. 321–336.
[52] Bing Zhou, Zongxing Xie, Yinuo Zhang, Jay Lohokare, Ruipeng Gao, and Fan Ye. 2021. Robust Human Face Authentication Leveraging Acoustic Sensing on Smartphones. *IEEE Transactions on Mobile Computing* (2021).
[53] Man Zhou, Qian Wang, Jingxiao Yang, Qi Li, Feng Xiao, Zhibo Wang, and Xiaofeng Chen. 2018. Patternlistener: Cracking android pattern lock using acoustic signals. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*.

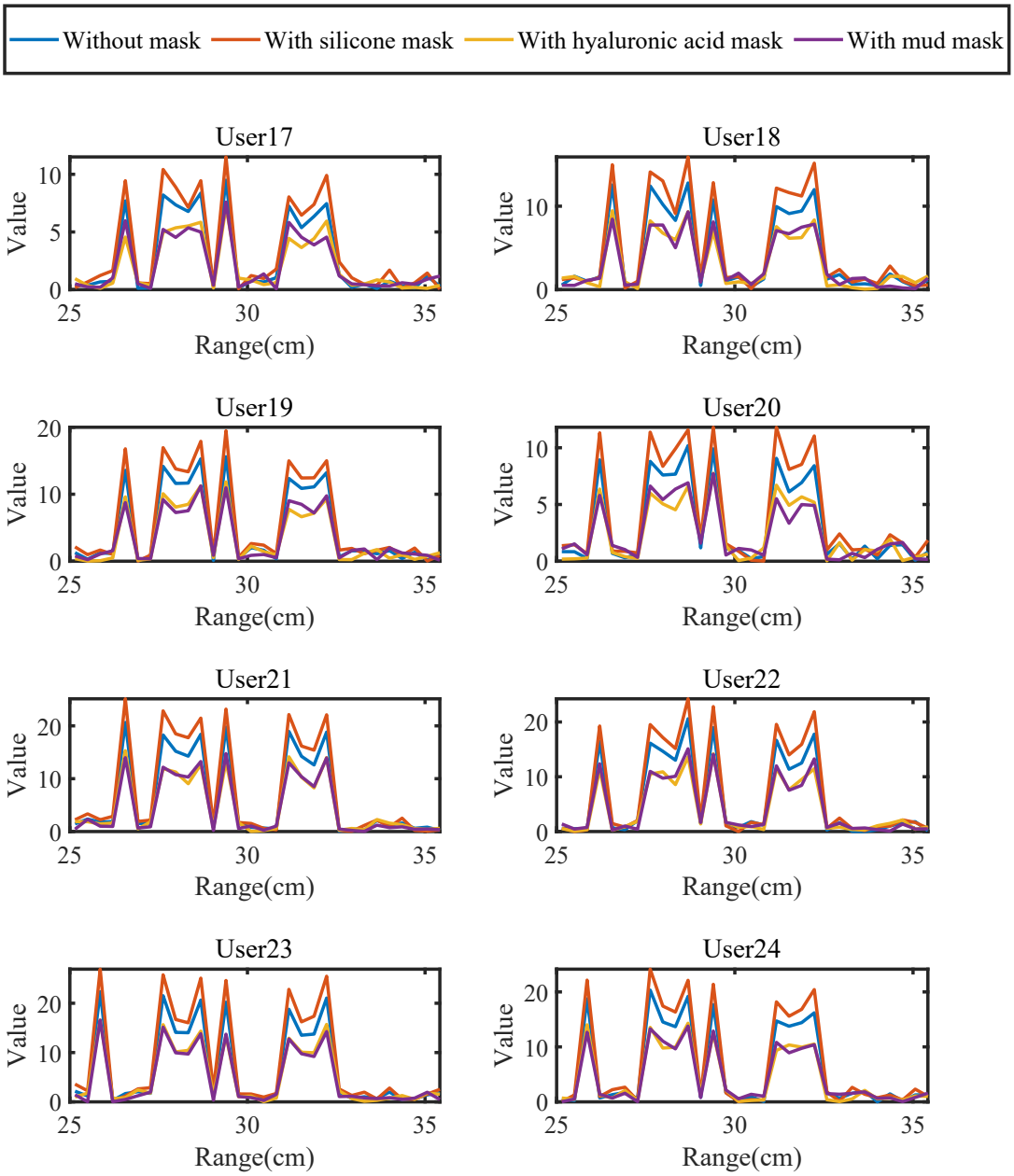
1775–1787.

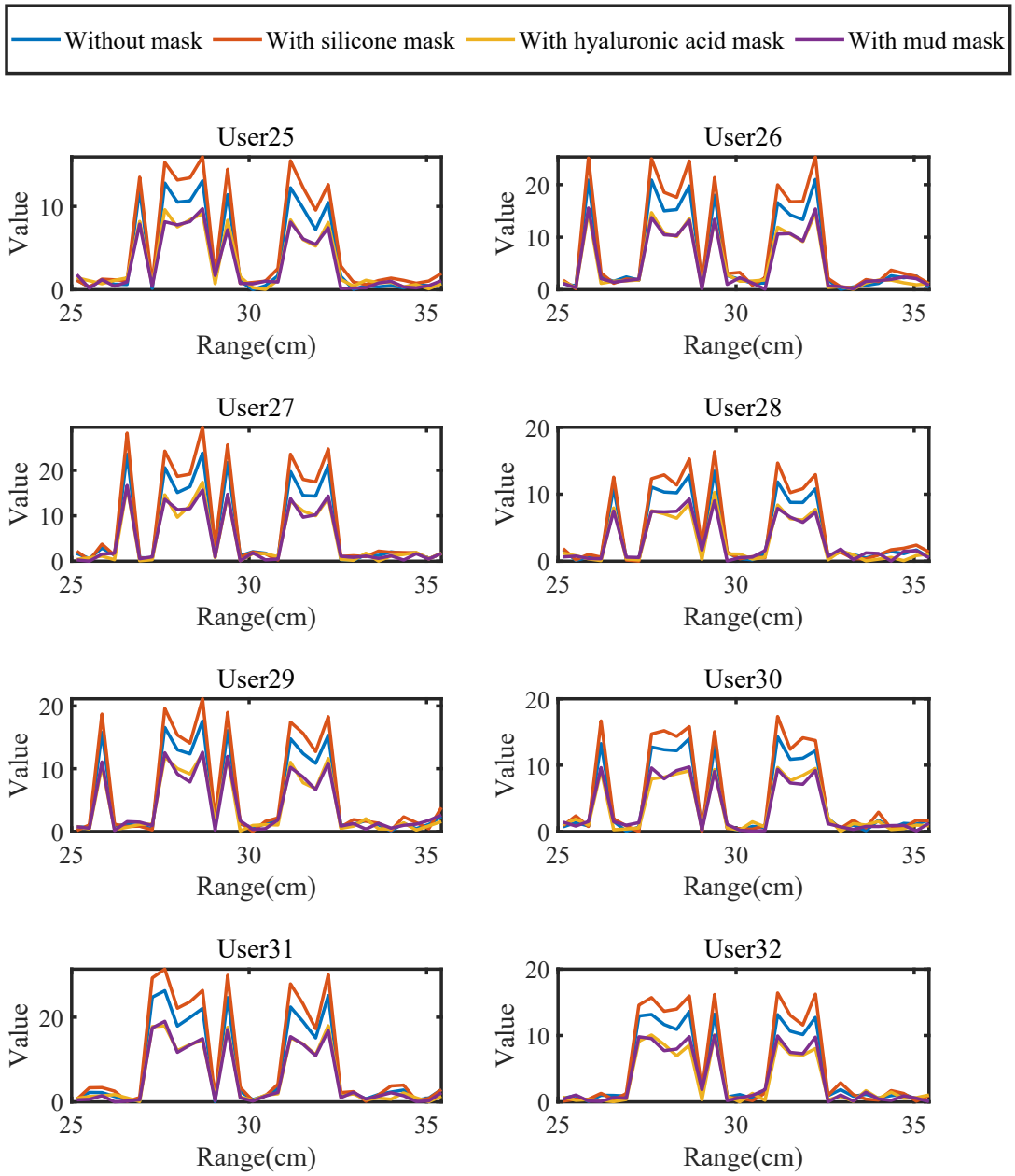
A APPENDIX: REFLECTIVE ENERGY CONTRAST

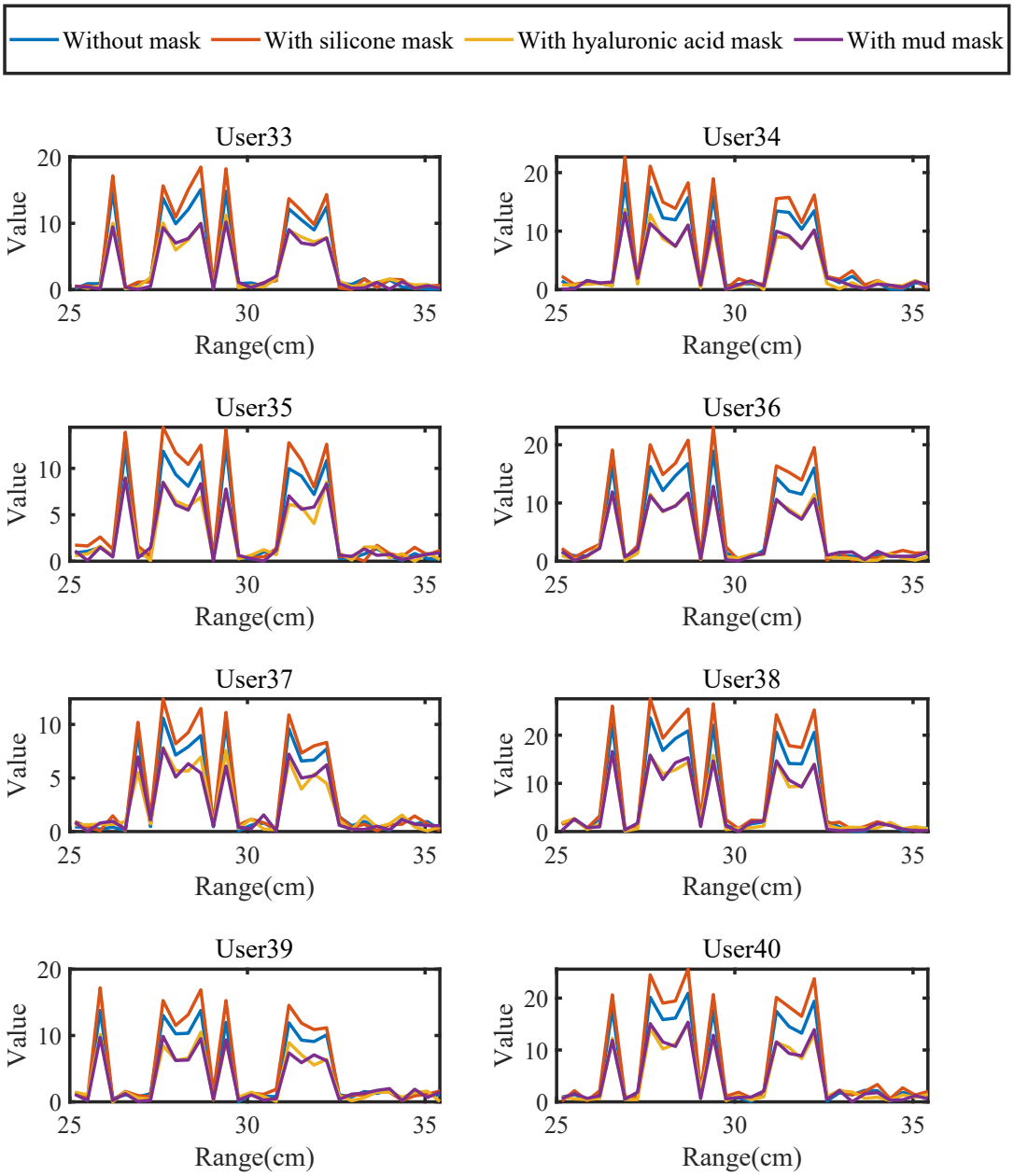
In this appendix, we provide comparison figures of the reflex signals of 40 users wearing masks of 3 different materials and without wearing masks.




Without mask
With silicone mask
With hyaluronic acid mask
With mud mask
User9
User10
User11
User12
User13
User14
User15
User16
Value
Range(cm)


Without mask
With silicone mask
With hyaluronic acid mask
With mud mask
User17
User18
User19
User20
User21
User22
User23
User24
Value
Range(cm)


Without mask
With silicone mask
With hyaluronic acid mask
With mud mask
User25
User26
User27
User28
User29
User30
User31
User32
Value
Range(cm)


Without mask
With silicone mask
With hyaluronic acid mask
With mud mask
User33
User34
User35
User36
User37
User38
User39
User40
Value
Range(cm)
25
30
35