UltraFace: Secure User-friendly Facial Authentication on Smartphones Using Ultrasound

Xinyue Fang[1], Jianwei Liu[1,2], Yike Chen[1], and Jinsong Han[1]

[1]Zhejiang University, China
[2]Hangzhou City University, China
{xinyuefang, jianweiliu, chenyike, hanjinsong}@zju.edu.cn

***Abstract*—With a wide range of common and privacy-sensitive applications, smartphones are frequently accessed for substantial personal information. Therefore, user-friendliness and security are crucial for user authentication on smartphones. Recently, convenient and secure biometric-based authentication is widely employed for smartphones, where the facial authentication stands out due to its potential for advancements in both user-friendliness and security. However, existing facial authentication methods possess some defects. For example, camera-based methods require good illumination conditions and are susceptible to 2D spoofing attacks. Moreover, previous acoustic-based methods either require camera assistance, or still suffer from 3D spoofing attacks. Even worse, some acoustic-based methods use audible sound waves, causing discomfort to users. To solve these questions, in this paper we propose *UltraFace*, an anti-spoofing and user-friendly facial authentication system on smartphones. It extracts facial geometry features and acoustic impedance features from imperceptible ultrasound. Leveraging the principle of ultrasound propagation, *UltraFace* correlates spectrograms of reflected signals with facial biometrics. Utilizing a deep learning model as a feature extractor, *UltraFace* mines fine-grained facial geometry features and acoustic impedance features from the spectrograms for accurate user authentication. Extensive experiments show that *UltraFace* achieves 97.2% accuracy in user authentication and can effectively defend against spoofing attacks. Furthermore, *UltraFace* exhibits robustness for long-term usage.**

***Index Terms*—Wireless Sensing, User Authentication, Facial Biometrics**

I. Introduction

Recent years have witnessed the prevalence of smartphones with various applications, containing significant amounts of personal information, such as social networks, user interests, and mobile banking [1], [2]. The user authentication system on smartphones is a critical factor in safeguarding privacy and ensuring the security of personal property. Traditional authentication methods of smartphones rely on "what users know", i.e., PIN [3]. But passwords set in such a knowledge-based approach can be easily forgotten and are vulnerable to shoulder-surfing attacks [4]. To avoid this dilemma, "something users are" are leveraged to provide more convenient and secure authentication, namely biometric-based authentication. Among biometric-based authentication, the facial recognition stands out due to its potential for advancements in both security and user-friendliness. However, existing facial authentication methods exhibit shortcomings. For example, camera-based face recognition requires specific lighting conditions [5] and is susceptible to 2D spoofing attacks by photos or videos [6].

Jinsong Han is the corresponding author.

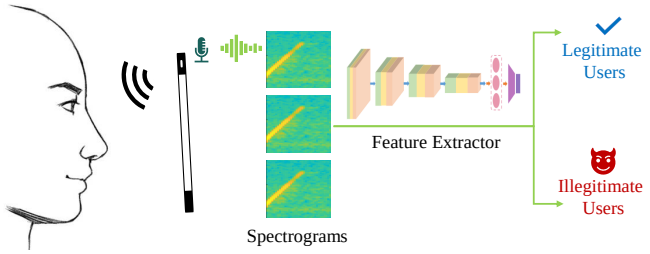


Fig. 1. Illustration of *UltraFace*.

In order to utilize existing smartphone sensors for reliable and 2D spoofing-resistant facial authentication, previous researches have explored facial features through acoustic sensing using built-in speakers and microphones. For example, Zhou et al. [1] introduce a two-factor authentication that combines acoustic features and vision facial landmark features as the joint feature description of the user. Chen et al. [7] analyze the similarity between reflected signals of two microphones to obtain the uneven stereostructure of the user's face. Kong et al. [8] design a two-branch framework that utilizes global and local frequency clues of acoustic signals for anti-2D spoofing facial authentication. Xu et al. [9] propose an iso-depth model and a range-adaptive algorithm to extract facial structure features and biological material features. However, these methods either require the use of cameras [1], which can be difficult to implement under poor illumination conditions, or emit sound waves in the audible frequency range [7]–[9], causing discomfort to users. Additionally, some methods only capture the 3D structure of the user's face, which still could be deceived by 3D masks [1], [7], [8]. Therefore, there is an urgent demand for a secure and user-friendly authentication method on smartphones.

In this paper, we propose a novel anti-spoofing facial authentication system on the smartphone named *UltraFace*. Leveraging the smartphone's built-in speaker and microphone, *UltraFace* emits and captures ultrasound to probe facial geometry and impedance for user-friendly authentication. As shown in Fig. 1, when an authentication attempt is initiated, *UltraFace* emits a specifically designed ultrasonic signal. Then, the ultrasound reflected from the user's face is received by *UltraFace* to extract facial biometrics, i.e. facial

geometry features and acoustic impedance features, for user authentication. Compared with previous works, *UltraFace* can work solely with the microphone and speaker, independently of cameras, which makes it effective in poor illumination conditions. Besides, *UltraFace* is highly secure and capable of defending against both 2D and 3D spoofing attacks. Since ultrasound is imperceptible to humans and can be transmitted automatically without user cooperation, *UltraFace* also has outstanding usability and user-friendliness.

To realize *UltraFace* in practice, we address the following questions: (1) **How to establish a correlation between the ultrasound reflection and facial biometrics?** Unlike vision data such as images and videos, ultrasonic signals are inherently less intuitive. Thus, identifying signal parts that encapsulate facial features is challenging. To investigate the manifestation of facial characteristics in the reflected ultrasound, we build a theoretical connection between them based on the ultrasound propagation principle. Upon this theoretical model, we perform frequency-domain analyses on the ultrasonic signals to obtain spectrograms that capture both geometry features and acoustic impedance features of the user's face. (2) **How to extract effective features from noisy spectrograms to achieve accurate user identification?** The ultrasonic signals received by the microphone contain both reflection from the user's face and other environmental noise. To realize the effective feature extraction for user authentication, we first design a signal processing method and then develop a deep learning model to extract fine-grained facial geometry features and acoustic impedance features of individuals for accurate authentication.

We build a prototype of *UltraFace* on a commercial off-the-shelf (COTS) smartphone and perform extensive experiments. We evaluate *UltraFace* with 24 volunteers under different conditions. The experiment results show that *UltraFace* can achieve an authentication success rate of 97.2%. Meanwhile, *UltraFace* can defend against various attacks. Moreover, the robustness study demonstrates that the performance of *UltraFace* will not degrade over time. In summary, our contributions are as follows:

- We propose a spoofing-resistant and user-friendly authentication system on smartphones, namely *UltraFace*. It achieves non-intrusive user identification by capturing facial biometrics (i.e., facial geometry features and acoustic impedance features) from reflected ultrasound.
- We establish a theoretical correlation between the ultrasonic signals and facial biometrics using the principle of ultrasound propagation. Based on this foundation, we propose a learning-based approach to further extract facial biometrics from ultrasonic signals.
- We implement *UltraFace* with COTS smartphones and conduct extensive experiments. The experiment results indicate that *UltraFace* can achieve accurate and robust authentication, while defending against various attacks.

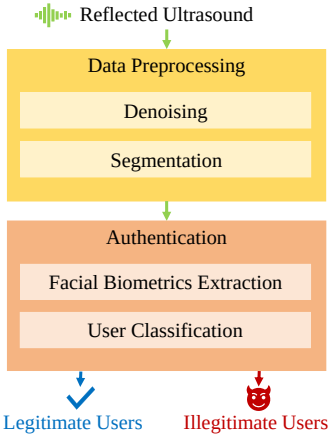


Fig. 2. Workflow of *UltraFace*.

II. Threat Model

Security is paramount for an authentication system. In an adversarial environment against smartphone user authentication, the attacker aims to steal private information or carry out unauthorized operations. In the threat model, we assume that the attacker has gained physical access to the user's smartphone. We focus on three types of most threatening and common attacks on smartphone authentication.

Zero-effort attack. In this attack, we assume that the attacker is unfamiliar with *UltraFace*. Instead of following the authentication process to present facial biometrics, the attacker attempts to bypass the authentication system effortlessly.

2D spoofing attack. For this attack, we presume that the attacker holds 2D photos or videos of a legitimate user's face. S/he tries to deceive *UltraFace* by using 2D photos or videos in place of live facial biometrics.

3D spoofing attack. In this attack, we consider that the attacker has acquired the depth information of a legitimate user's face using the infrared scanner. S/he utilizes 3D printing technology to create a replica of the user's face. This replica is subsequently employed to trick *UltraFace*.

III. System Overview

To realize a secure user-friendly authentication system on smartphones, *UltraFace* extracts facial geometry features and acoustic impedance features from the received ultrasonic signals to distinguish individuals. As shown in Fig. 2, *UltraFace* is primarily composed of two modules: data preprocessing and user authentication. When an individual initiates an authentication request on the smartphone, *UltraFace* emits an ultrasonic wave as the stimulus signal. Subsequently, the emitted ultrasound is reflected by the face and received by the smartphone. The data preprocessing module then performs signal denoising and segmentation to obtain a clean signal. After that, the clean signal is fed into the user authentication module. In this module, *UltraFace* extracts facial biometrics (i.e., facial geometry features and acoustic impedance features) from the clean signal and employs a deep learning model to achieve user identification. Particularly, the emitted ultrasound is a chirp signal ranging from 17kHz to 22kHz with a duration

of 0.1s. To prevent echo overlap, a 0.4s interval is added between consecutive signals. Such a frequency band is barely detectable by the human ear, making the system user-friendly.

Data preprocessing. The goal of this module is to provide clean ultrasound samples. It first removes the noise outside the interested frequency band through filtering, and then segments the signals to obtain each complete chirp signal. We will elaborate on this module in Sec. IV.

User authentication. In this module, *UltraFace* initially derives the signal spectrograms from the clean ultrasound samples, as spectrograms contain the user's facial biometrics. Then, *UltraFace* utilizes a well-designed deep learning model to extract fine-grained facial features from the spectrograms and perform user classification. If the probabilities output by the deep learning model for all classes fall below a predefined acceptance threshold, *UltraFace* identifies the individual as an illegitimate user and denies the access. Otherwise, the authentication is considered to be initiated by a legitimate user associated with the highest probability. The user authentication approach will be detailed in Sec. V and Sec. VI.

IV. Data Preprocessing

Before analyzing the ultrasonic data, we first perform denoising and segmentation for preprocessing. Raw ultrasonic signals reflected by the face include environmental and hardware noise that is irrelevant to the facial biometrics. Therefore, we utilize a denoising method to obtain clean ultrasound measurements. Additionally, to obtain each chirp signal for subsequent feature extraction, we locate each chirp profile and perform signal segmentation.

A. Signal Denoising

Besides capturing facial reflected signals, the smartphone's microphone also records environmental and hardware noise, which can affect authentication performance. To address this issue, we apply a bandpass Butterworth filter to remove noise, thereby producing stable amplitude-frequency characteristics that embody facial biometrics. The passband frequency range of the filter is set to 17-22kHz. Moreover, we utilize a Hamming window [10] to smooth the reflected signals and reduce frequency leakage caused by signal truncation.

B. Signal Segmentation

Next, we adopt a segmentation method to locate and isolate the chirp signals from denoised ultrasonic data. Time stamp-based approaches are inadequate for handling unintentional signal interruptions caused by hardware instability. So we employ a phase-locked loop [11], which is effective in locating signals at specific frequencies. Specifically, we detect 17kHz and 22kHz frequency points within signals. These cutting points are then used to segment each chirp signal accurately.

V. Facial Biometrics Extraction from Ultrasound Reflection

To extract effective and representative facial biometrics, we first analyze the relationship between the reflected ultrasound and facial geometric features, as well as facial acoustic impedance features. Based on the theoretical model, we then perform a qualitative analysis of these features using signal frequency-domain analysis.

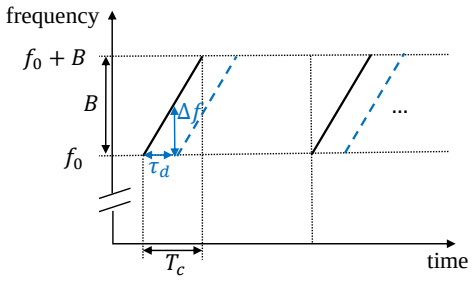


Fig. 3. Chirp signal.

A. Facial Geometry Features

Facial geometry features are one type of the facial biometrics that indicate the characteristics of facial structures. To acquire accurate geometry features, we establish a theoretical correlation based on the principle of ultrasound propagation to analyze the geometry information in ultrasound reflection.

As shown in Fig. 3, the carrier frequency $f_c(t)$ of chirp signals increases linearly over time during a signal period T_c, expressed as $f_c(t) = f_0 + \frac{B}{T_c}t$, where f_0 is the initial frequency and B is the sweep bandwidth. Since the chirp signal is initially transmitted by the speaker and subsequently reflected by the target object back to the microphone, the reflected signal is received after a round-trip time delay τ_d with a proportional frequency shift Δf. Based on the geometry similarity principle of a triangle, the delay τ_d can be derived as:

$$\tau_d = \frac{\Delta f}{B} T_c. \tag{1}$$

Consequently, the path distance d between the smartphone and the target object can be represented with the frequency shift between transmitted and received signals as:

$$d = \frac{v}{2} \cdot \tau_d = \frac{v}{2} \cdot \frac{\Delta f}{B} \cdot T_c, \tag{2}$$

where v is the speed of acoustic signals in the air.

It can be deduced from Eq. 2 that the distance between the smartphone and the target object is proportional to the frequency shift Δf. Therefore, the geometry features of the target object can be embodied by the frequency information of the reflected ultrasound. Further, since different people possess different facial geometry features, there exists the potential to help user authentication by extracting the facial geometry features through ultrasonic sensing.

B. Facial Acoustic Impedance Features

Besides facial geometry features, facial acoustic impedance features are another type of facial biometrics that indicate the characteristics of facial biological materials. According to [12], the acoustic impedance information of skin can be embodied by the intensity of the reflected ultrasound I_r, which can be expressed as:

$$I_r = rI_i = (\frac{Z_{n2} - Z_{n1}}{Z_{n2} + Z_{n1}})^2 I_i, \tag{3}$$

where r is the reflection coefficient of ultrasound intensity, I_i is the intensity of incident ultrasound, Z_{n1} and Z_{n2} are the perpendicular acoustic impedance of the air and the skin at the interface, respectively. Due to the variability in facial acoustic impedance features among individuals, ultrasound sensing can be utilized to extract facial acoustic impedance features for user authentication.

C. Spectrogram Derivation

According to the above theoretical analysis, the frequency and intensity information of the reflected ultrasound can reveal facial geometry features and acoustic impedance features, respectively. Meanwhile, we noticed that the Short-Time Fourier Transform (STFT) can make the variation of frequency and intensity over time more intuitive and prominent. Therefore, we apply the STFT technique to transform the ultrasonic signals into the spectrograms.

Specifically, the spectrogram $s(\tau)$ is calculated based on a window function $h(\tau)$ with length T. The signal is segmented into multiple short-time windows, and the Fourier Transform is applied to each window to derive spectral information. By moving the window along the timeline, the spectral characteristics of the signal at different time can be obtained. This process can be calculated as follows:

$$STFT(t,f) = \sum_{\tau=t}^{t+T-1} s(\tau)h(\tau - T)\exp(-j2\pi f\tau), \quad (4)$$

where t and f are the time and the frequency index, respectively. To get the intensity of signals, we compute the magnitude of the STFT result as follows:

$$spectrum(t,f) = |STFT(t,f)|. \quad (5)$$

Now we convert the reflected signals into a two-dimensional matrix with time, frequency, and intensity information. To further extract facial geometry features and acoustic impedance features, this matrix is fed into a deep learning model. For ease of training, we transform the matrix into a greyscale spectrogram by the way of square root mapping. Each pixel at the spectrogram position (t, f) describes how each spectral point of the reflected signal is influenced by the user's face over time in terms of frequency and intensity. After the above processing, we obtain a two-dimensional signal spectrogram with sufficient and prominent facial geometry features and acoustic impedance features for further authentication.

VI. User Authentication

So far, we have derived spectrograms from reflected ultrasonic signals, which contain sufficient facial geometry features and acoustic impedance features. Next, we map the facial biometrics into user identities for authentication. Due to the exceptional feature extraction capabilities of the Convolutional Neural Networks (CNN) [13], as well as the effective ability of Convolutional Block Attention Module (CBAM) [14] to enhance CNN's representation power, we develop a CNN with the CBAM to deeply extract facial biometrics. Besides, the obtained spectrograms contain temporal information of frequency and intensity, thus we employ the Long Short-Term Memory (LSTM) [15] after the CNN to capture the temporal characteristics for further facial biometrics extraction. Therefore, we develop a CNN-LSTM-based feature extractor to mine fine-grained facial geometry features and acoustic impedance features. In this section, we first introduce the architecture of our CNN-LSTM model, and then detail the training strategy.

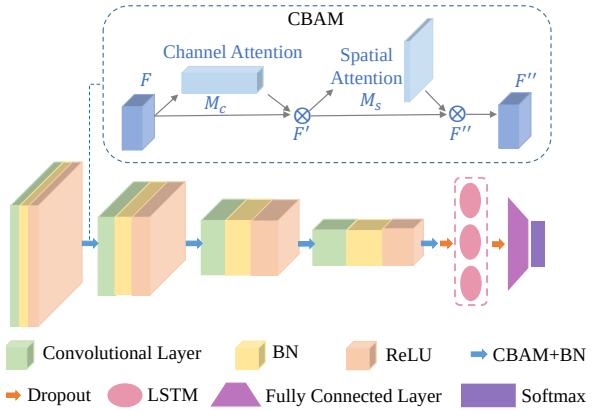


Fig. 4. The architecture of our feature extractor.

A. Architecture of Feature Extractor

We develop a CNN-LSTM-based feature extractor for facial biometrics extraction, comprising four convolutional layers with CBAM, one LSTM layer, one fully connected layer, one softmax layer, and two dropout layers applied after the last convolutional layer and the LSTM layer, respectively. The extractor takes spectrograms as inputs, while outputs probabilities indicating the likelihood that the facial biometrics correspond to each legitimate user. The identity with the highest probability is considered as the final result.

The detailed structure of the CNN with CBAM is shown in Fig. 4. Each convolutional layer is followed by a rectified linear unit (ReLU) and a max pooling layer. The convolutional kernel size and the slide stride are empirically set to 3×3 and 1×1, respectively. The ReLU is utilized to reduce inter-neuronal dependencies, while the max pooling layer is arranged to reduce the feature map size for further feature extraction. The pooling kernel is set to 2×2 empirically.

Following the convolution operations that produce the intermediate feature map from the input spectrogram, we employ the CBAM to refine this feature map by emphasizing meaningful features and suppressing irrelevant ones. To do this, CBAM computes attention maps along two separate dimensions: channel and spatial. These attention maps are then multiplied to the intermediate feature map for adaptive feature refinement. The overall CBAM process can be summarized as:

$$\begin{aligned} F^{'} &= M_c(F) \oplus F, \\ F^{''} &= M_s(F^{'}) \oplus F^{'}, \end{aligned} \quad (6)$$

where F is the intermediate feature map output by former convolution operations, M_c is the 1D channel attention map, M_s is the 2D spatial attention map, $\oplus$ denotes element-wise multiplication, and $F^{''}$ is the final refined output of the CBAM.

Specifically, in the channel attention module, average-pooling and max-pooling operations aggregate spatial information of the intermediate feature map F. The generated descriptors are then processed by a shared multi-layer perceptron (MLP) network and merged using element-wise summation to produce the channel attention map M_c. In the spatial attention module, average-pooling and max-pooling operations aggregate channel information, creating two 2D maps. These maps are then concatenated and processed by a convolution layer to produce the spatial attention map M_s. With channel and spatial attention, CBAM refines the intermediate features, thereby enhancing the feature extractor's capability to capture facial biometrics. After CBAM, we apply a batch normalization (BN) function to stabilize the distribution of feature maps, thus improving the robustness of the feature extractor.

After the two-dimensional CNN with CBAM, the output is fed into the LSTM layer with a hidden size of 128. The LSTM is used to extract the coarse-grained temporal features from the fine-grained features output by the former operations. The temporal features embody the information of frequency and intensity variations over time. By capturing temporal dependencies, the LSTM facilitates coarse-grained feature fusion. Following the LSTM layer is a fully connected layer with 128 units. The fully connected layer linearly combines the output features from the LSTM layer and integrates the local features into global features, thus mapping high-dimensional features to the target class space (i.e., different person IDs). Finally, a softmax layer converts the output of the fully connected layer into a probability distribution, where the class with the highest probability is the predicted result of the corresponding input spectrogram. Note that two dropout layers with a rate of 0.5 are applied after the LSTM layer and the last convolutional layer to mitigate overfitting in the feature extractor, respectively.

B. Training Strategy

To ensure the effectiveness of our feature extractor in capturing facial biometrics, we utilize a supervised learning approach to update parameters. Specifically, we adopt cross-entropy as the loss function, which can be calculated as:

$$CrossEntropy = -\sum_{x}(p(x)\log q(x)), \tag{7}$$

where the probability distribution $p(x)$ and $q(x)$ are the expected output and the actual output, respectively. Besides, we utilize Adam as the optimizer due to its adaptive learning rates, which allow it to adjust learning rates for each parameter individually.

In the registration phase, each user needs to provide several ultrasound samples to train the feature extractor. In the authentication process, spectrograms obtained from ultrasound are fed into the well-trained extractor to mine facial biometrics, which outputs the probabilities that the input spectrograms belongs to each legitimate user. If all probabilities are smaller than a specified threshold, the individual is regarded as an unauthorized intruder. Otherwise, the authentication request is approved and the individual is recognized as a legitimate user corresponding to the identity with the highest probability.

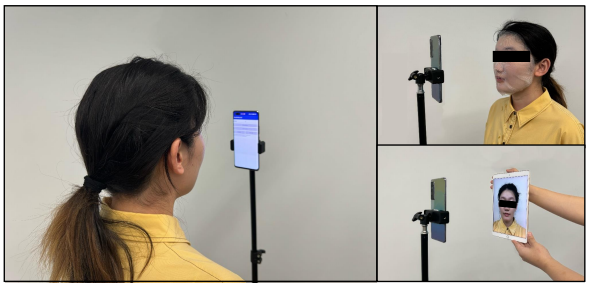


Fig. 5. Experiment setup.

VII. Evaluation

This section first describes the implementation of *UltraFace*, and then details its quantitative performance in real world environment.

Experiment setup. As shown in Fig. 5, we build a proof-of-concept prototype of *UltraFace* to evaluate its authentication and security performance. This prototype is implemented on a COTS smartphone, i.e., HUAWEI nova 7 Pro. We develop an Android App for the smartphone to emit the ultrasonic chirp signals with its built-in speaker and simultaneously record the reflected ultrasound via its built-in microphone. The sampling rate of the acoustic signal is set to 48kHz. All experiments are conducted by adhering to the approval of our university's Institutional Review Board (IRB).

Data collection. We invite 24 volunteers (11 females and 13 males) to participate in our experiments, with heights ranging from 157cm to 185cm and ages ranging from 23 to 28. Among these volunteers, we randomly choose 4 volunteers as illegitimate users and the rest 20 volunteers as legitimate users. We set an acceptance threshold (empirically set to 0.93) to determine users as legitimate or not. Each participant is asked to collect at least 250 signal samples for evaluation. Considering typical smartphone usage habits, the data collection distance between the face and the smartphone is set at about 25 cm.

Metrics. We define four metrics to quantify the performance of *UltraFace*: Authentication Success Rate (ASR), Defense Success Rate (DSR), False Reject Rate (FRR), and False Accept Rate (FAR). ASR indicates the probability that *UltraFace* identifies a legitimate user correctly. It can be represented as:

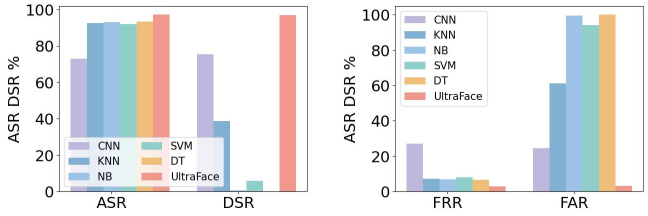
$$ASR = 100\% \times \frac{N_l^{suc}}{N_l}, \tag{8}$$

where N_l^{suc} is the number of successfully accepted legitimate authentication attempts and and N_l is the number of all legitimate authentication attempts. DSR describes the probability that an illegitimate user is successfully detected. It can be calculated as:

$$DSR = 100\% \times \frac{N_{il}^{suc}}{N_{il}}, \tag{9}$$

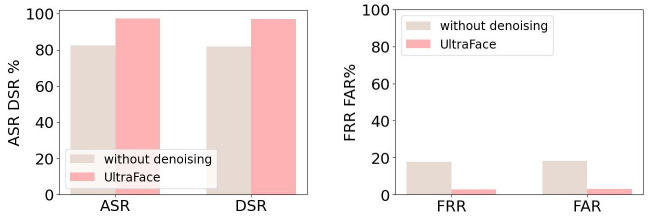
where N_{il}^{suc} and N_{il} are the number of correctly detected illegitimate authentication attempts and the number of all illegitimate authentication attempts, respectively. FRR is the probability that *UltraFace* mistakenly authenticates a legitimate user as an illegitimate one. It can be formulated as:

$$FRR = 100\% \times \frac{N_l^{mis}}{N_l}, \tag{10}$$



(a) ASR and DSR of classifiers. (b) FRR and FAR of classifiers.

Fig. 6. Overall performance of classifiers.



(a) ASR and DSR of different signals. (b) FRR and FAR of different signals.

Fig. 7. Overall performance of without denoising signals and *UltraFace*.

where N_l^{mis} is the number of wrongly rejected legitimate authentication attempts. FAR represents the probability that *UltraFace* falsely accepts an illegitimate user as a legitimate one. It can be expressed as:

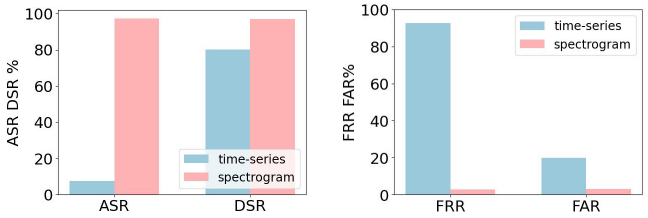
$$FAR = 100\% \times \frac{N_{il}^{mis}}{N_{il}}, \tag{11}$$

where N_{il}^{mis} is the number of wrongly accepted illegitimate authentication attempts. The higher the ASR and DSR, as well as the lower the FAR and FRR, the better the authentication capability and security of *UltraFace*.

A. Overall Performance

We first evaluate the performance of our feature extractor by comparing it with five commonly-used classifiers, including CNN, Support Vector Machine (SVM), Naive Bayes classifier (NB), K-Nearest Neighbours (KNN), and Decision Tree (DT). The dataset is randomly split into [80%, 20%] for data training and testing. Each classifier runs ten times, and we compute the average of these metrics across the ten runs to obtain the final metrics for each classifier. The ASR, DSR, FRR, and FAR of these classifiers are illustrated in Fig. 6. The overall ASR and DSR of our feature extractor are 97.2% and 96.9% respectively, which outperforms other classifiers. Meanwhile, our feature extractor also achieves the lowest FRR and FAR of 2.8% and 3.1%, respectively. These results demonstrate that our feature extractor is able to mine fine-grained facial biometrics to achieve accurate user authentication, whereas traditional classifiers do not possess such ability.

Next, to prove the effectiveness of our denoising method, we compare the performance of *UltraFace* with that of signals without data denoising. As shown in Fig. 7, the ASRs for *UltraFace* and signals without denoising are 97.2% and 82.4%, respectively. Meanwhile, the DSRs of *UltraFace* and signals without denoising are 96.9% and 81.8%, respectively. These results indicate that *UltraFace* can effectively remove noise in the raw signals to realize high-performance authentication.



(a) ASR, DSR of different signals. (b) FRR, FAR of different signals.

Fig. 8. Overall performance of time-series signals and spectrograms.

TABLE I
DEFENSE ABILITY OF *UltraFace*.

Attack	DSR	FAR
Zero-effort attack	100	0
2D spoofing attack	99.6	0.4
3D spoofing attack	99.8	0.2

Lastly, to show the superiority of our spectrogram derivation, we compare its performance with that of time-series ultrasonic signals. As shown in Fig. 8, the ASRs for spectrograms and time-series signals are 97.2% and 7.4%, respectively. Meanwhile, the DSRs of spectrograms and time-series signals are 96.9% and 80.1%, respectively. The comparison results indicate that our signal spectrograms are effective in making facial biometrics more prominent.

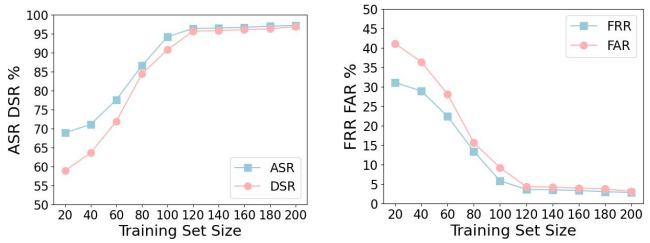
B. Security Study

In this part, we evaluate the security of *UltraFace* against the three attacks introduced in the threat model.

Zero-effort attack. In this attack, the attacker attempts to bypass authentication by simply picking up the smartphone. Since *UltraFace* necessitates facial biometrics to initiate the authentication session, the attacker who is unfamiliar with the principle cannot pass the verification. In the defense experiment, we invite three volunteers as attackers to initiate malicious authentication requests. The result is shown in Table I, where the mean DSR for attackers is 100%. Therefore, *UltraFace* is capable of defending against zero-effort attacks.

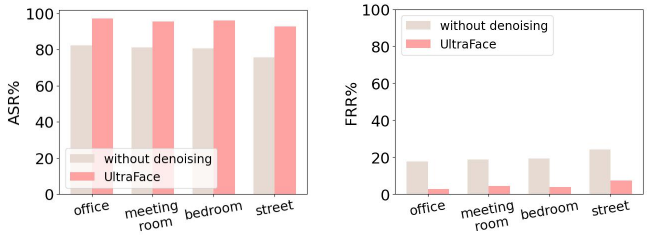
2D spoofing attack. In this attack, the attacker tries to deceive *UltraFace* by using 2D photos or videos of a legitimate user. However, *UltraFace* extracts facial geometry features that are absent in 2D images or videos, rendering the attack ineffective. To assess the defense ability of *UltraFace* against this attack, we use the photos and videos of three legitimate users displayed on an iPad to attack *UltraFace*. As shown in Table I, the mean DSR is 99.6%. Hence, *UltraFace* is able to defend against 2D spoofing attacks effectively.

3D spoofing attack. In this attack, the attacker endeavors to be accepted by *UltraFace* through a replica of a legitimate user's face. Due to the lack of acoustic impedance features in the replica, the attacker will fail to trick *UltraFace*. Given that there are slight differences between the 3D-printed model and the actual user's facial structure, We invite three legitimate users to wear masks to simulate 3D attacks. The experiment result (Table I) shows a mean DSR for the 3D spoofing attack of 99.8%. Thus, *UltraFace* can also defend against 3D spoofing attacks.



(a) Effect of the training set size on ASR and DSR. (b) Effect of the training set size on FRR and FAR.

Fig. 9. Effect of training set size.



(a) ASR of different signals. (b) FRR of different signals.

Fig. 10. Impact of environmental noise.

C. Effect of Training Set Size

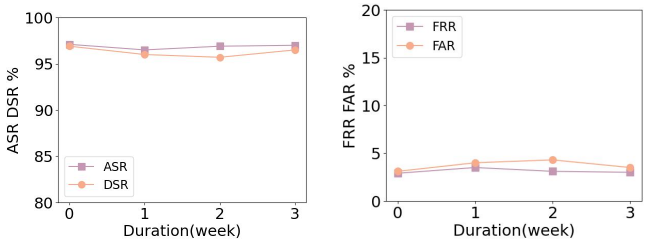
The size of the training set, i.e., the number of training samples (chirp sounds) provided by each user, significantly impacts the user-friendliness of *UltraFace*. Thus, in this experiment, we explore the effect of the training set size on *UltraFace*'s performance. Specifically, we vary the number of training samples per user from 20 to 200 in step of 20. The results are illustrated in Fig. 9. It can be found that the ASRs and DSRs increase as the training set size grows, while the FRRs and FARs decrease. Once the training set size reaches 120, the ASRs and DSRs gradually become saturated. The performance under 140 training samples is only slightly better than that under 120 training samples. Therefore, collecting 140 data for model training is sufficient to ensure a good authentication service. Since collecting 140 chirp sounds only consumes 70 seconds, *UltraFace* is highly user-friendly.

D. Robustness Analysis

In this part, we evaluate the robustness of *UltraFace* across different environments and over extended periods.

Impact of environmental noise. Smartphones are used in various scenarios with diverse environmental noise. To assess *UltraFace*'s robustness against noise interference, we conduct experiments in four typical scenarios: an office, a meeting room, a bedroom, and a street side. As shown in Fig. 10, *UltraFace* maintains high ASRs across various scenarios. Although the ASR drops slightly at the street side, it remains above 92%, significantly higher than the ASRs of the signals without denoising. This suggests that our data denoising method can effectively eliminate noise and enhance the quality of biometrics. More importantly, *UltraFace* is robust against environmental interference.

Long-term observation. We conduct a long-term study for three weeks to verify the stability of the used facial biometrics. Three participants (two legitimate users and one illegitimate user) take part in the experiments. They are tested four times over a three-week period with a one-week interval. As shown in Fig. 11, the ASRs of *UltraFace* are consistently higher than 96%, while the DSRs of *UltraFace* maintain above 95%. This indicates that the performance of *UltraFace* will not degrade significantly over time.



(a) ASR and DSR of different time. (b) FRR and FAR of different time.

Fig. 11. Long-term observation.

TABLE II
COMPARISON WITH EXISTING METHODS.

System	User-imperceptible	One Modality Only	2D Spoofing-resistant	3D Spoofing-resistant
EchoPrint [1]	✓	×	✓	×
EchoFace [7]	×	✓	✓	×
Echo-FAS [8]	×	✓	✓	×
AFace [9]	×	✓	✓	✓
UltraFace	✓	✓	✓	✓

E. Overhead

In this part, we evaluate the storage overhead of *UltraFace*, which primarily comes from three components: the Android APP, the authentication signal, and the feature extractor. Firstly, the Android APP only takes up 3.78MB. Besides, one authentication signal sample consumes about 48KB. As for the feature extractor, it occupies about 16.9MB. Thus, the total storage overhead of *UltraFace* is approximately 21MB, which is far less than the smartphone's memory capacity. Therefore, the storage overhead of *UltraFace* is acceptable.

F. Comparison with Existing Methods

We compare *UltraFace* with four state-of-the-art acoustic-based facial authentication methods on smartphones: EchoPrint [1], EchoFace [7], Echo-FAS [8], and AFace [9], in terms of user-friendliness, sensor modality, and spoofing resilience. As depicted in Table II, EchoFace, Echo-FAS, and AFace utilize acoustic signals within the audible range for authentication, which would cause discomfort to users. In contrast, *UltraFace* emits imperceptible ultrasound, enhancing its usability and user-friendliness. Besides, EchoPrint requires the assistance of cameras, whereas *UltraFace* can work independently of cameras, making it effective in poor illumination conditions. Additionally, the five methods can defend against 2D spoofing attacks. However, EchoPrint, EchoFace, and Echo-FAS only capture the 3D structure of the user's face, leaving them still vulnerable to 3D spoofing attacks. But *UltraFace* can extract both facial geometry features and acoustic impedance features, making it resistant to 3D spoofing. In summary, compared with existing methods,

UltraFace stands out as the only system enabling secure and user-friendly authentication on COTS smartphones.

VIII. Related work

This work is mainly related to two kinds of techniques: ultrasonic sensing and facial authentication on smartphones.

A. Ultrasonic Sensing

Ultrasound is widely utilized in various applications due to its inaudibility, accessibility, and low-cost. For example, ultrasound can be used for recognizing finger/hand gestures. RobuCIR [16] proposes a contact-free gesture recognition system based on ultrasonic sensing. VSkin [17] utilizes ultrasound to capture finger tapping and movements. Besides, ultrasonic sensing enables devices to support health monitoring. UltraMotion [18] employs the paired smartwatch and smartphone for real-time arm motion tracking. Ubi-Asthma [19] implements an asthma detection system on the smartwatch using ultrasonic signals to obtain health-related breathing data. Different from previous works, we leverage ultrasound to capture user's facial biometrics to achieve user-friendly authentication.

B. Facial Authentication on Smartphones

As indispensable devices in daily life, smartphones are frequently accessed for personal information, thus necessitating user-friendly and secure facial authentication. However, existing facial authentication methods have shortcomings. For example, camera-based methods require good lighting [5] and are susceptible to 2D spoofing attacks [6]. To achieve reliable and secure facial authentication, previous works have explored acoustic sensing techniques. Echoprint [1] combines the camera with microphone and speaker to capture joint features. Echoface [7] and Echo-FAS [8] use acoustic signals to extract facial 3D structures. AFace [9] captures facial structure and biological material features. However, these methods either require cameras or remain susceptible to 3D spoofing attacks. Additionally, some methods emit audible signals, causing users discomfort. Different from prior approaches, *UltraFace* achieves spoofing-resistant and user-friendly authentication by collecting facial geometry features and acoustic impedance features through imperceptible ultrasound.

IX. Conclusion

In this paper, we propose a spoofing-resistant and user-friendly authentication system on smartphones, namely *UltraFace*. It extracts facial geometry features and acoustic impedance features from ultrasonic signals reflected off the face to identify users. To do so, we initially build a theoretical model to figure out the correlation between the facial biometrics and the ultrasound reflection. Based on this model, we perform frequency-domain analysis on the signal to get spectrograms. Then, we employ a deep learning model to mine fine-grained facial biometrics from the spectrograms and achieve accurate user authentication. Extensive experiments show that *UltraFace* can achieve an authentication success rate of 97.2%. Meanwhile, *UltraFace* is secure and robust for long-term use.

Acknowledgement

This paper is supported by the National Natural Science Foundation of China under grant No. U21A20462 and 62372400, "Pioneer" and "Leading Goose" R&D Program of Zhejiang under grant No. 2023C01033, and the Postdoctoral Fellowship Program of CPSF under grant No. GZC20241488.

References

[1] B. Zhou, J. Lohokare, R. Gao, and F. Ye, "Echoprint: Two-factor authentication using acoustics and vision on smartphones," in *Proceedings of the ACM International Conference on Mobile Computing and Networking (MobiCom)*, 2019.

[2] GMSA, "The mobile economy 2024," https://www.gsma.com/solutions-and-impact/connectivity-for-good/mobile-economy/.

[3] S. Rajarajan and P. Priyadarsini, "UTP: a novel PIN number based user authentication scheme," *The International Arab Journal of Information Technology (IAJIT)*, vol. 16, no. 5, pp. 904–913, 2019.

[4] F. Tari, A. A. Ozok, and S. H. Holden, "A comparison of perceived and real shoulder-surfing risks between alphanumeric and graphical passwords," in *Proceedings of the ACM Symposium on Usable Privacy and Security (SOUPS)*, 2006.

[5] D. Ai, W. Fan, K. Jia, M. Lu, and Y. Liu, "A method of dual-spectrum feature fusion for face recognition under non-ideal lighting conditions," in *Proceedings of the ACM International Symposium on Signal Processing Systems (SSPS)*, 2022.

[6] W. Xu, J. Liu, S. Zhang, Y. Zheng, F. Lin, J. Han, F. Xiao, and K. Ren, "Rface: Anti-spoofing facial authentication using COTS RFID," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)*, 2021.

[7] H. Chen, W. Wang, J. Zhang, and Q. Zhang, "Echoface: Acoustic sensor-based media attack detection for face authentication," *IEEE Internet of Things Journal*, vol. 7, no. 3, pp. 2152–2159, 2020.

[8] C. Kong, K. Zheng, S. Wang, A. Rocha, and H. Li, "Beyond the pixel world: A novel acoustic-based face anti-spoofing system for smartphones," *IEEE Transactions on Information Forensics and Security (TIFS)*, vol. 17, pp. 3238–3253, 2022.

[9] Z. Xu, T. Liu, R. Jiang, P. Hu, Z. Guo, and C. Liu, "Aface: Range-flexible anti-spoofing face authentication via smartphone acoustic sensing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (UbiComp)*, vol. 8, no. 1, pp. 26:1–26:33, 2024.

[10] R. Wang, L. Huang, and C. Wang, "Low-effort VR headset user authentication using head-reverberated sounds with replay resistance," in *Proceedings of the IEEE Symposium on Security and Privacy*, 2023.

[11] G. Hsieh and J. C. Hung, "Phase-locked loop techniques. A survey," *IEEE Trans. Ind. Electron.*, vol. 43, no. 6, pp. 609–615, 1996.

[12] X. Fang, J. Liu, Y. Chen, X. Xu, and J. Han, "Wristpass: Secure wearable continuous authentication via ultrasonic sensing," in *Proceedings of the IEEE International Symposium on Quality of Service (IWQoS)*, 2024.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2012.

[14] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," in *European Conference on Computer Vision (ECCV)*. Springer, 2018.

[15] R. C. Staudemeyer and E. R. Morris, "Understanding LSTM - a tutorial into long short-term memory recurrent neural networks," *CoRR*, vol. abs/1909.09586, 2019.

[16] Y. Wang, J. Shen, and Y. Zheng, "Push the limit of acoustic gesture recognition," *IEEE Transactions Mobile Computing (TMC)*, vol. 21, no. 5, pp. 1798–1811, 2022.

[17] K. Sun, T. Zhao, W. Wang, and L. Xie, "Vskin: Sensing touch gestures on surfaces of mobile devices using acoustic signals," in *Proceedings of the ACM International Conference on Mobile Computing and Networking (MobiCom)*, 2018.

[18] X. Niu, K. Zou, D. Shen, S. Drew, S. Wu, G. Guo, and R. Chen, "Ultramotion: High-precision ultrasonic arm tracking for real-world exercises," *IEEE Transactions Mobile Computing (TMC)*, vol. 23, no. 2, pp. 1846–1862, 2024.

[19] Y. Wu, J. Zhang, Y. Chen, J. Wang, W. Shi, and Q. Zhang, "Ubi-asthma: Toward ubiquitous asthma detection using the smartwatch," *IEEE Internet Things J.*, vol. 10, no. 13, pp. 11 576–11 587, 2023.