electronics

MDPI

Article

MetaEar: Imperceptible Acoustic Side Channel Continuous Authentication Based on ERTF

Zhuo Chang [1,2], Lin Wang [1], Binbin Li [3] and Wenyuan Liu [1,*]

[1] School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China
[2] School of Cyber Security and Computer, Hebei University, Baoding 071000, China
[3] School of Economics and Management, Yanshan University, Qinhuangdao 066004, China
* Correspondence: wyliu@ysu.edu.cn

Abstract: With the development of ubiquitous mobile devices, biometrics authentication has received much attention from researchers. For immersive experiences in AR (augmented reality), convenient continuous biometric authentication technologies are required to provide security for electronic assets and transactions through head-mounted devices. Existing fingerprint or face authentication methods are vulnerable to spoof attacks and replay attacks. In this paper, we propose *MetaEar*, which harnesses head-mounted devices to send FMCW (Frequency-Modulated Continuous Wave) ultrasonic signals for continuous biometric authentication of the human ear. CIR (channel impulse response) leveraged the channel estimation theory to model the physiological structure of the human ear, called the Ear Related Transfer Function (ERTF). It extracts unique representations of the human ear's intrinsic and extrinsic biometric features. To overcome the data dependency of Deep Learning and improve its deployability in mobile devices, we use the lightweight learning approach for classification and authentication. Our implementation and evaluation show that the average accuracy can reach about 96% in different scenarios with small amounts of data. *MetaEar* enables one to handle immersive deployable authentication and be more sensitive to replay and impersonation attacks.

Keywords: acoustic sense; continuous authentication; Ear Related Transfer Function; FMCW; CIR

Citation: Chang, Z.; Wang, L.; Li, B.; Liu, W. MetaEar: Imperceptible Acoustic Side Channel Continuous Authentication Based on ERTF. *Electronics* **2022**, *11*, 3401. https://doi.org/10.3390/electronics11203401

Academic Editor: Akshya Swain

Received: 27 September 2022
Accepted: 19 October 2022
Published: 20 October 2022

1. Introduction

Ubiquitous smart devices, such as smart phones, smart headphones, and smart watches, with their rich built-in sensors, have become an important access carrier for the Cyber–Physical Human System (CPHS) [1]. On the one hand, the requirements of work and entertainment during the mobile process make wearing headphones a daily behavior, which spawns a new computing model, earable computing [2]. On the other hand, taking mobility, ubiquity, and human-centricity into account, CPHS and even the Metaverse have higher requirements for continuous identity authentication [3,4]. The global biometric authentication market size is projected to reach 30.5 billion U.S. dollars by 2026 [5].

Digital assets and commercial transactions require highly secure and reliable continuous authentication to protect user property. These transactions require ongoing multi-factor authentication to ensure their security and validity. In addition, continuous, safe and efficient biometric authentication can also be applied to scenarios such as telemedicine, disabled services, and virtual production.

Biometric authentication has the advantages of stochastic variation within the different people and no dependence on shared secrets. The existing face, fingerprint, iris, and voiceprint authentication methods are based on extrinsic biometrics. Authentication based on face, fingerprint, and iris is vulnerable to presentation attacks and also raises privacy concerns [6], using a photo [7], video [8], mask [9], or silicone fingertip [10] to impersonate a victim. In addition, the voiceprint-based authentication method is vulnerable to replay attacks [11,12] of voice recording. So, can we design a continuous and transparent biometric authentication method to ensure the system's security?

At present, ultrasonic authentication through the human ear is a novel solution. The head-mounted device sends ultrasonic sound to the human ear and performs continuous identity authentication through the feedback of the human ear. Collecting the FMCW (Frequency-Modulated Continuous Wave) ultrasonic signals reflected from the human ear uses the signal channel characteristics to describe the human body's anatomy and extract the unique biological features [13] for authentication. The inaudible ultrasound can perform continuous passive perceptual authentication without affecting human hearing and without interfering with the user's immersive experience, releasing the user from the authentication frequent interaction process. Furthermore, unlike the existing face identification and fingerprint authentication, ultrasonic authentication can prevent counterfeiting and replay attacks, thus ensuring the security of transactions in Metaverse. On this basis, we aim to build an explainable acoustic authentication model that can be applied in daily scenarios with zero effort and high accuracy.

However, we face three major technical challenges to achieving a one-fits-all model. First of all, the features originally used in acoustic action recognition, such as doppler [14] and phase [15], are aimed at dynamic behaviors and cannot be used to describe static object features. So for the uniqueness of the biological structures of auricles and ear canals, how do we design a model that extracts the unique biological geometric features of static auricles and ear canals? Second, the FMCW acoustic signal received by two microphones in real time is not synchronized. How can the multi-microphone signal be effectively synchronized? Third, ultrasound is a vibration wave, and for a monostatic sensing device with co-located transceivers, vibration will cause self-interference of the received signal. How do we remove the self-interference and extract high-resolution characteristic signals?

To overcome these challenges, we propose *MetaEar*, a continuous authentication system based on imperceptible acoustic fingerprints, which uses FMCW ultrasound to model the unique biometrics of the human ear and conduct authentication, as shown in Figure 1. The key component of *MetaEar* is ERTF (Ear Related Transfer Function), which uses ERTF to identify and extract the unique biometrics of the human ear. Our observation is that each person's auricle and the ear structure have different responses to the delay and magnitude of FMCW ultrasound. The auricle and ear canal modulate the sound signal to produce different delays. The eardrum and the cochlea convert the different magnitude mechanical sound waves to neural electrical signals.

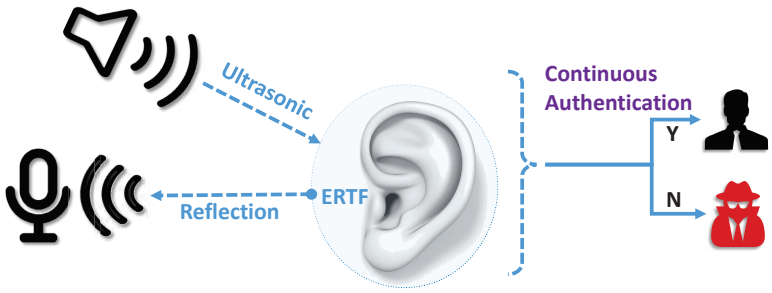


Figure 1. MetaEar, an acoustic side channel continuous authentication system.

Furthermore, the channel impulse response of the human ear can be modeled to extract the ear's unique features and authenticate it. The speaker on the headset sends inaudible FMCW ultrasonic waves, and the co-located microphone receives the reflected sound signal. After passing the band-pass filter, the phase slope change is harnessed to align the signals due to the problem of fuzzy peak judgment of cross-correlation. MetaEar employs a dual-microphone differential denoising to eliminate the self-interference caused by the co-located physical transmission of the vibration wave. Finally, the channel impulse response of the signal is calculated by modeling the ERTF so that the unique biometric vector of the human ear can be achieved. This process is equivalent to model-based human ear anatomy feature embedding. For better application deployment, instead of using a Deep Learning model that requires numerous data, we utilize a traditional SVM (Support Vector

Machine) to perform one-class authentication. Due to *MetaEar*'s modeling of the complex structure of the human ear and the relatively slow transmission speed and low power of ultrasonic signals, it prevents co-located signal collection and resists counterfeiting attacks and replay attacks.

In a nutshell, our core contributions are three-fold.

- We propose the *MetaEar* system, which uses the ultrasonic reflection signals of the auricle and ear canal to continuously authenticate users with ERTF, which effectively prevents replay attacks and counterfeiting attacks.
- We design a characteristic function to represent the auricle and ear canal biometric features through the principle of the impulse response of channel estimation, which effectively expresses the unique biological characteristics of the human auricle and ear canal.
- We build a prototype of *MetaEar* with commercial off-the-shelf smartphones and evaluate its effectiveness and security in different settings. Extensive experiments show that it authentication accuracy over 96.8%.

The rest of this paper is organized as follows. In Section 2, we review the related works. Section 3 introduces the adversary model. Section 4 gives the mathematical derivation on how ERTF could model the unique feature of the human ear, and demonstrates the feasibility study. Section 5 presents the MetaEar architecture design. Section 6 is the implementation setup. Section 7 show the evaluation result, followed by a conclusion in Section 8.

2. Related Works

2.1. Earable Computing

The era of earable computing is coming, and "earable" [2,16] refers to wearable devices around the ears, such as mobile phones, headsets, headphones, and smart glasses. These devices can utilize acoustic signals to interact with people, such as listening to music, sensing oral activities [14,17,18], and even using the ear canal for authentication [19–21].

2.2. Acoustic Authentication

Acoustic waves can measure temperature [22], tracking [23–25], gesture sensing [26,27], and activity recognition [28] to breathe monitoring [15]. The authentication based on the sound signal mainly uses the acoustic signal to extract corresponding unique biometric features to achieve authentication, including using audible sounds, such as voiceprints [29], for authentication. Some use the FMCW ultrasonic signal to extract the feature of teeth actions for authentication [20,30], and some use the sound signal to extract the characteristics of the throat movement [31] for authentication. The FMCW ultrasonic signal is also used for lip motion [32] authentication or face liveness [33] detection.

2.3. Continue Authentication

As augmented reality evolves, continuous authentication ensures system security without interfering with the user's immersive experience. Some use WiFi signals for continuous authentication [34,35], and the sensitivity of RF signals to location and orientation seriously reduces generalization. Some existing works using heart rate [36–38] and respiration [39,40] require the user to remain motionless, prone to severe interference from multipath environments, and cannot be applied to multi-person scenarios. Behaviors [41] can also be used for continuous authentication but are vulnerable to impersonation attacks. Continuous authentication uses eye movement [42,43] but is sensitive to ambient light and requires high equipment costs. There is also work using the ear canal [19,21], but the types of equipment are all earbuds, which will not only cause irreversible damage to the user's hearing but also have higher needs on the depth and position of the earplugs in the ear canal.

3. Threat Model

We assume that the attackers are not resourceful in the headset hardware. The attack succeeds if the attacker can implant malware in the headset and obtain any private data he wants. In other words, augmented reality's software and hardware resources are secure. Based on the above assumptions, two types of adversarial behaviors are considered below.

Replay attack: The replay attack scenario in which the adversary is physically near the victim and his/her enrolled handset, such as in a crowded campus cafe or public vehicle. The collected signals can be disguised as the victim entering the system to invade the victim's Metaverse property and privacy [44].

Impersonate attack: The adversary sends and receives sound wave signals from its ears in different modes, impersonating the victim, intending to deceive the security authentication into attacking the system. Alternatively, various silicone bionic materials are used to imitate the biological structure of the human body, for example, imitating faces and imitating fingerprints with silicone. However, both faces and fingerprints are explicit biometrics that can be easily forged. The counterfeiting attack uses forged 3D printed artificial dummy ears of particular anthropomorphic materials to complete the invasion. It deceives the system to destroy and take the virtual property of the owner [45,46].

4. ERTF Model

4.1. Ear Structure

The human ear is an essential auditory organ of the human body and has unique biological characteristics that are different for each person [47]. The human ear structure is divided into three parts, as shown in Figure 2, in which the auricle and the ear canal collect and transmit sound. The human ear canal is about 25∼35 mm long and about 8 mm diameter. The eardrum and cochlea are the perceptions of sound. Different people have different geometric shapes of the external auricle and ear canal, and the response of the eardrum and cochlea to sound is unique for different people. Therefore, continuous authentication can be realized by modeling the characteristics of the ear's intrinsic physiological structure and extrinsic geometric shape.

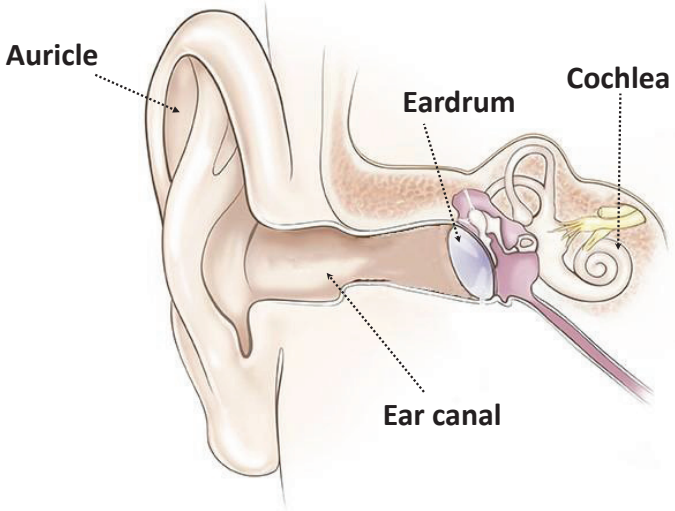


Figure 2. The physiological structure of human ear.

4.2. ERTF

In Metaverse, in addition to vision, sound immersion is also essential, allowing users to maintain the immersion of sound and perceive spatial positioning. Individualized or generic HRTFs (Head Related Transfer Functions) are usually employed to render spatialized sounds within an AR headset.

HRTF is a phenomenon that describes how an ear and head receive sound from a sound source. ERTF is very similar but describes the ear profile to the FMCW acoustic

signal. Most notably, the shape of the pinna, the length and the diameter of the ear canal, and the subtle differences in biological properties influence the incoming ultrasonic signal by boosting some delays and phases. When the reflected sound signal reaches the microphone, the intrinsic unique biometric features of the human ear can be collected. So ERTF is the change of the sound's response profile to the unique characteristics of the user's ear. It is mainly produced by the pinna, ear canal, and cochlea, and we define it as ERTF:

$$h_{\mathrm{ERTF}}(t) = IFFT(H(f)). \tag{1}$$

where $H(f)$ is the channel frequency response produced by the disparate anatomy of the ear, and IFFT is Inverse Fast Fourier Transform.

We imagine the whole human acoustic sensing process as communication through an RF signal. The acoustic sensing channel can be modeled as a linear time-invariant system, which effectively models propagation delay and signal attenuation along multiple propagation paths. So, all the channel parameter changes by the user's organic textures are modeled for the channel state estimation. In this way, we could easily extract the unique features of the subject's ear.

To achieve that, we borrow the idea from channel estimation to determine the fading and path loss of the wireless channel. The received signal can be mathematically represented as $r(t) = h(t) \times s(t)$, where $h(t)$ represents the CIR (Channel Impulse Response) of the acoustic channel; $r(t)$ and $s(t)$ represent the received signal and transmitted signal, respectively. The linear frequency of the FMCW acoustic signal is expressed as:

$$f(t) = f_c + \frac{Bt}{T}, \tag{2}$$

where T is FMCW chirp duration time, B is the chirp bandwidth, and f_c is initial frequency.

Phase $u(t)$ is derived as:

$$u(t) = \int_0^t f(t')dt' = 2\pi(f_c t + B\frac{t^2}{2T}), \tag{3}$$

So the send signal is:

$$s_{FMCW}(t) = e^{-j2\pi\left(f_c t + \frac{B}{2T}t^2\right)}, \tag{4}$$

Then the frequency domain expression is:

$$S_k = \sum_{i=0}^{N-1} e^{-j2\pi\left(f_c t + \frac{B}{2T}t^2\right)} e^{-j\frac{2\pi}{N}kn}. \tag{5}$$

Ultrasonic sound signals are transmitted in the ear canal and inner ear in a multipath environment and reflected in the microphone for the reception. Assuming there are P paths, the delay of each path is $\tau_i, i \in [1, P]$, and the multipath delay of the received signal is:

$$r_{FMCW}(t) = \sum_{i=1}^{P} \theta_i e^{-j2\pi\left(f_c(t-\tau_i) + \frac{B}{2T}(t-\tau_i)^2\right)}, \tag{6}$$

where θ_i is attenuation coefficient. The frequency-domain formula is:

$$\begin{aligned} R_k &= \sum_{n=0}^{N-1} \sum_{i=1}^{P} \theta_i e^{-j2\pi\left(f_c(t-\tau_i) + \frac{B}{2T}(t-\tau_i)^2\right)} e^{-j\frac{2\pi}{N}kn} \\ &= \sum_{i=1}^{P} \theta_i S_k e^{-j\frac{2\pi}{N}k\tau_i}, \end{aligned} \tag{7}$$

S_k represents the transmitted acoustic signal. CIR-based ERTF is:

$$h_{\mathrm{ERTF}}(t) = IFFT\left(\frac{R_k}{S_k}\right) = IFFT\left(\sum_{i=1}^{P}\theta_i e^{-j\frac{2\pi}{N}k\tau_i}\right)$$
$$= \sum_{n=0}^{N-1}\sum_{i=1}^{P}\theta_i e^{-j\frac{2\pi}{N}k\tau_i}e^{j\frac{2\pi}{N}kn} \quad (8)$$
$$= \sum_{n=0}^{N-1}\sum_{i=1}^{P}\theta_i e^{-j\frac{2\pi}{N}k(n-\tau_i)}.$$

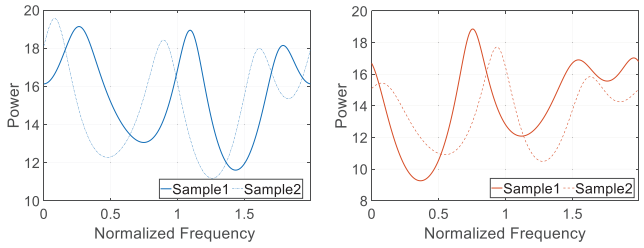
In practice, CIR measurement is represented with a set of complex values. Each complex value measures the channel information of a specific propagation delay range and the corresponding magnitudes and phases of the CIR. It can be seen from Equation (8) that ERTF is related to attenuation θ_{i} and time delay τ_i. θ_{i} expresses the magnitude attenuation caused by the conversion of sound mechanical waves into electronic signals by the intrinsic biological structure, and τ_i expresses the multipath delay caused by the extrinsic geometric structure. That is why ERTF can express the human ear's physiological and physical unique characteristics.

4.3. Feasibility Analysis

In a quiet hall, we used the Samsung NEXUS 6 mobile phone to collect the data on the FMCW sound signals. The frequency band of the ultrasonic wave is 18∼22 kHz. By calculating the received signal's PSD (Power Spectral Density) $P_r(\omega)$ of different subjects, as shown in Figure 3, it can be seen that the features can be distinguished.

$$P_r(\omega) = \left|H_{ERTF}\left(\mathrm{e}^{\mathrm{j}\omega}\right)\right|^2 P_s(\omega). \quad (9)$$

Because in the channel estimation, for the autoregressive time series model, the relationship between the PSD and the ERTF amplitude is as in Equation (9), we can show the unique features of different subjects by the PSD.



(**a**) PDF of user1 (**b**) PDF of user2

Figure 3. Feasibility analysis by power spectral density.

5. System Design

5.1. Overview

MetaEar is a system for continuous biometric-based authentication using FMCW ultrasound. As shown in Figure 4, the system consists of four modules, data collection, signal processing, feature extraction, and authentication. The registration process is the same as the authentication process. First, the speaker sends out FMCW ultrasonic waves with a time slot facing the human ear, and the two microphones receive the reflected chirp signal.

MetaEar contains two primary parts: signal processing and feature extraction modules. Since the device's microphones have a hardware startup time, the front empty window period caused by the hardware delay should be removed first. Then, for the co-located transmitting and receiving devices, the self-interference cancellation is performed using the differential technique of the dual microphones. After denoising, it is critical to align each chirp to perform feature extraction. Here we use a method based on the phase slope

so that the two received signals can be satisfactorily aligned. We only sample the signal within the corresponding frequency band and carry out a Hanning filter to remove the noise signal outside the frequency band. There is a particular time slot between every two chirps, and the actual signal segment of each chirp needs to be extracted. The segmentation is implemented through a power spectrum envelope. Subsequently, the segments of these chirps are directly input into the feature extraction module.

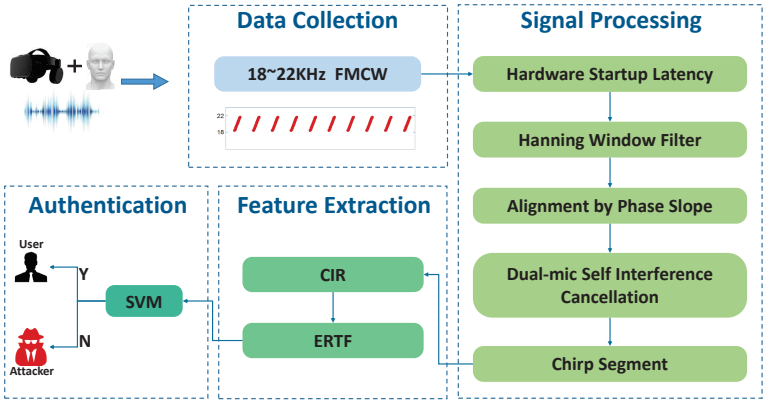


Figure 4. The authentication framework of *MetaEar* consists of four components: data collection, signal processing, feature extraction, and authentication.

The feature extraction module mainly uses the proposed ultrasonic chirp of a specific frequency band to perform feature extraction. First, the calculation of the channel frequency response is fulfilled. Secondly, the channel impulse response is converted into the channel impulse response through IFFT. The uniqueness of the human ear's geometric structure and endoplasmic structure can be represented as the channel impulse parameters of the sound transmission channel. Then PCA (Principle Component Analysis) extracts principal components to form feature vectors. These feature vectors are fed to the SVM of the authentication module for one-classification to achieve efficient continuous authentication.

5.2. Acoustic Design

On the one hand, the ultrasonic signal cannot produce audible noise that interferes with the user. Due to the limitation of mobile phone hardware, the highest sound frequency of prevalent mobile phones is 24 kHz, and the maximum sample rate is 48 kHz. In order to increase the bandwidth of the sound signal, we employ the ultrasonic signal above 18 kHz. The FMCW ultrasonic signal frequency is 18∼22 kHz, FC = 18 kHz, B = 4 kHz, and the wavelength is 16∼21 mm. The ear canal length is about 25∼35 mm. Furthermore, combined with the reflection distance, the overall length can reach the range of 1∼2 sound wavelengths, which naturally satisfies the near-field requirements.

As shown in Figure 5, the chirp duration of FMCW ultrasound is 10 ms because the smaller the FMCW period, the higher the range resolution. To avoid ISI (Inter Symbol Interference) between each chirp, we added two 10 ms idle time slots between every two chirps. The time delay between the transmitted signal and the received signal is τ.

5.3. Denoise

Hardware delay elimination. Because transceivers of mobile phones have a hardware run-up time, the 10 ms period FMCW sound signal is sensitive to the short startup time. Therefore, we need to remove the empty sampling points at the beginning of the acquisition signal. According to experience, this part of the time is 500 samples, which is 500/48 = 10.4167 ms.

Ambient noise. Typically, common noise in the daily circumstances declines sharply in the frequency band above 8 kHz and lower than 18 kHz [48]. Therefore, we use the Hanning window filter for the corresponding filtering and only maintain the signal within

18∼22 kHz, as shown in Figure 6, to remove the high- and low-frequency noise in the environment and reduce the frequency leakage.

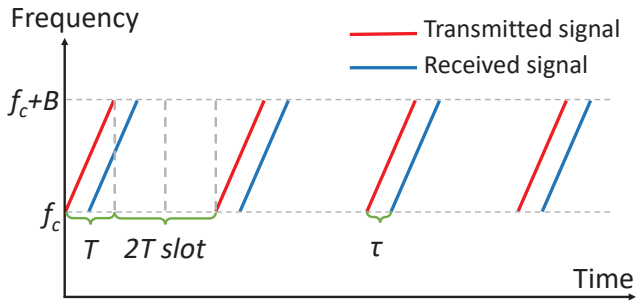


Figure 5. The FMCW chirp signal design.

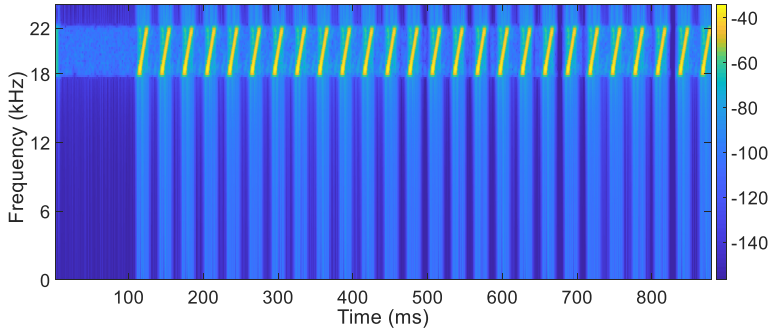


Figure 6. The frequency domain denoised FMCW chirp signal.

Dynamic interference. Because the sound wave propagates slowly, the transmission power is weak. It attenuates quickly, the propagation distance is exceptionally compact, and the transmission and reception area is close to the ear. There is almost no need to remove the interference of moving objects in the long-distance environment to the signal. For very close moving objects, such as other people doing random behaviors close to the user's ears, since the action frequency is generally lower than 18 kHz, the above Hanning bandpass filter can filter out this part of the noise.

5.4. Synchronization

The two microphone channels of sound signals have a time difference due to the distance between the two microphones, so time alignment and synchronization are required. Time alignment is obtained by transforming the signals into the frequency domain and applying the linear phase shifts corresponding to the time delay.

Existing [49] work in RF sensing uses two transmit antennas to cancel the direct path signal at the receiver. However, this approach does not work for acoustic-based ERTF. The reasons are two-fold. First, a smartphone typically has two pairs of co-located speakers and microphones. Playing sounds with any speaker may saturate the corresponding microphone by directly transferring the acoustic signal, making it infeasible to sense reflections. Second, the speakers on the smartphone are designed for different usages (e.g., communication, playing sound) and thus are highly heterogeneous. It is hard to perform equalization on a commercial audio system for FMCW chirp signals.

Instead, we leverage two microphones available on smartphones to achieve self-interference cancellation. Specifically, suppose one speaker plays the FMCW signal, and two microphones receive $r_1(t)$ and $r_2(t)$, respectively. Then ERTF estimates the subsample delay ε_t with the phase slope changing in the frequency domain and further calculates the correlation between two aligned signals [37].

$$\begin{aligned} \varepsilon_t &= \min_{\varepsilon} |\angle(\mathcal{F}[r_1(t)]\mathcal{F}^*[r_2(t)]) + 2\pi f\varepsilon| \\ r_2^{\text{shift}}(t) &= \mathcal{F}^{-1}\left[\mathcal{F}[r_2(t)] \cdot e^{-j2\pi f\varepsilon_t}\right], \end{aligned} \tag{10}$$

where $\mathcal{F}[\cdot]$ denotes Fourier transform. Since the direct signal from the speaker to the microphones is the most potent component, we can approximate the estimation above as the delay and amplitude ratio of the direct acoustic of the two microphones.

5.5. Dual-Mic Subtraction

After alignment, we want to remove the self-interfering signal received by the microphone due to direct physical propagation. Since this part of the signal propagates through solid hardware, the propagation speed is 15 times the speed in the air, so the first received signal component is the self-interference signal. The most commonly used method utilizes autocorrelation [39,50] to eliminate the self-interference through the corresponding tap peak search. However, due to autocorrelation, there will be the disadvantage that the tap peak is ambiguous. The distance is too close to the regular reflection signal peak, so the blur is too high to distinguish the peak of the direct path. We use the dual microphone cancellation method. Differentiating the signals of the two microphones after alignment can eliminate the self-interference signal and retain the relevant dynamic and adequate information about the ERTF. For the noise in the environment, both microphones can receive it, so the dual-microphone differential method can also remove the ambient noise. From the above, we know that:

$$\rho = \frac{r_1(t) \cdot r_2^{\text{shift}}(t)}{|r_1(t)||r_2^{\text{shift}}(t)|}, \tag{11}$$

Thus, ERTF scales $r_2^{\text{shift}}(t)$ with ρ, and subtracts it from $r_1(t)$:

$$r(t)^{\text{cancel}} = r_1(t) - \rho r_2^{\text{shift}}(t), \tag{12}$$

Then the signals from each microphone are time synchronization. The synthesized signal in the time domain is formulated below.

$$r(t)^{\text{cancel}} = \frac{1}{N}\sum_{i=1}^{N} \theta_i \omega_i x_i(t - \Delta t_{\varepsilon_t}). \tag{13}$$

where Δt_{ε_t} indicates the time delay.

5.6. Chirp Segment

After completing the above procedures, each cycle transmitted FMCW sound signal has been segmented and extracted. Because the guard gap slot in the middle is futile, it is only necessary to calculate the ERTF for the transmit and receive chirps in each cycle. After trying several segment methods such as VAD (Voice Activity Detection), variance, and power accumulation, we eventually chose signal envelope, which is more efficient in computation. First, find all peaks in the signal, and obtain local maxima that are separated by N points at least, then use spline interpolation to return the peak envelope of the signal, as formulated in Equation (14).

$$Envelope_{r(t)} = \left| r(t)_{\text{cancel}} \right|. \tag{14}$$

After the power spectrum envelope is obtained, we set the threshold to be the overall expectation and detect the start and end points of the chirp signal of each cycle. The algorithm is shown in Algorithm 1.

Algorithm 1: Chirp Segmentation.

Input: Denoised & Synchronized FMCW Sequence: $r[n]$
Output: Segmented Chirp Sequence: $Chirp\_Matrix[n]$

// Get the envelope of r[n].
$Peak[i] = FindTop(r[i])$;
$Envelope[i] = Interpolation(Peak[i])$;

// Find peaks of envelope[i].
$up\_peaks[j] = FindPeaks\_Envelope(Envelope[i])$;
$threshold = Average(r[n])$;

// Detect start/end points of chirp segment .
$m = n = up\_peaks[i]$;
while $r[m] >= threshold$ **do**
  $m--$;
end
$start\_point[k] = m$;
while $r[n] >= threshold$ **do**
  $n++$;
end
$end\_point[k] = n$; $Segments = Fusion(start\_point, end\_point)$;
$Chirp\_Matrix[n] = Segments[0:19]$;

The result is shown in Figure 7, which can segment each chirp ideally.

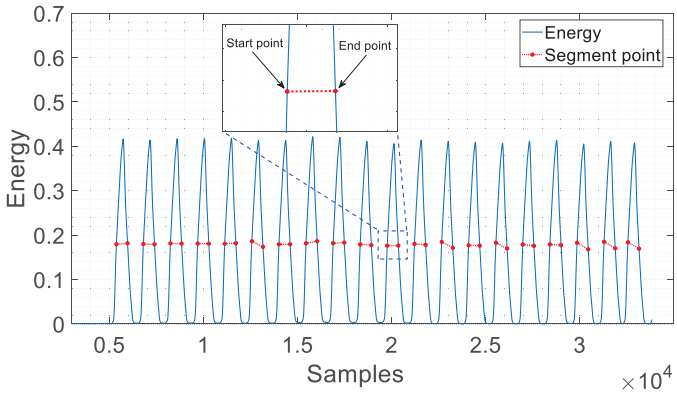


Figure 7. The energy envelope and boundary points of segmentation.

5.7. ERTF

Previous works [51,52] have demonstrated that the microphone's frequency response (especially in the ultrasonic band) is a stable and feasible feature over time, even sensitive enough to distinguish among millions of smartphones of the same brand and model. Considering it from another view, attenuation and time delay still express the geometric structure and intrinsic biological features of the auricle and ear canal besides the frequency response. The overall biologically unique characteristics of the expression occupy a principal place.

However, the time-domain equalization calculation complexity is too high, and the result is not accurate enough. In addition, the CIR calculation in the time domain utilizes a complex cross-correlation calculation. Moreover, most acoustic signal processing is implemented in the frequency domain. In order to avoid multiple conversions between the frequency and time domain and obtain the corresponding CIR using FFT (Fast Fourier Transform) for high efficiency, we first calculate the frequency-domain response CFR (Channel Frequency Response) and then use the computationally efficient IFFT to obtain $CIR = IFFT(CFR) = IFFT(H(f))$. $H(f)$ represents the acoustic frequency-domain signal.

The CIRs obtained from a set of chirp signals are stacked to form a matrix, which is the echo matrix of the signal generated by a sound cycle sequence.

To this end, the system applies a PCA (Principal Component Analysis) to compress the twenty-channel matrix signals into one channel. Each of these channels is input as an observation of the PCA algorithm. We use the covariance analysis of PCA to decorrelate the data, projecting the data to the direction with the most significant variance as the main feature, to achieve the purpose of dimensionality reduction. The component with the highest eigenvalue carries the most critical information: the user's unique biometric vector embedding.

5.8. Authentication

Finally, the feature vector obtained by PCA is saved to assemble the unique feature embedding of the human ear. Since the authentication issue is typically a classification problem, SVM is the most suitable method. We do not choose a data-driven deep neural network for three reasons. First, it can ensure the efficiency and applicability of the *MetaEar*. Second, it avoids much heavy work of collecting user data. Third, users can quickly sign up without retraining the network or fine-tuning. Transfer learning and these Deep Learning algorithms are almost unattainable to apply. We input the feature vector into the SVM for training and eventually output the authentication result of 1 (legitimate user) or 0 (illegal user). MetaEar uses LibSVM [53] and sets up an RBF (Radial Basis Function) kernel with the one-classification function. The dataset is divided into two parts, the training set and the test set, which account for 75% and 25%, respectively, and then utilize 10-fold cross-validation to train the model, and finally obtain the certification accuracy of the test set.

6. Implementation Setup

6.1. Mobile APP

We developed an Android APP to send and collect FMCW acoustic signals within 18∼22 kHz, as shown in Figure 8. The Android APP could config the sample rate to 48 kHz and the highest frequency to 20 kHz. The lowest frequency is 18 kHz, the chirp period is 0.01 s, and the upper speaker is employed to send the FMCW acoustic signal. Before starting to collect data, the preparation time is 1000 ms, and the total number of samples to be collected is 1.

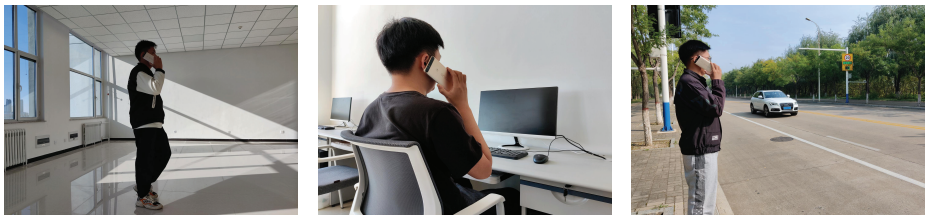
Figure 8. The Android APP and three smartphones.

As shown in Figure 8, we chose three smartphones, Huawei Nexus 6P, Samsung Galaxy S6, and OnePlus 8. They have more device diversity than headsets and can collect data at different angles, distances, and postures so that the ERTF model can be effectively verified in multiple dimensions. On the one hand, there will not be a significant deviation due to the depth or angle of the inserted earbuds. Furthermore, it also avoids the harmful effect of long-term in-ear headphones on the physiological function of the ear. In the experiment, to verify our point of view, we used a desktop computer in the background,

with AMD Ryzen 3 2200G, 16G RAM, which can ensure that our data is thoroughly and efficiently processed.

6.2. Environment

We collected FMCW sound signals with mobile phones in three scenarios, hall, laboratory, and street, as shown in Figure 9. The laboratory is relatively clean, with less noise and few people walking around. There is human noise in the hall and interference of the sound signal by moving objects. The street is the scene with the most complicated environment. Not only is there a large number of moving object interference, but also the most significant noise interference.



(**a**) The hall scenario (**b**) The laboratory scenario (**c**) The street scenario

Figure 9. Three scenarios for data collection.

7. Evaluation

7.1. Dataset

We recruited 17 people aged 22 to 25, eleven males and six females. Each person uses three mobile phones to collect data at different distances, angles, and behaviors in each scene. The collection was performed 780 times by each person, and a total of $780 \times 17 = 13260$ samples were collected. The dataset is sufficient to train the SVM for classification. During SVM training, the data of the current legitimate user are labeled as positive, while the data of other users are labeled as negative. These negative samples, randomly selected from all labeled negative users, have the same number of positive ones. We divide the current dataset into three parts, 70% for training, 15% for validation, and 15% for testing.

7.2. Overall Accuracy

Because the system aims to authenticate users, it can be attributed to a one-classification problem with traditional OC-SVM (One-Class SVM). The overall confusion matrix is shown in Figure 10. It is worth noting that this confusion matrix is different from the traditional one. The x-axis represents different users, while the ordinate represents different individual models. Each row in the figure denotes the average authentication accuracy produced after inputting different user data into models. Therefore, for each system legitimate user, a corresponding individual model is trained to constitute a model library. It can be seen that the minimum authentication accuracy is 92.8%, the highest accuracy is 100%, and the average accuracy is 96.48%.

7.3. Impact of Environment

We collected data in three different scenarios, as shown in Figure 9: the hall, the laboratory, and the noisy street. In the environment, there are different noises and disturbances, respectively. The indoor environment is relatively less noisy, but there will be reflection interference from close-range moving objects. The outdoor environment selected the noisiest and most complex street environment, which aims to test whether the system can work typically in the most complex and noisy environment. These three environments basically cover daily production and life scenarios. As shown in Figure 11, no matter what kind of environment, our system can denoise sufficiently and perform safety authentication. In most cases, the authentication accuracy in the hall and lab is better than in the street environment. In the hall and the laboratory, people often walk, and there are human voices

and keyboard and mouse tapping noises in the laboratory. The accuracy in the hall is the highest, achieving on average 96.19%. Noisy street authentication accuracy is slightly lower, and average accuracy is also achieved at 92.02%. The loud street is not only contained by high-decibel motor vehicle noise but also the voice of people. For user7 and user9, the authentication accuracy of the street environment is also relatively good due to some uncontrollable factors in the environment noise. For example, there are few vehicles and low noise during data collection.

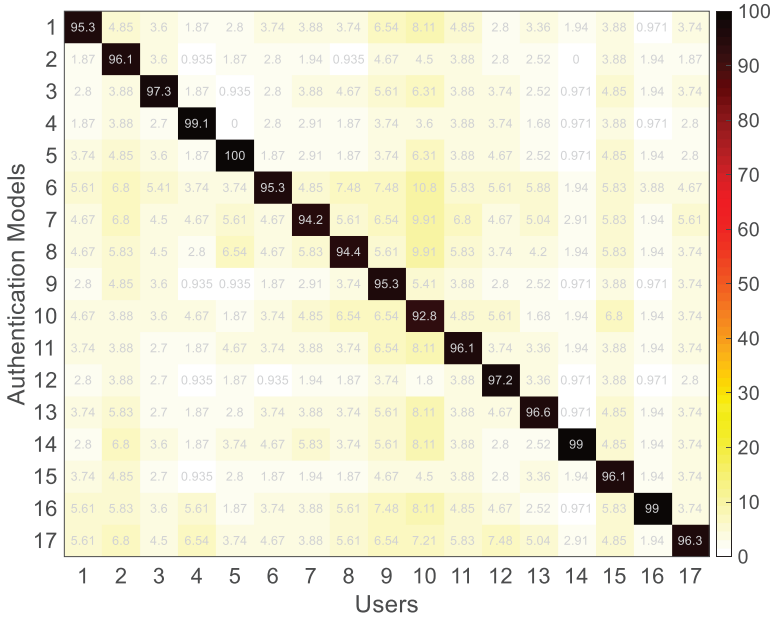


Figure 10. The fusion matrix of overall accuracy.

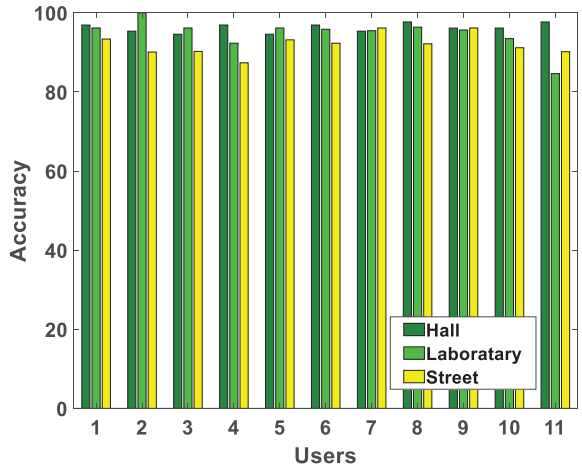
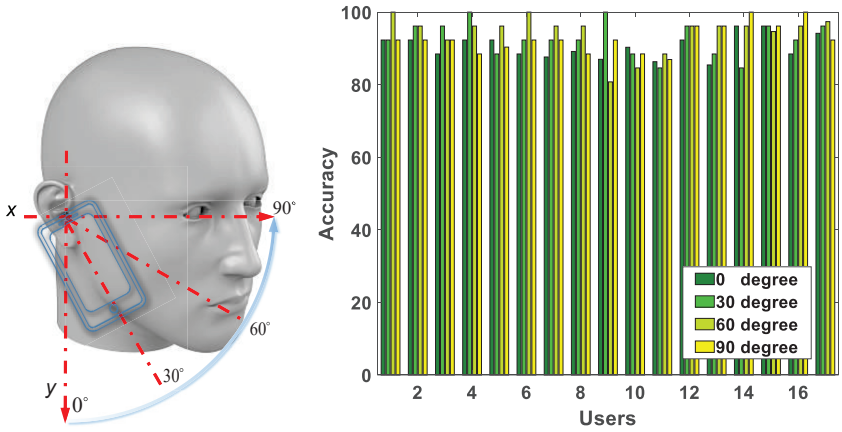


Figure 11. Accuracy at different circumstances.

7.4. Impact of Angel

Different angles of the mobile phone have different effects on the authentication accuracy. We measured different angles with a semicircular protractor in the hall and collected data when the mobile phone was at four different angles, as shown in Figure 12. The coordinate system is set to the auricle plane, the x-axis is from back to front, and the y-axis is from top to bottom. When the mobile phone and the y-axis coincide, the angle is 0 degrees. The corresponding rotation to the x-axis can reach 30 degrees, 60 degrees, and 90 degrees, four angles for experimentation in total. As can be seen from the figure, most users hold the highest authentication accuracy at 30 degrees and 60 degrees, and the average accuracy is about 93.54%. The reason is that the FMCW ultrasonic signal is most

apparent in terms of fetching the characteristic expression of the pinna and ear canal at 30 degrees and 60 degrees. In addition, the mobile phone's microphone also receives the highest signal strength, and the SNR (Signal to Noise Ratio) is optimal so that it can achieve the best authentication effect. However, some users have the highest accuracy at 90-degree angles, primarily because different users hold their mobile phones in different postures. At 90 degrees, the mobile phone is not very close to the face, so the expression of human ear biometrics is accurate.



(**a**) The angular coordinate (**b**) Accuracy at different angels

Figure 12. Impact of different angels.

7.5. Impact of Distance

We verified the authentication sensitivity to the distance between the human ear and the transceiver. The authentication accuracy was measured when the mobile phone was 1 cm, 3 cm, and 5 cm away from the human ear, as shown in Figure 13. It can be concluded from the figure that the accuracy at 1 cm is the highest, and the average accuracy is achieved at 92.65%. At a 5 cm distance, the accuracy is lower, with an average of 92.3%. This is mainly because when the device is far away from the human ear, the SNR decreases, and the characteristic signals of the ear canal and auricle feedback cannot be well received. Furthermore, hair and earrings will increase negative influence so that the accuracy will decrease. We did not test more extended distance scenarios because if the signal selective fading increases sharply with distance, the SNR decreases, and the accuracy drops severely.

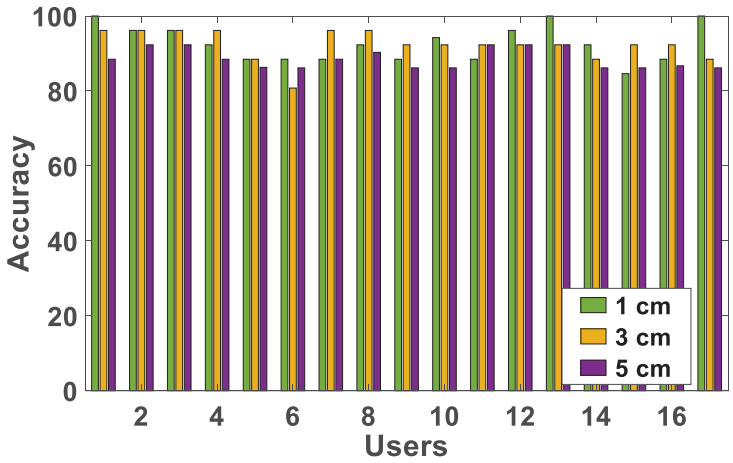


Figure 13. Accuracy at different distances.

7.6. Impact of Different Behavior

During the authentication process, users may be sitting motionless and doing some other body movements in daily life. We test the most potential activities, including static

state, shaking, and walking. as shown in Figure 14. Among them, the authentication accuracy in the static state is 96.89%. It is 96.15% in the shaking head state, and while walking, the average accuracy is 91.60%. Because the vibration of the bones and muscles of the human body will affect the feedback of the signal during the walking process, the accuracy will decline. The authentication accuracy is reduced due to the change of the mobile phone position caused by non-vigorous head shaking. In the static state, the authentication effect is the best.

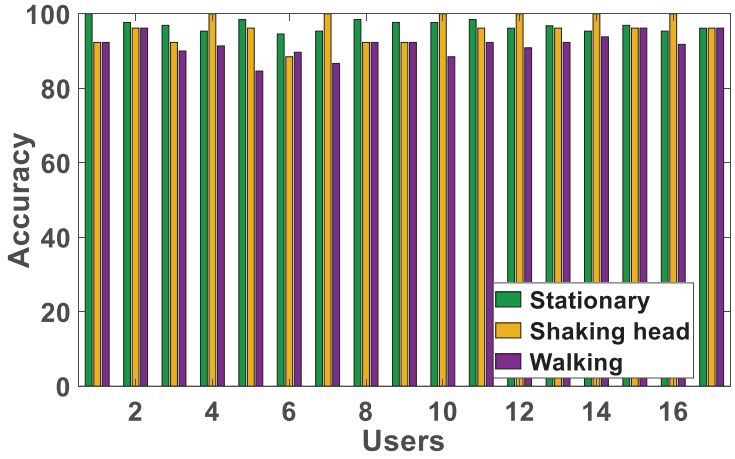


Figure 14. Accuracy of different behaviors.

7.7. Impact of Different Devices

In order to measure the impact of different hardware on the authentication, we use three brands of mobile phones for experimental verification, Oneplus 8, Samsung galaxy s6, and Huawei nexus 6p. When collecting data in the early stage, due to the difference in the microphone ADC (Analog-to-Digital Converter) circuit and nonlinear processing methods and numerical precision of different devices, the collected data will be different, so different devices will cause accuracy differences. We use AMD Ryzen PC to simulate the same technical data process. As shown in Figure 15, the average accuracy of the three devices is almost the same, showing that our system is adaptable to different hardware. Overall, the effect of Oneplus 8 is slightly better, with accuracy reaching 98.44%, principally because the smartphone was produced in 2020 with improved hardware and performance. However, the result of user10 using Oneplus 8 is low, which is mainly caused by the extensive construction noise interference in the environment when the user collects data.

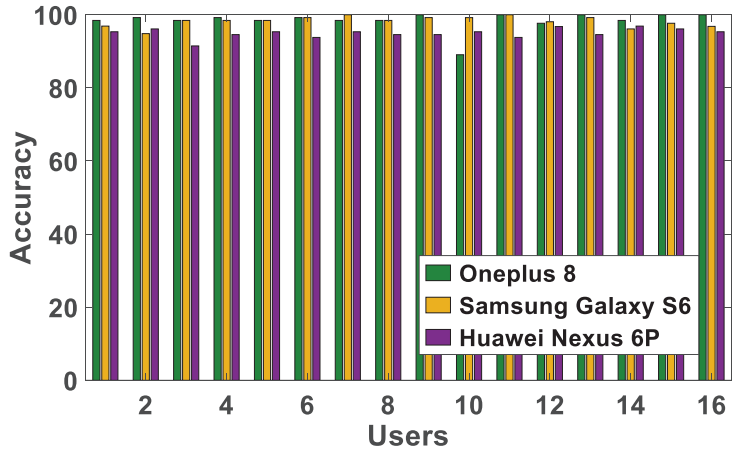


Figure 15. Authentication accuracy by different smartphones.

7.8. Impact of Data Quantity

Machine learning is a data-driven learning model. Although OC-SVM uses much less data for training than DNN, it also obtains different authentication results for different data amounts. We train the model with different amounts of data and then use the test data to test the authentication accuracy of the model, as shown in Figure 16. As the volume of data increases, the accuracy steadily increases. When using 30 data samples, the accuracy reaches 86%, and when the number of training data samples is 90, the test accuracy reaches 95.56%, which can meet the daily requirements. Ultimately, when using 590 training samples, the accuracy can reach 98.98%.

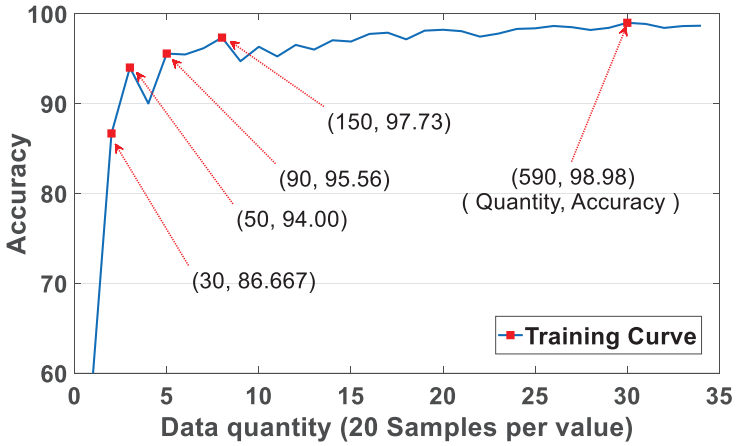


Figure 16. Training accuracy by different data quantities.

7.9. Efficiency of Attack Defense

We conducted an attack defense test on the *MetaEar*. For impersonation attacks, the adversaries utilize their biological signal to imitate the signal of the legitimate user, intending to deceive the system. We take one person as a legitimate user, and then 16 other people imitate this person's habits and behavioral characteristics to collect data. In the hall, the voice data of 17 people were collected at the same location and time. For the model trained for a legitimate user, we take the data of the remaining 16 people as the attacker, input it into the model, and the output is shown in Figure 17. It is concluded that the median FAR of other users is below 4%, except for user2, user9, user10, and user15. In particular, the FAR of user10 is 11%. The main reason for the deviation of these users is that the legitimate users did not follow the dictated actions when collecting data, which caused the distance between the device and the ear to change, and the extracted biometric characteristics of the human ear changed significantly, thereby increasing attack success probability. Nonetheless, attackers have a low probability of successfully executing impersonation attacks. Compared with face or fingerprint authentication, the registration process also requires a fixed posture and process. If our system can allow legitimate users to develop fixed habits when collecting data and adopt dictated actions, the results of resisting attacks will be better.

7.10. Analysis of Time Efficiency

We measured the time efficiency of each system module, as shown in Table 1. The table is divided into two stages, one is the registration stage, and the other is the login stage. In the registration stage, the collected data need to be trained by SVM, so the time in the table is the consumption for training 90 samples. The time statistics of each phase in the login is the time consumption of one sampling. 'pre' denotes data preprocessing, including some data access time and the time to remove the hardware delay, and its time efficiency is 0.364 s. 'Denoise' runtime is 0.414 s. The signal alignment described by 'Ali' has a running time of 0.88 s. 'SIC' is self-interference cancel, consuming 0.883 s. Furthermore, 'seg' is the chirp segment, which takes 20.056 s, and includes the total running time of segmenting samples of 90 times. This time consumption is slightly higher than the other parts. However,

it is much less than the time spent on fingerprint scanning when performing fingerprint authentication registration. So it still has an overall advantage in time efficiency. 'FE' is feature extraction, which takes 0.544 s. The SVM training phase took 1.335 s, which shows that the training efficiency of SVM is very high. Finally, the total time spent in the training stage is about 24 s, which is less than the registration time consumed by some existing fingerprint or face biometric authentication methods. When logging in, the total time consumption is 0.939 s. Moreover, in the MALAB execution environment, the complexity of all programs is O(n). After program optimization, the processing time should be able to meet the actual needs. Overall, *MetaEar* is an efficient and deployable system.

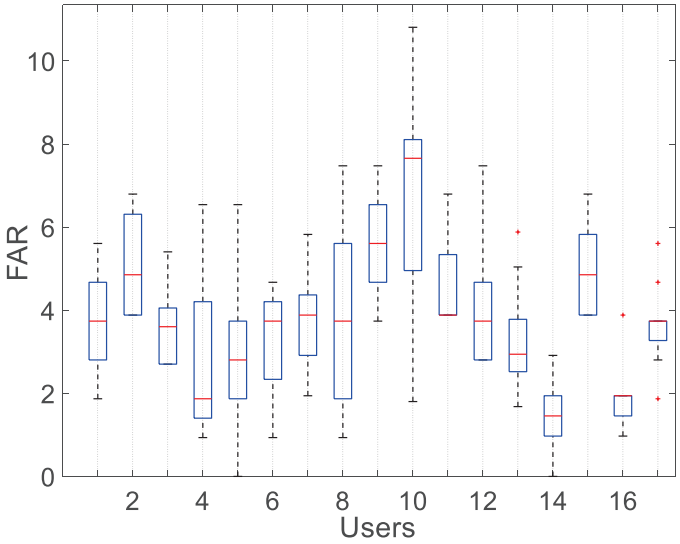


Figure 17. Resistance capability to impersonate attack (The red plus signs indicate outliers).

Table 1. Average Time Efficiency.

Phase	Pre	Denoise	Ali	SIC	Seg	FE	Train	Auth	Total
Reg(s)	0.364	0.414	0.880	0.883	20.056	0.544	1.335	None	**24.476**
Login(s)	0.015	0.018	0.008	0.009	0.273	0.022	None	0.594	**0.939**

7.11. Comparison of Different System

Furthermore, we compared the other four systems in three main aspects: device usage, biometric location, and average accuracy, as shown in Table 2. The average authentication accuracy of our system is the highest, and we use not only the intrinsic biometrics of the ear canal but also the extrinsic biometrics of the auricle, making the biometrics' uniqueness more accurate. Moreover, the smartphones we use do not need to be modified and use the machine learning algorithm SVM, which can be efficiently integrated with existing systems and deployed in practical scenarios.

Table 2. Comparison of Different Acoustic Biometric Authentication.

AcousticAuth	EarPrint [46]	VocalLock [31]	EarEcho [19]	EarDynamic [54]	MetaEar
Devices	Earphone	Smartphone	Earphone	Earphone	Smartphone
Features	Body sound	Vocal Tract	Ear Canal	Ear Canal	Ear Canal+Pinna
Accuracy	96.36%	91.1%	94.2%	93.04%	96.48%

8. Conclusions

This paper proposes *MetaEar* for modeling and authenticating human ear ERTF biometrics using FMCW ultrasonics. By sending FMCW ultrasonic waves to the ear, the dual microphones receive the feedback sound wave, extract the features through ERTF, and then feed into the SVM for one-class authentication. A large number of experiments verify that our average authentication accuracy can reach 96.48%, which can effectively strengthen biometric authentication and resist replay attacks and imitation attacks. From the overall authentication accuracy result, we do not achieve the best authentication accuracy, and the attack resistance test FAR also does not achieve the optimal level. Therefore, the next step is further improving the authentication accuracy and achieving practical deployment. First, we will use the combinatorial optimization method to perform more accurate feature extraction on ultrasound to improve the authentication accuracy; second, through multi-modal fusion biometric authentication, we will obtain a more robust performance against attacks.

Author Contributions: Conceptualization, methodology, software, validation, writing—original draft preparation, Z.C.; formal analysis, investigation, resources, funding acquisition, writing—review and editing, L.W.; data curation, visualization, B.L.; supervision, project administration, W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Science Foundation of China (NSFC) grant number 61772453, and the Natural Science Foundation of Hebei Province grant number F2022203045, F2018203444.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chen, K.; Zhang, D.; Yao, L.; Guo, B.; Yu, Z.; Liu, Y. Deep Learning for Sensor-Based Human Activity Recognition: Overview, Challenges, and Opportunities. *ACM Comput. Surv.* **2021**, *54*, 1–40. [CrossRef]
2. Choudhury, R.R. Earable Computing: A New Area to Think About. In Proceedings of the 22nd International Workshop on Mobile Computing Systems and Applications, Virtual, 24–26 February 2021; Association for Computing Machinery: New York, NY, USA, 2021; HotMobile '21, pp. 147–153.
3. Bhalla, A.; Sluganovic, I.; Krawiecka, K.; Martinovic, I. MoveAR: Continuous Biometric Authentication for Augmented Reality Headsets. In Proceedings of the 7th ACM on Cyber–Physical System Security Workshop, Hong Kong, China, 7 June 2021; pp. 41–52.
4. Lee, L.H.; Braud, T.; Zhou, P.; Wang, L.; Xu, D.; Lin, Z.; Kumar, A.; Bermejo, C.; Hui, P. All one needs to know about metaverse: A complete survey on technological singularity, virtual ecosystem, and research agenda. *arXiv* **2021**, arXiv:2110.05352.
5. *Biometric Authentication & Identification Market 2022 with Growth Opportunities, Top Countries Data, Future Trends and Business Size and Share with Revenue Forecast to 2026*; MarketWatch: New York, NY, USA, 2022.
6. Acquisti, A.; Gross, R.; Stutzman, F.D. Face recognition and privacy in the age of augmented reality. *J. Priv. Confid.* **2014**, *6*, 1. [CrossRef]
7. Raghavendra, R.; Raja, K.B.; Busch, C. Presentation attack detection for face recognition using light field camera. *IEEE Trans. Image Process.* **2015**, *24*, 1060–1075. [CrossRef] [PubMed]
8. Boulkenafet, Z.; Komulainen, J.; Li, L.; Feng, X.; Hadid, A. OULU-NPU: A mobile face presentation attack database with real-world variations. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 612–618.
9. Komkov, S.; Petiushko, A. Advhat: Real-world adversarial attack on arcface face id system. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 819–826.
10. ISO—ISO/IEC 30107-1:2016, I.I. Information Technology—Biometric Presentation Attack Detection—Part 1: Framework. Available online: https://www.iso.org/standard/53227.html (accessed on 26 September 2022).
11. Zhang, L.; Tan, S.; Yang, J.; Chen, Y. Voicelive: A phoneme localization based liveness detection for voice authentication on smartphones. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, 24–28 October 2016; pp. 1080–1091.
12. Zhang, L.; Tan, S.; Wang, Z.; Ren, Y.; Wang, Z.; Yang, J. Viblive: A continuous liveness detection for secure voice user interface in iot environment. In Proceedings of the Annual Computer Security Applications Conference, Austin, TX, USA, 7–11 December 2020; pp. 884–896.

13. Tuyls, P.T.; Verbitskiy, E.; Ignatenko, T.; Schobben, D.; Akkermans, T.H. Privacy-protected biometric templates: Acoustic ear identification. In *Biometric Technology for Human Identification*; International Society for Optics and Photonics: Bellingham, DC, USA, 2004; Volume 5404, pp. 176–182.
14. Xu, X.; Gao, H.; Yu, J.; Chen, Y.; Zhu, Y.; Xue, G.; Li, M. ER: Early recognition of inattentive driving leveraging audio devices on smartphones. In Proceedings of the IEEE INFOCOM 2017-IEEE Conference on Computer Communications, Atlanta, GA, USA, 1–4 May 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–9.
15. Xu, X.; Yu, J.; Chen, Y.; Zhu, Y.; Kong, L.; Li, M. Breathlistener: Fine-grained breathing monitoring in driving environments utilizing acoustic signals. In Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services, Seoul, Korea, 17–21 June 2019; pp. 54–66.
16. Yang, Z.; Choudhury, R.R. Personalizing Head Related Transfer Functions for Earables. In Proceedings of the 2021 ACM SIGCOMM 2021 Conference, Virtual, 23–27 August 2021; Association for Computing Machinery: New York, NY, USA, 2021; SIGCOMM '21, pp. 137–150.
17. Prakash, J.; Yang, Z.; Wei, Y.L.; Hassanieh, H.; Choudhury, R.R. EarSense: Earphones as a Teeth Activity Sensor. In Proceedings of the 26th Annual International Conference on Mobile Computing and Networking, London, UK, 21–25 September 2020; Association for Computing Machinery: New York, NY, USA, 2020.
18. Yang, Z.; Wei, Y.L.; Shen, S.; Choudhury, R.R. Ear-AR: Indoor Acoustic Augmented Reality on Earphones. In Proceedings of the 26th Annual International Conference on Mobile Computing and Networking, London, UK, 21–25 September 2020; Association for Computing Machinery: New York, NY, USA, 2020.
19. Gao, Y.; Wang, W.; Phoha, V.V.; Sun, W.; Jin, Z. EarEcho: Using ear canal echo for wearable authentication. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2019**, *3*, 1–24. [CrossRef]
20. Wang, Z.; Ren, Y.; Chen, Y.; Yang, J. Earable Authentication via Acoustic Toothprint. In Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual, 15–19 November 2021; Association for Computing Machinery: New York, NY, USA, 2021; CCS '21, pp. 2390–2392.
21. Wang, Z.; Tan, S.; Zhang, L.; Ren, Y.; Wang, Z.; Yang, J. An Ear Canal Deformation Based Continuous User Authentication Using Earables. In Proceedings of the 27th Annual International Conference on Mobile Computing and Networking, New Orleans, LA, USA, 25–29 October 2021; Association for Computing Machinery: New York, NY, USA, 2021; MobiCom '21, pp. 819–821.
22. Cai, C.; Pu, H.; Ye, L.; Jiang, H.; Luo, J. Active Acoustic Sensing for Hearing Temperature under Acoustic Interference. In *IEEE Transactions on Mobile Computing*; IEEE: Piscataway, NJ, USA, 2021; p. 1.
23. Yun, S.; Chen, Y.C.; Zheng, H.; Qiu, L.; Mao, W. Strata: Fine-grained acoustic-based device-free tracking. In Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services, New York, NY, USA, 19–23 June 2017; pp. 15–28.
24. Wang, W.; Liu, A.X.; Sun, K. Device-free gesture tracking using acoustic signals. In Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking, New York, NY, USA, 3–7 October 2016; pp. 82–94.
25. Cai, C.; Pu, H.; Wang, P.; Chen, Z.; Luo, J. We Hear Your PACE: Passive Acoustic Localization of Multiple Walking Persons. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2021**, *5*, 1–24. [CrossRef]
26. Ling, K.; Dai, H.; Liu, Y.; Liu, A.X.; Wang, W.; Gu, Q. Ultragesture: Fine-grained gesture sensing and recognition. *IEEE Trans. Mob. Comput.* **2020**, *21*, 2620–2636. [CrossRef]
27. Ruan, W.; Sheng, Q.Z.; Yang, L.; Gu, T.; Xu, P.; Shangguan, L. AudioGest: Enabling fine-grained hand gesture detection by decoding echo signal. In Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Heidelberg, Germany, 12–16 September 2016; pp. 474–485.
28. Zheng, T.; Chao, C.; Chen, Z.; Luo, J. Sound of Motion: Real-time Wrist Tracking with A Smart Watch-Phone Pair. In Proceedings of the IEEE INFOCOM 2022-IEEE Conference on Computer Communications, London, UK, 2–5 May 2022.
29. Feng, H.; Fawaz, K.; Shin, K.G. Continuous Authentication for Voice Assistants. In Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking, Snowbird, UT, USA, 16–20 October 2017; Association for Computing Machinery: New York, NY, USA, 2017; MobiCom '17, pp. 343–355.
30. Xie, Y.; Li, F.; Wu, Y.; Chen, H.; Zhao, Z.; Wang, Y. TeethPass: Dental Occlusion-based User Authentication via In-ear Acoustic Sensing. In Proceedings of the IEEE INFOCOM 2022-IEEE Conference on Computer Communications, London, UK, 2–5 May 2022.
31. Lu, L.; Yu, J.; Chen, Y.; Wang, Y. Vocallock: Sensing vocal tract for passphrase-independent user authentication leveraging acoustic signals on smartphones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2020**, *4*, 1–24. [CrossRef]
32. Lu, L.; Yu, J.; Chen, Y.; Liu, H.; Zhu, Y.; Kong, L.; Li, M. Lip Reading-Based User Authentication Through Acoustic Sensing on Smartphones. *IEEE/ACM Trans. Netw.* **2019**, *27*, 447–460. [CrossRef]
33. Zhou, M.; Wang, Q.; Li, Q.; Jiang, P.; Yang, J.; Shen, C.; Wang, C.; Ding, S. Securing Face Liveness Detection Using Unforgeable Lip Motion Patterns. *arXiv* **2021**, arXiv:2106.08013.
34. Kong, H.; Lu, L.; Yu, J.; Chen, Y.; Xu, X.; Tang, F.; Chen, Y.C. MultiAuth: Enable Multi-User Authentication with Single Commodity WiFi Device. In Proceedings of the Twenty-Second International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing, Shanghai, China, 26–29 July 2021; Association for Computing Machinery: New York, NY, USA, 2021; MobiHoc '21, pp. 31–40.
35. Shi, C.; Liu, J.; Liu, H.; Chen, Y. WiFi-Enabled User Authentication through Deep Learning in Daily Activities. *ACM Trans. Internet Things* **2021**, *2*, 1–25. [CrossRef]

36. Zhao, T.; Wang, Y.; Liu, J.; Chen, Y. Your Heart Won't Lie: PPG-Based Continuous Authentication on Wrist-Worn Wearable Devices. In Proceedings of the 24th Annual International Conference on Mobile Computing and Networking, New Delhi, India, 29 October–2 November 2018; Association for Computing Machinery: New York, NY, USA, 2018; MobiCom '18, pp. 783–785.
37. Qian, K.; Wu, C.; Xiao, F.; Zheng, Y.; Zhang, Y.; Yang, Z.; Liu, Y. Acousticcardiogram: Monitoring heartbeats using acoustic signals on smart devices. In Proceedings of the IEEE INFOCOM 2018-IEEE Conference on Computer Communications, Honolulu, HI, USA, 16–19 April 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1574–1582.
38. Lin, F.; Song, C.; Zhuang, Y.; Xu, W.; Li, C.; Ren, K. Cardiac Scan: A Non-Contact and Continuous Heart-Based User Authentication System. In Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking, Snowbird, UT, USA, 16–20 October 2017; Association for Computing Machinery: New York, NY, USA, 2017; MobiCom '17, pp. 315–328.
39. Wang, T.; Zhang, D.; Zheng, Y.; Gu, T.; Zhou, X.; Dorizzi, B. C-FMCW Based Contactless Respiration Detection Using Acoustic Signal. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2018**, *1*, 1–20. [CrossRef]
40. Chen, Y.; Xue, M.; Zhang, J.; Guan, Q.; Wang, Z.; Zhang, Q.; Wang, W. ChestLive: Fortifying Voice-based Authentication with Chest Motion Biometric on Smart Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2021**, *5*, 1–25. [CrossRef]
41. Eberz, S.; Rasmussen, K.B.; Lenders, V.; Martinovic, I. Evaluating Behavioral Biometrics for Continuous Authentication: Challenges and Metrics. In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, Abu Dhabi, UAE, 2–6 April 2017; Association for Computing Machinery: New York, NY, USA, 2017; ASIA CCS '17, pp. 386–399.
42. Zhang, Y.; Hu, W.; Xu, W.; Chou, C.T.; Hu, J. Continuous Authentication Using Eye Movement Response of Implicit Visual Stimuli. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2018**, *1*, 1–22. [CrossRef]
43. Liu, J.; Li, D.; Wang, L.; Xiong, J. BlinkListener: "Listen" to Your Eye Blink Using Your Smartphone. *Proc. ACM Interactive Mob. Wearable Ubiquitous Technol.* **2021**, *5*, 1–27. [CrossRef]
44. Kinnunen, T.; Sahidullah, M.; Delgado, H.; Todisco, M.; Evans, N.; Yamagishi, J.; Lee, K.A. The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection. In Proceedings of the 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, 20–24 August 2017; pp. 2–6.
45. Huang, W.; Tang, W.; Jiang, H.; Luo, J.; Zhang, Y. Stop Deceiving! An Effective Defense Scheme Against Voice Impersonation Attacks on Smart Devices. *IEEE Internet Things J.* **2022**, *9*, 5304–5314. [CrossRef]
46. Gao, Y.; Jin, Y.; Chauhan, J.; Choi, S.; Li, J.; Jin, Z. Voice In Ear: Spoofing-Resistant and Passphrase-Independent Body Sound Authentication. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2021**, *5*, 1–25. [CrossRef]
47. Kates, J.M. A computer simulation of hearing aid response and the effects of ear canal size. *J. Acoust. Soc. Am.* **1988**, *83*, 1952–1963. [CrossRef] [PubMed]
48. Rappaport, T.S. *Wireless Communications: Principles and Practice*; Prentice Hall PTR: Hoboken, NJ, USA, 1996; Volume 2.
49. Adib, F.; Katabi, D. See through walls with WiFi! In Proceedings of the ACM SIGCOMM 2013 Conference on SIGCOMM, Hong Kong, China, 12–16 August 2013; pp. 75–86.
50. Sun, K.; Zhao, T.; Wang, W.; Xie, L. Vskin: Sensing touch gestures on surfaces of mobile devices using acoustic signals. In Proceedings of the 24th Annual International Conference on Mobile Computing and Networking, New Delhi, India, 29 October–2 November 2018; pp. 591–605.
51. Zhou, Z.; Diao, W.; Liu, X.; Zhang, K. Acoustic fingerprinting revisited: Generate stable device id stealthily with inaudible sound. In Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, Scottsdale, AZ, USA, 3–7 November 2014; pp. 429–440.
52. Han, D.; Chen, Y.; Li, T.; Zhang, R.; Zhang, Y.; Hedgpeth, T. Proximity-proof: Secure and usable mobile two-factor authentication. In Proceedings of the 24th Annual International Conference on Mobile Computing and Networking, New Delhi, India, 29 October–2 November 2018; pp. 401–415.
53. Chang, C.C.; Lin, C.J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol. (TIST)* **2011**, *2*, 1–27. [CrossRef]
54. Wang, Z.; Tan, S.; Zhang, L.; Ren, Y.; Wang, Z.; Yang, J. EarDynamic: An Ear Canal Deformation Based Continuous User Authentication Using In-Ear Wearables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2021**, *5*, 1–27. [CrossRef]