2025 IEEE 45th International Conference on Distributed Computing Systems (ICDCS) | 979-8-3315-1723-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICDCS63083.2025.00020

BaroAuth: Harnessing Ear Canal Deformation for Speaking User Authentication on Earbuds

Luo Zhou*, Shan Chang*✉, Jiusong Luo*, Huixiang Wen†, Hongzi Zhu‡, and Li Lu§
*Donghua University †Bengbu Medical University
‡Shanghai Jiao Tong University §University of Electronic Science and Technology of China
zhouluo@mail.dhu.edu.cn, changshan@dhu.edu.cn, jiusongluo@mail.dhu.edu.cn
huixiangwen@bbmu.edu.cn, hongzi@cs.sjtu.edu.cn, luli2009@uestc.edu.cn

***Abstract*—The growing adoption of smart wearable devices (e.g., earbuds and smart watches) poses new challenges for secure and seamless user authentication due to their limited interaction interfaces. Conventional biometric methods, including fingerprints, voice, and facial recognition, often suffer from usability constraints, noise interference, or susceptibility to spoofing attacks. In this paper, we propose BaroAuth, a novel authentication system that utilizes the stable and distinctive patterns of Speech-aware Pressure Sequences (SPSs) captured by miniature MEMS barometers embedded in earbuds. The design of BaroAuth is based on two key observations. First, speech production involves coordinated movements of articulatory organs, such as the jaw, tongue, and soft palate, which reshape the ear canal geometry via the temporomandibular joint (TMJ), generating subtle pressure variations that encode the speaker's unique articulatory dynamics and physiological traits. Second, SPSs demonstrate strong intra-individual consistency and notable inter-individual variability, which can be effectively captured by barometers. We implement the prototype of BaroAuth and conduct comprehensive experiments on it. Experimental results demonstrate that BaroAuth achieves a mean false-acceptance rate (FAR) and false-rejection rate (FRR) of 1.62% and 1.74%, respectively, even under sophisticated attack scenarios.**

***Index Terms*—speaking user authentication, earbuds, ear canal deformation, barometer**

I. Introduction

In recent years, smart wearable devices have gained widespread adoption, with earbuds experiencing particularly rapid growth. The global earbuds is expected to reach $145 billion by 2031, with a compound annual growth rate of 17.0% [1]. Beyond providing convenient communication and entertainment, earbuds are becoming a key interaction gateway in IoT ecosystem. They are playing a more significant role in identity authentication for applications such as smart home control, password-free payments, and intelligent voice assistants.

User authentication methods on earbuds can be broadly categorized into three types. First, activity-based methods extract dynamic features generated during everyday human activities, with one prominent approach being voiceprint-based authentication. It captures unique vocal characteristics such as pitch, tone, and rhythm during speech. While widely used in earbuds due to easy integration with microphones, its reliability is often undermined by environmental noise, mimicry, and adversarial attacks [2]–[4]. Other activity-based methods leverage facial muscle movements [5] and acoustic resonance [6]–[9] to enable user authentication.

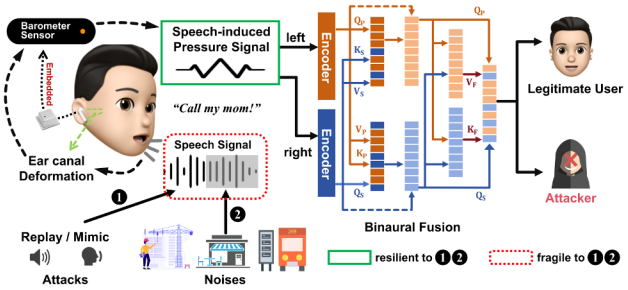


Fig. 1: BaroAuth utilizes earbuds to capture speech-induced pressure variations in ear canals and then transmits the pressure signals to paired device for speaker authentication.

These methods also maintain the usability by eliminating the need for user input, while they struggle with significant body motion artifacts due to the Inertial Measurement Unit (IMU) sensor characteristic. Second, interaction-based methods rely on predefined user actions, such as dental occlusion [10] or sliding fingers on the face with specific gestures [11] [12]. These methods require active user participation, reducing practicality in daily use. Additionally, repetitive actions tend to be less reliable, prone to operational errors, and more vulnerable to spoofing attacks. Third, biometric-based methods utilize unique physiological signals like breathing [13], heartbeat [14], [15], or ear canal geometry [16] [17]. These methods offer high uniqueness and privacy, but they typically rely on ultrasonic modulation, which increases energy consumption and may interfere with the normal audio output function of the earbuds. Moreover, concerns regarding long-term exposure to ultrasonic signals limit their scalability.

In this paper, we introduce BaroAuth, a novel user authentication system that harnesses speech-induced pressure variations within the ear canals as a robust dynamic biometric. The intuition behind BaroAuth is that *Ear Canal Dynamic Motion (ECDM) [18] is related to the complex interplay of articulatory organs, including the lower jaw, tongue, and oral cavity. This results in strong intra-individual consistency and significant inter-individual variability, making ECDM a promising dynamic biometric [19].* BaroAuth integrates a miniature MEMS pressure sensor into the earbud housing

✉ corresponding author.

2575-8411/25/$31.00 ©2025 IEEE
DOI 10.1109/ICDCS63083.2025.00020

to continuously and passively sense speech-induced pressure fluctuations. As illustrated in Figure 1, the Speech-aware Pressure Sequences (SPSs) captured by the sensors are then used by BaroAuth for authentication. SPSs offer a unique biometric solution that balances strong privacy, convenience, and reliability. Their inherent characteristics ensure security, while the use of natural speech enables seamless and user-friendly authentication.

BaroAuth operates in two phases. During offline enrollment, the user repeatedly utters a predefined passphrase while BaroAuth records the corresponding SPSs. These samples are used to create a training dataset for an end-to-end deep learning model that learns user-specific patterns to enable user authentication, eliminating the need for manual feature engineering. During online authentication, the user speaks the passphrase once, and BaroAuth determines whether the user is legitimate or not.

Challenges and Contributions. The design of BaroAuth faces three challenges, as follows:

1) Mitigating Baseline Drift: Environmental variations, such as changes in altitude and temperature, can cause baseline drift in pressure signals, which negatively impacts authentication accuracy. To address this, we apply Empirical Mode Decomposition (EMD) to adaptively isolate and remove low-frequency components, resulting in a stable and drift-free pressure signal.

2) Capturing Time and Frequency Domain Features Complementarily: SPSs reflect how dynamic speech behavior affects the static physiological structure. Besides mouth shape changes, vocal cord vibration and oral cavity resonance also induce subtle pressure fluctuations in the ear canal. Both factors impact authentication accuracy, but their complementary features are difficult to capture. we design an end-to-end deep learning model that adaptively integrates time-domain features (reflecting mouth shape changes over time) and frequency-domain features (capturing pressure fluctuations caused by vocal cord vibrations and oral cavity resonance), allowing the model to capture both global trends and local characteristics for improved authentication accuracy.

3) Addressing Variations in Speaking Rhythm and Limited Training Data: Varying rhythms of speaking can reduce the model distinguishably. To tackle this, we augment training dataset by applying time-stretching and compression operations to the recorded audio signals, simulating diverse speaking rhythms and enriching the training dataset. By producing multiple rhythm-transformed samples from limited original data, we significantly enhance the model's robustness and adaptability to real-world variations.

Our contributions are threefold:

- We propose BaroAuth, a novel user authentication method based on dynamic biometric SPSs. BaroAuth achieves high accuracy and demonstrates robustness in challenging mobile scenarios (e.g., walking, climbing stairs, and taking an elevator) and offers strong resistance against speech impersonation attacks.
- We embed miniature MEMS pressure sensors (Bosch BMP390) into a pair of earbuds, creating a prototype of BaroAuth, that ensures high hardware compatibility and low power consumption, while maintaining the normal audio output of the earphones.
- We conduct extensive real-world experiments with 24 volunteers on BaroAuth, evaluating impact factors such as speaking rhythm, command length, and motion scenarios, which demonstrate its efficacy and robustness. Furthermore, the experimental results show that, even under speech impersonation attacks, BaroAuth achieves a low false acceptance rate (FAR) and a false rejection rate (FRR) of 1.62% and 1.74%, respectively.

II. Related Work

A. Activity-based Methods

Activity-based methods capture dynamic features from daily human activities. For instance, EarDynamic [17] combines ear canal geometry with speech motion, while MandiPass [5] analyzes jaw structure via speech-induced vibrations. EarPrint [6] examines sound conduction patterns from the throat to the ear canal, and EarGate [9] captures gait characteristics from foot impacts. While effective, these methods are sensitive to ambient noise and body motion artifacts.

B. Interaction-based Methods

Interaction-based methods require predefined actions for authentication. TeethPass [10] and ToothSonic [20] use tooth movements, with the former relying on bone-conducted sounds from teeth occlusion and the latter on acoustic signals from tooth gestures. Similarly, EarSlide [11] and BudsAuth [12] capture tissue features during finger sliding across the face. While effective, these methods often suffer from poor user experience due to complex actions and low efficiency.

C. Biometrics-based Methods

Biometric methods utilize inherent human biosignals, such as breathing, heartbeat, or ear canal acoustic properties. BreathSign [13] captures in-ear sounds from breathing, HeartPrint [14] records heartbeat sounds, and EarPass [15] extracts heartbeat patterns from blood pulse signals. LR-Auth [16] analyzes ear canal modulation in response to sound. While these methods demonstrate high accuracy, they generally face challenges in reliably extracting the weak physiological signals, which are highly susceptible to noise and motion interference. This makes them less suitable for real-time authentication.

III. Preliminaries

A. System and Threat Model

BaroAuth leverages barometers embedded in earbuds to capture SPSs, providing a novel and secure approach to user authentication. Since the barometers passively and consistently detect pressure variances in the ear canals, authentication can be performed continuously or triggered on-demand based on mobile application requirements. During registration, speakers are required to utter a predefined passphrase, which will also be used for future authentication. Users are not required to perform special actions or undergo additional learning during either the registration or authentication process. During speaking, users can move freely, such as walking, climbing stairs, or taking an elevator.

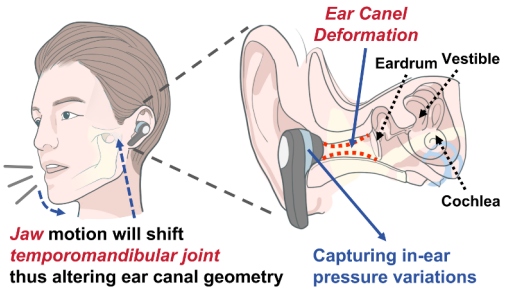


Fig. 2: Speech-induced jaw motion shifts the TMJ, causing ear canal deformation and in-ear pressure variances, captured for user authentication.

In this work, we assume a threat model where the attacker fully understands how BaroAuth operates and has gained relay access to the victim's earbuds. We consider two types of attack scenarios:

Scenario-A: Zero-knowledge Attack. The attacker knows the victim's voice command used for authentication and tries to mimic it by saying the same command without the victim's awareness. However, the attacker cannot replicate the full range of the victim's articulatory motions involved in producing the command.

Scenario-B: Knowledge-based Attack. This attack scenario follows the same process as the zero-knowledge attack but assumes the attacker secretly observes or records how the victim speaks the target command. Through extensive observation and study, the attacker becomes proficient in mimicking the victim's articulatory motions.

B. Speech-aware Pressure Sequence

Speech is produced through the coordinated movements of articulatory organs, such as the jawbone, tongue, and lips, supported by the temporomandibular joint (TMJ), which connects the mandible to the temporal bone. These movements dynamically shape the vocal tract and significantly influence the geometry of the ear canals, a phenomenon known as ear canal dynamic deformation. Driven primarily by jaw and tongue activity, these deformations alter the ear canal's geometric properties, including volume and diameter. For instance, jaw movements can cause ear canal volume changes ranging from 10 mm^3 to 25 mm^3 [19], or diameter variations of up to 2.5 mm [21], depending on movement intensity and direction. Such deformations exhibit high inter-individual variability but strong intra-individual consistency, reflecting unique anatomical and behavioral traits. These characteristics establish ear canal dynamic deformation as a promising biometric feature for user authentication.

Definition 1. A **Speech-aware Pressure Sequence** refers to the dynamic ear canal deformation caused by speech production. It consists of the following two components: 1) Immediate movements of articulatory organs. These include activities of the jaw and tongue, which drive TMJ movements, directly causing geometric changes in the ear canal, such as expansion or compression; 2) Secondary ear canal geometry deformation induced by traction. These deformations result from the coordinated activities of related muscles, such as the masseter, zygomaticus, and sternocleidomastoid, which reshape the ear canal geometry through traction effects.

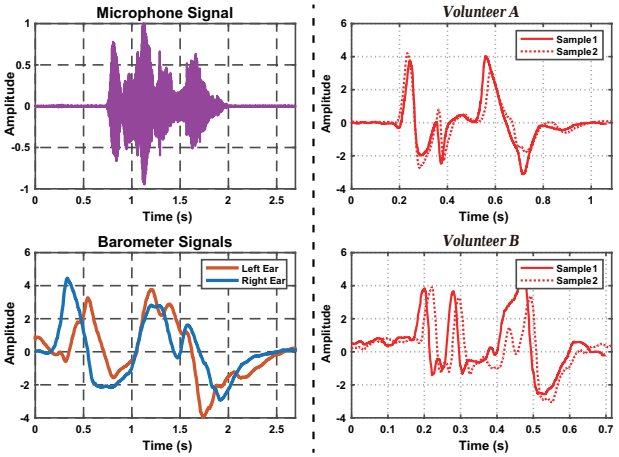


Fig. 3: (a) When a speaker utters "Navigate to my office", the microphone picks up the raw audio data; meanwhile, the barometer records the corresponding SPS. (b) Pressure signals from left ear sampled during volunteer A (above) and B (below) uttering command "Video my mom" twice.

C. Distinguishability of SPSs

Unlike active acoustics-based ear canal deformation detection methods, we develop an in-ear earphone with pressure sensors named Barobuds (see Figure 10a). Barobuds passively captures real-time dynamic pressure variances in the ear canals for robust speaker authentication. As shown in Figure 3a, when a participant utters "Navigate to my office", pressure signals from the left and right ear canals precede audio signals recorded by a microphone, which occurs due to preparatory actions like jaw adjustment and mouth opening before vocalization, and post-speech movements such as jaw retraction and mouth closure, which cause a lag in pressure signals. These pressure changes reflect articulatory patterns, revealing individual speaking habits and physiological traits. Even if an attacker mimics the victim's articulatory motion, structural differences in the ear canal produce distinct pressure patterns, ensuring reliable authentication.

Additionally, analysis of data from multiple participants reveals consistent differences in dynamic pressure between the left and right ear canals of the same individual (after excluding device performance variations). This likely stems from asymmetric facial muscle exertion during speech or anatomical differences in the ear canals. Studies [22] [23] confirm that ear canals exhibit slight but significant geometric asymmetries, such as variations in length, curvature, and cross-sectional area. These asymmetries, combined with individual speaking patterns, contribute to the observed pressure differences. Based on these findings, we hypothesize that the SPS for specific speech content reflects unique and repeatable spatiotemporal features for each speaker. As shown in Figure 3b, SPSs exhibit strong intra-speaker consistency and inter-speaker variability. For example, when two volunteers (A and B) each pronounce

"Video my mom" twice, samples from the same speaker show high similarity, while samples from different speakers vary significantly. These results highlight the potential of BaroAuth to effectively leverage SPSs for speaker authentication.

IV. Data Collection

We develop a Python application running on a PC to collect microphone and barometer signals simultaneously via serial communication with an Arduino sensing board (sampling rates of 44,100 Hz and 108 Hz, respectively). We recruit 12 volunteers ($four$ women and $eight$ men) aged 21 to 50. As shown in Figure 10b, each volunteer wears our Barobuds and either stands or sits while uttering voice commands, ensuring minimal head movement. Additionally, volunteers are free to rest or exit at any time.

Each volunteer preselects ten commands from Table I, which lists 30 commands divided into $three$ groups by length, ensuring an even distribution across groups. Volunteers coordinate to ensure each command is selected $four$ to six times. We then collect $three$ types of data:

Trace A: Collected over 30 days. Each volunteer speaks each command ten times daily with different speeds and vocal amplitude. Specifically, each command is spoken six times with varying speeds (two times at normal, slow, and fast speeds) and twice at normal speed with small and large vocal amplitude. To study the robustness of BaroAuth in motion scenarios, we conduct data collection during the last 15 days in $three$ different mobile scenarios: walking, climbing stairs, and riding an elevator, each for $five$ days. This results in 300 microphone signals and corresponding pressure sensor signals per command per volunteer. We also record the entire process with a video camera for future mimic attacks.

Trace B: For mimic attacks, we select $five$ volunteers as victims, with others as attackers. Each attacker performs ten zero-knowledge attacks (see Scenario A) for each victim's command. Then, after observing the victims' mouth movements while speaking the target command in videos, the attackers attempt to mimic the commands ten times (see Scenario B). Each command undergoes 100 attacks for both scenarios.

Trace C: Collected one week (six months after Trace A). Volunteers speak each command ten times daily at normal speed and articulatory motion amplitude.

TABLE I: Command List

Command		
2-3 words (*nine* commands)		
1. OK Google.	2. Hey Siri.	3. Take photo.
4. Call Nancy.	5. Turn off Wi-Fi.	6. Video my mom.
7. Check e-mail.	8. Lock screen.	9. Disable all alarms.
10. Open google maps.		
4-5 words (*ten* commands)		
11. Navigate to my office.	12. How's the weather today?	
13. Find my way home.	14. How are you today?	
15. Turn on the lights.	16. Turn off my reminders.	
17. What's on my calendar next?	18. Send a message to mom.	
19. Open up google scholar website.	20. Take me to nearby supermarket.	
6-7 words (*ten* commands)		
21. Where is the closest grocery store?	22. Can you translate this into Chinese?	
23. Can you show me nearby hotels?	24. Play my favorite song by Apple Music.	
25. Who won the final game last night?	26. I think you look really cool today.	
27. Help me find the nearest gas station.	28. What's the current temperature in Shanghai?	
29. I would like to order KFC takeout.	30. Show me around the local tourist attractions.	

V. Design of BaroAuth

A. Overview

The core concept of BaroAuth is to use in-ear wearable devices to capture the unique dynamic deformation characteristics of the ear canal during speech. When a user speaks a command, the system simultaneously records voice and pressure signals from the left and right ear canals using a microphone and the pressure sensor. It is important to note that the voice data is only used for data preprocessing and augmentation, and is not involved in the model training.

BaroAuth authentication has two phases: offline training and online authentication. During the training phase, users repeat the voice command multiple times to collect both speech and pressure data. In the authentication phase, the pressure data is fed into an end-to-end deep learning model to verify whether it originates from a registered user.

BaroAuth includes three modules: *Data Preprocessing*, *Data Augmentation*, and *Authentication Network*, detailed below:

B. Data Preprocessing

In our model, we use pressure data from both the left and right ear canals, which follow the same preprocessing pipeline. For simplicity, we treat the SPSs as a pair in the following sections. Given a pair of time-synchronized sequences from the microphone and barometer, we apply the following preprocessing steps:

1) *Voice Activity Detection.* Since we only focus on the pressure time series associated with human speech activities, we utilize WebRTC [24] to extract speech segments in the audio signal. Specifically, given an audio signal V, it can be segmented into multiple fragments based on the voice commands, and the corresponding SPS can be divided synchronously, denoted as P. This process generates several pairs of microphone and barometer data segments. Each of these fragments is represented as $S_i = (V_i, P_i), i \geqslant 1$, where V_i and P_i denote the audio sequence and the corresponding SPS, respectively.

Notably, during segmentation, S_i exhibits a certain degree of lead and lag relative to P_i (as illustrated in Figure 3(a)). Assuming $V_i = \{v_{i,t_0}, v_{i,t_1}, \ldots, v_{i,t_T}\}$, where t_0 and t_T represent the start and end time points of V_i, respectively, we extract the complete SPS as $P_i = \{p_{i,t_0-\sigma}, \ldots, p_{i,t_0}, \ldots, p_{i,t_T}, \ldots, p_{i,t_T+\sigma}\}$, where the parameter σ adjusts the time boundaries.

2) *Baseline Drift Elimination.* Due to the characteristics of the pressure sensor, sudden changes in altitude or external temperature cause baseline drift in the SPSs detected by Barobuds, resulting in varying degrees of distortion in P_i and reducing its fidelity. To ensure practicality, this study focuses on $three$ typical mobile scenarios: walking, climbing stairs, and taking an elevator. As shown in Figure 4, during

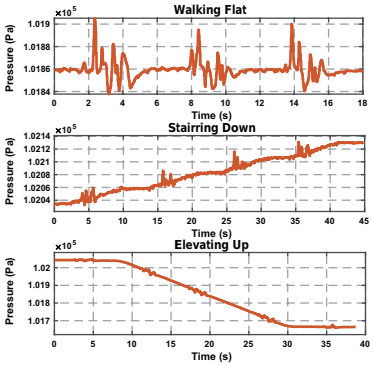


Fig. 4: SPSs under three classical modalities.

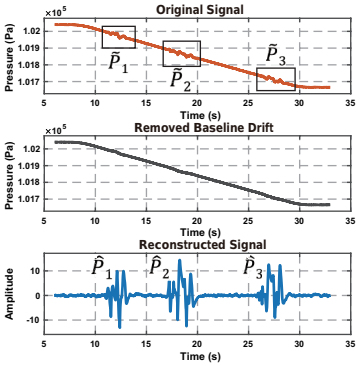


Fig. 5: An example of removed baseline drift and reconstructed SPSs

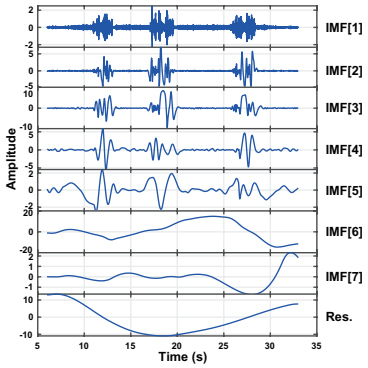


Fig. 6: An example of EMD decomposition

walking on flat ground, the pressure signals remain stable with no baseline drift. Additionally, minor head movements while speaking do not affect ear canal pressure measurements, as the Bosch BMP390 pressure sensor used in BaroAuth can detect altitude changes as small as 8 cm. However, during stair climbing and elevator rides, significant baseline drift occurs, distorting the SPSs. Notably, the ear pressure decreases with increasing altitude and increases as altitude decreases.

We use Empirical Mode Decomposition (EMD) [25] as a filter bank to remove low-frequency baseline drift signals and extract drift-free SPSs. Unlike traditional methods that rely on predefined basis functions (e.g., sine functions in Fourier transforms or mother wavelets in wavelet transforms), EMD is data-driven and adaptively decomposes the input signal into multiple frequency components based on its intrinsic characteristics. Each frequency component represents a specific oscillatory mode. Specifically, EMD assumes that any complex signal $f(x)$ is composed of x sub-signals, known as Intrinsic Mode Functions (IMFs). Each IMF is a zero-mean, narrowband oscillatory signal that reflects the oscillatory behavior of the signal in different frequency ranges.

To decompose the SPS with baseline drift $\tilde{P}_i(t)$ into a finite number of IMFs, EMD proceeds as follows:

- Identify all extrema of $\tilde{P}_i(t)$.
- Interpolate between all local minima (or maxima) to obtain the envelopes, $E_{\min}(t)$ and $E_{\max}(t)$.
- Compute the mean $R(t) = [E_{\min}(t) + E_{\max}(t)]/2$.
- Extract the detail signal $D(t) = \tilde{P}_i(t) - R(t)$.
- Iterate on the residual signal $R(t)$ until it no longer satisfies the zero-mean and narrowband oscillation conditions.

Finally, $\tilde{P}_i(t)$ is decomposed as $\tilde{P}_i(t) = \sum_{m=1}^{M} D_m(t) + R_M(t)$, where $D_m(t)$ represents the m^{th} IMF corresponding to different frequency-local oscillations, and $R_M(t)$ represents the final residual component, capturing the global low-frequency trend (e.g., baseline drift). By removing the low-frequency residual $R_M(t)$, we can extract the target signal $P_i(t)$ with no baseline drift.

Figure 5 shows the EMD decomposition results of an ear canal pressure signal captured by Barobuds while the user is speaking in an elevator. After EMD decomposition, BaroAuth obtains $seven$ IMFs and one residual. IMF corresponds to high-frequency noise from the sensor. IMF[2-5] contain the frequency components associated with the SPS. The remaining IMFs correspond to low-frequency baseline drift components. As shown in Figure 6, using IMF[2-5], we reconstruct the drift-free SPS, denoted as $\hat{P}_i(t)$. To validate the effectiveness of the EMD algorithm, we compare $\hat{P}_i(t)$ with $P_i(t)$, recorded in a stationary state under the assumption that the user speaks the same content with the same rhythm. Figure 7 presents the comparison results, showing a Pearson correlation coefficient [26] of 0.81 between the two signals, indicating a high degree of similarity between $\hat{P}_i(t)$ and $P_i(t)$. Additionally, EMD inherently removes high-frequency sensor noise during decomposition.

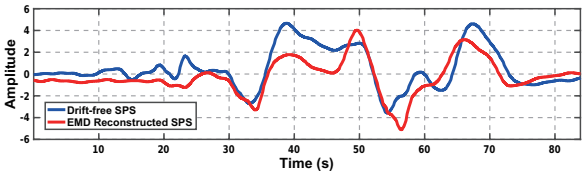


Fig. 7: An example of SPSs decomposed via EMD and from a drift-free barometer.

3) *Normalization.* Since the barometer directly reflects surrounding atmospheric pressure, the baseline pressure detected by BaroAuth varies across different sessions. Additionally, the varying amplitude of articulation motions during speech leads to different degrees of ear canal deformation, causing differences in ear canal pressure oscillation. To mitigate these effects, we apply Min-Max normalization to each pressure segment P_i, scaling the values to the range [-1, 1].

C. Data Augmentation

The SPSs are strongly affected by the user's speaking rhythm. Using raw data directly for training reduces the model's ability to adapt to inputs with varying rhythms, which weakens its generalization. To tackle this issue, we design a rhythm transformation-based data augmentation method. This method transforms each segment V_i to varying degrees, generating rhythm-diverse P_i and significantly enhancing the model's robustness to different rhythms. Specifically, the process includes the following steps:

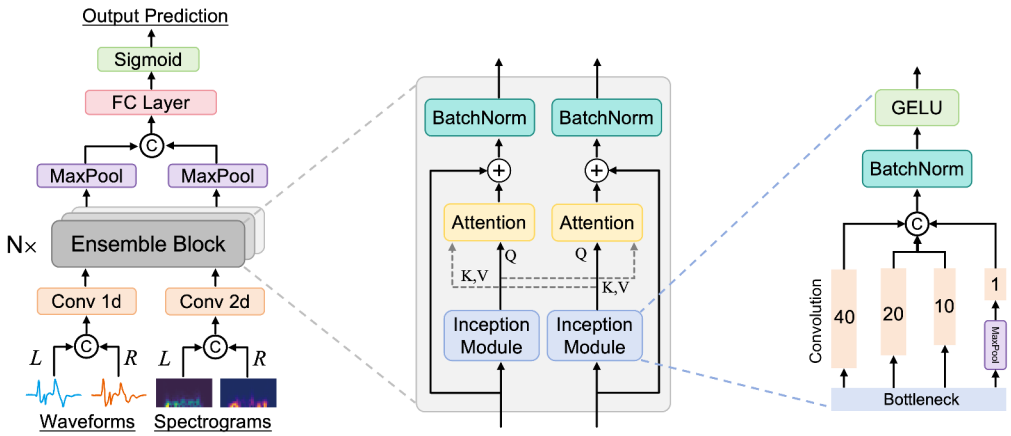


Fig. 8: The architecture of BaroIDNet.

1) *Alignment.* As mentioned in Section V-B, the SPS exhibits varying degrees of lead and lag compared to the synchronously sampled audio sequence. To align the pressure signal P_i with V_i on the time axis and facilitate subsequent rhythm transformations of the SPS, we resample P_i , adjusting its time length to match V_i . Specifically, assuming the length of the pressure signal P_i is n_P and the length of the audio signal V_i is n_V , we proceed as follows:

$$t_k^P = \frac{k}{n_P - 1}, k \in [0, n_P - 1] \tag{1}$$

$$t_j^V = \frac{j}{n_V - 1}, j \in [0, n_V - 1] \tag{2}$$

To achieve alignment, we map the time points of the audio signal t_j^V onto the time axis of the pressure signal, obtaining the corresponding positions:

$$t_j^P = \frac{j}{n_V - 1} \times (n_P - 1) \tag{3}$$

Since t_j^P may not exactly coincide with the sampling points of the pressure signal, we calculate its corresponding value through interpolation. If t_j^P lies between t_k^P and t_{k+1}^P , the corresponding pressure signal value p'_{i,t_j} is computed using linear interpolation:

$$p'_{i,t_j} = p_{i,t_k} + \frac{p_{i,t_{k+1}} - p_{i,t_k}}{t_{k+1}^P - t_k^P} \times (t_j^P - t_k^P) \tag{4}$$

We repeat the above process for all t_j^V values, ultimately obtaining the resampled pressure signal:

$$P'_i = \{p'_{i,t_0}, p'_{i,t_1}, ..., p'_{i,t_T}\} \tag{5}$$

2) *Rhythm Transformation.* Since the rhythm characteristics of audio signals in the time domain (e.g., syllables and pauses) are more distinct, we apply rhythm transformations, including time stretching and compression, to V_i, simulating speech patterns at various speaking speeds. Using a phase vocoder [27], we perform time scaling on V_i while preserving its pitch. First, we apply Short-Time Fourier Transform (STFT) to obtain the spectral representation $\mathbb{F}(V_i)$. Then, we adjust the time axis of the spectrum using a speed ratio $\beta \in [0.7, 1.3]$ with a step size of 0.1, generating a set of transformed spectra $\{\mathbb{F}(V_i'^{\beta})\}$. Finally, for each β, we reconstruct the corresponding rhythm-transformed audio signals $\{V_i'^{\beta}\}$ using the inverse STFT: $V_i'^{\beta} = \mathbb{F}^{-1}(\mathbb{F}(V_i'^{\beta}))$.

3) *SPS Synchronization.* Since V_i and P'_i are synchronized on the time axis (after *Alignment*), after applying rhythm transformations to the audio signal V_i, we obtain a set of transformed signals $\{V_i'^{\beta} | \beta \in [0.7, 1.3]\}$, each corresponding to a different speaking rhythm. For each $V_i'^{\beta}$, P'_i undergoes the same alignment process to generate $P_i'^{\beta}$. Ultimately, we form a set of synchronized pairs $S'_i = \{(V_i'^{\beta}, P_i'^{\beta}) | \beta \in [0.7, 1.3]\}$. As a result, the original S_i is augmented to produce six additional copies with different speaking rhythms (except $\beta = 1.0$), forming the augmented set S'_i.

D. BaroIDNet for Authentication

Then, S'_i are fed into **BaroIDNet** (as shown in Figure 8), which integrates time- and frequency-domain features at the initial stage, leveraging multi-scale convolution and cross-domain attention mechanisms for deep feature interaction and fusion. A residual regularization strategy enhances training stability, transforming raw signals into compact and efficient feature vectors for authentication.

1) *Parallel Representation Construction.* By analyzing raw time series data from numerous participants (after accounting for variations in earbud performance), we observe that SPSs from the binaural channel vary significantly across individuals. While some participants exhibit high similarity between their left and right ear canals, others show notable differences. This inter-ear canal variability is highly individual-specific, reflecting subtle personal characteristics and providing valuable information for identity recognition. To fully exploit this property, we concatenate the SPSs from the binaural channel in both the time and frequency domains and apply convolutional operations for preliminary feature extraction.

Let $\mathbf{x}_t^{(L)}, \mathbf{x}_t^{(R)} \in \mathbb{R}^T$ denote the time-domain pressure sequences from the left and right ear canals, each of length T. Using STFT, these signals are converted into spectrograms $\mathbf{X}_f^{(L)}, \mathbf{X}_f^{(R)} \in \mathbb{R}^{F \times T'}$, where F represents the frequency resolution and T' is the number of time steps after framing. The concatenated SPSs from the left and right ear canals are then passed through convolutional layers for feature extraction.

In the time domain, one-dimensional convolution (Conv1D) is used to extract features:

$$\mathbf{H}_t = \text{Conv1D}([\mathbf{x}_t^{(L)}, \mathbf{x}_t^{(R)}]) \in \mathbb{R}^{d \times T} \tag{6}$$

In the frequency domain, two-dimensional convolution (Conv2D) is applied to the spectrograms, yielding:

$$\mathbf{H}_f = \text{Conv2D}([\mathbf{X}_f^{(L)}, \mathbf{X}_f^{(R)}]) \in \mathbb{R}^{d \times T''} \tag{7}$$

where d is the feature dimension, and T'' represents the time steps after convolution and pooling operations.

2) *Multi-scale Feature Extractor.* To capture features at multiple temporal scales in ear pressure signals, we apply Inception modules [28] to both $\mathbf{H}_t$ and $\mathbf{H}_f$, which enables multi-scale modeling through parallel branches with different convolution kernel sizes and a pooling branch. The kernel sizes are dynamically defined as $[k, k/2, k/4]$, where k represents the initial receptive field size. Each convolution branch produces feature maps with a channel size of $d_{\text{model}}/4$, balancing model complexity and representation power. For the pooling branch, we apply 3×3 max pooling (MaxPool) followed by a 1×1 convolution to map feature dimensions and capture global context. The final module output integrates all branches, forming feature representations as:

$$\mathbf{Z} = [\mathbf{z}^k; \mathbf{z}^{k/2}; \mathbf{z}^{k/4}; \mathbf{z}^{\text{pool}}] \tag{8}$$

where $\mathbf{z}^k$ corresponds to the outputs from convolution branches with different kernel sizes, and $\mathbf{z}^{\text{pool}}$ denotes the output of the pooling branch.

3) *Feature Fusion.* SPSs exhibit complex time- and frequency-domain characteristics, including instantaneous dynamic changes and harmonic energy distributions. To integrate these complementary features, we introduce a cross-attention-based fusion mechanism that dynamically aligns and models interactions between time- and frequency-domain features. This approach ensures efficient and synergistic feature optimization.

Given time-domain features $\mathbf{Z}_t \in \mathbb{R}^{d_t \times T}$ and frequency-domain features $\mathbf{Z}_f \in \mathbb{R}^{d_f \times F}$, the mechanism uses *Query*, *Key*, and *Value* mappings to establish dependencies between the two domains. For example, when time-domain features attend to frequency-domain features, $\mathbf{Z}_t$ serves as the *Query* (Q_t), while $\mathbf{Z}_f$ acts as the *Key* (K_f) and *Value* (V_f). The attention mechanism computes similarity scores to generate weights, capturing dependencies between temporal and spectral components:

$$\text{Attention}_{t \to f} = \text{Softmax}\left(\frac{\text{Q}_t \text{K}_f^T}{\sqrt{d_k}}\right) \text{V}_f \tag{9}$$

where $\text{Q}_t = \text{W}_q \mathbf{Z}_t$, $\text{K}_f = \text{W}_k \mathbf{Z}_f$, and $\text{V}_f = \text{W}_v \mathbf{Z}_f$. $\text{W}_q \in \mathbb{R}^{d_k \times d_t}$, $\text{W}_k, \text{W}_v \in \mathbb{R}^{d_k \times d_f}$ are learnable linear projection matrices. These matrices project input features into a unified feature space, with d_k representing the dimension of the projected features.

To enhance joint representation learning, the attention output is refined with residual connections and normalization, resulting in updated time-domain features $\mathbf{Z}_t'$. Similarly, cross-attention is computed for frequency-domain features attending to time-domain features, yielding $\mathbf{Z}_f'$. This bidirectional attention mechanism explores the complementarity between the two domains. Finally, the fused representation is formed by concatenating $\mathbf{Z}_t'$ and $\mathbf{Z}_f'$.

4) *Feature Combination and Classification.* After convolution, Inception modules, and cross-attention fusion, we apply a global pooling layer to compress and integrate time- and frequency-domain features into a unified representation $\mathbf{F}$. Global pooling, such as adaptive average pooling, reduces dimensionality while preserving essential temporal and frequency information. The resulting feature vector $\mathbf{F}$ captures both fine-grained dynamic changes and global spectral structures. This feature vector is then fed into a fully connected layer, mapping it to the classification space for determining user legitimacy.

During training, we use Binary Cross-Entropy loss as the optimization objective, defined as:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log \sigma(\hat{y}_i) + (1 - y_i) \log(1 - \sigma(\hat{y}_i))] \tag{10}$$

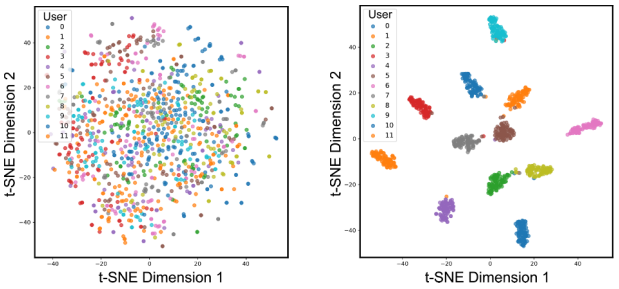
where $\hat{y}_i$ is the predicted output, $\sigma(\cdot)$ is the Sigmoid activation function, and $y_i \in \{0, 1\}$ denotes the ground-truth label.

To evaluate the ability of BaroAuth to extract user-discriminative features, we visualize both raw input data and model-extracted features using t-SNE [29]. Figure 9a shows the t-SNE visualization of raw input data, where samples from different users are intermixed, indicating a lack of discriminative properties. In contrast, after feature extraction, Figure 9b illustrates the model-extracted features, where samples from the same user form distinct, compact clusters. These results demonstrate the effectiveness of BaroAuth in transforming ambiguous raw inputs into a structured feature space with clear user-specific separability, highlighting its potential for identity verification.

VI. Implementation

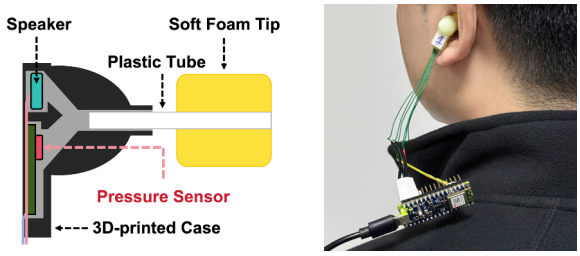
Hardware. BaroAuth consists of a custom-designed pair of earbuds, an Arduino sensing board, a Windows laptop, and a Linux server. As shown in Figure 10a, we develop a prototype of the earbuds with a 3D-printed case, adapting design parameters from OpenEarable [30]. We redesign the PCB circuit board to integrate a Bosch BMP390 pressure sensor, which providing a single-package solution (3mm x 3mm x 0.75mm). We use NANO 33 BLE Sense development board to acquire and transmit barometer signals at a sampling rate of 108 Hz. The Windows laptop, equipped with an Intel Core i7-8750H processor, 16 GB of RAM, and an NVIDIA RTX 1060 GPU, is used for real-time signal monitoring and initial preprocessing. Meanwhile, the Linux server, running Ubuntu 18.04, is equipped with dual AMD EPYC 7542 32-core processors, 64 GB of RAM, and an NVIDIA Tesla V100 GPU. It handles the execution of BaroIDNet for authentication.

Model Training and Testing. BaroIDNet is implemented in Python using the PyTorch framework, and its training and testing processes are designed to ensure a reliable evaluation



(a) Raw Data (b) Fusion Features

Fig. 9: The visualization of features extracted by (a) raw data and (b) BaroAuth.



(a) Hardware Schematic (b) Data-collection Setup

Fig. 10: (a) Hardware schematic of the Barobuds, (b) Experimental setup during data collection.

of its performance on the self-collected dataset (as described in Section IV). Typically, for all the datasets evaluated in the experiments, the split ratio of the training set and the test set is usually 8:2 with 5-fold cross-validation for reliable results. We adopt the Adam optimizer with an initial learning rate of 0.001, paired with a StepLR scheduler that reduces the learning rate by a factor of 0.5 per epoch. The training is conducted with a batch size of 64 and a maximum of 40 epochs. An early stopping strategy halts training if the validation loss shows no improvement for 10 consecutive epochs. To further enhance model stability and convergence, we apply Dropout with a rate of 0.1 and Batch Normalization.

VII. Performance Evaluation

A. Baselines and Metrics

Baselines. To thoroughly evaluate the performance of BaroIDNet, we compare it with baseline methods, including SVM (Support Vector Machine), KNN (K-Nearest Neighbors), CNN-LSTM (CNN-Long Short-Term Memory), and TCN (Temporal Convolutional Network). SVM and KNN are classical machine learning models relying on handcrafted features, while CNN-LSTM and TCN are deep learning architectures tailored for time series modeling. This comparison highlights the superiority and robustness of BaroIDNet in processing ear pressure signals. For fairness, all methods use the same dataset splits, training settings, and evaluation metrics.

Metrics. We employ the FAR and FRR as evaluation criteria. FAR is defined as the ratio between the number of falsely accepted illegitimate inputs and the number of all illegitimate testing inputs. FRR is defined as the ratio between the number of falsely rejected legitimate inputs and the number of all legitimate testing inputs. A Receiver Operating Characteristic (ROC) curve illustrates the diagnostic ability of a binary classifier as its discrimination threshold is varied. We obtain the Equal Error Rate (EER) from the ROC curve where FAR and FRR are equal.

B. Overall Performance

We first evaluate the overall performance of BaroAuth based on the traces described in Subsection IV by examining the FAR and FRR in the static scenario. During the experiment, we consider each participant as a valid user, and remaining participants as attackers. The experiment simulates the zero-knowledge attack where attackers expect their SPSs can fool the system. Overall, the average FAR and FRR over 12 participants are 1.57% and 1.31%, respectively. Figure 11 and Figure 12 show the detailed FAR and FRR results of each participant. We notice that BaroAuth achieves good performance and gives accurate classification labels based on the captured SPSs.

TABLE II: Performance comparison across different methods on monaural and binaural channel.

Methods	Monaural Channel		Binaural Channel	
	FAR	FRR	FAR	FRR
SVM	22.82%	8.05%	22.91%	8.15%
KNN	8.63%	7.53%	8.79%	7.79%
CNN-LSTM	45.1%	52.28%	45.25%	53.06%
TCN	5.13%	4.86%	5.23%	5.10%
BaroIDNet	**4.05%**	**3.25%**	**1.57%**	**1.31%**

Compared to the four baseline methods, BaroIDNet achieves lower FAR and FRR in both monaural and binaural channel, as shown in Table II. In the monaural channel, BaroIDNet achieves a FAR of 4.05% and FRR of 3.25%, while in the binaural channel, these values drop to 1.57% and 1.31%, respectively. Notably, the significant performance improvement in the binaural channel is due to BaroIDNet's ability to extract and fuse complementary features from left and right ear signals using convolution operations in both time and frequency domains. In contrast, the baselines simply concatenate SPSs from both ears, failing to leverage their collaborative features effectively.

C. Impact Factors

1) Training Data Size: We evaluate the binary classification performance of BaroAuth under different training dataset ratios. Specifically, for each predefined language commands, we trained and tested the model using 20%, 40%, 60%, 80%, and 100% in the training data. As shown in Figure 13, BaroAuth demonstrates exceptional performance across varying training data scales. Even with only 20% of the training data, the system achieves an Equal Error Rate (EER) of 6.10%, highlighting its robustness under limited data conditions. As the proportion of training data increases, the EER steadily decreases, leading to significant performance improvements. With 40% and 60% of the data, the EER drops to 3.22% and 2.32%, respectively. When the training data proportion

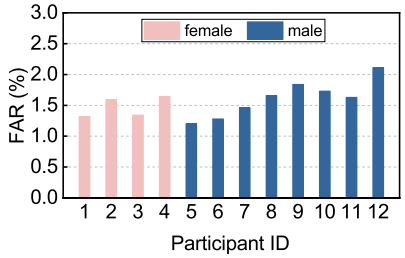


Fig. 11: FARs of participants.

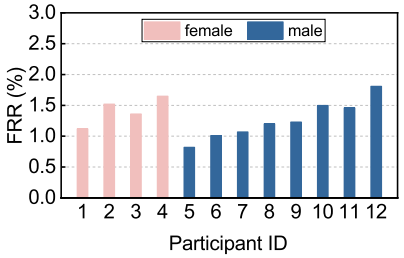


Fig. 12: FRRs of participants.

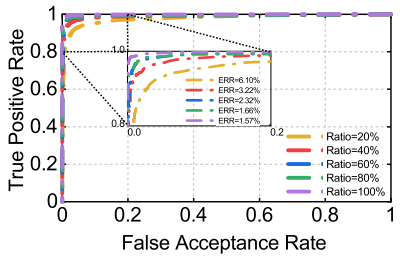


Fig. 13: The ROC curves and EERs under different training data ratios.

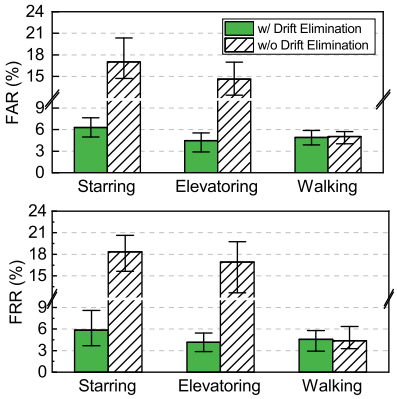


Fig. 14: FAR and FRR under different mobile scenarios

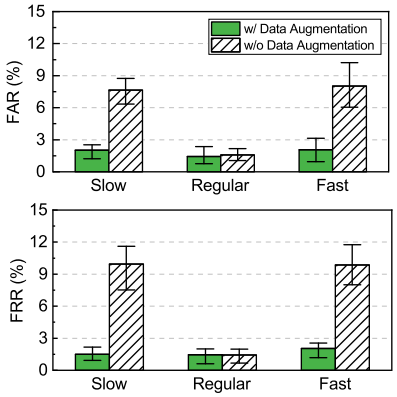


Fig. 15: FAR and FRR vs. Speaking rhythm

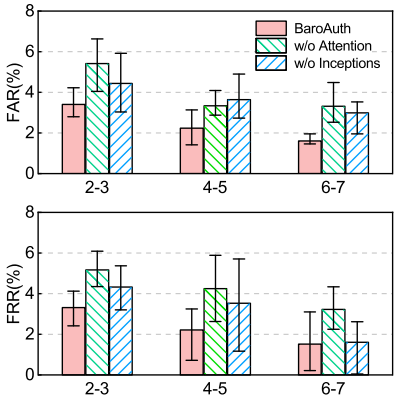


Fig. 16: FAR and FRR vs. Length of voice command

reaches 80%, the EER further decreases to 1.66%. Using the full 100% dataset, the system achieves its best performance, with an EER of just 1.57%. Overall, BaroAuth exhibits strong generalization capability under small-sample conditions, and its accuracy and robustness improve further with increasing training data. These results demonstrate its effectiveness and potential for practical applications.

2) Mobile Scenarios: We evaluate the performance of BaroAuth in three mobile scenarios (walking, climbing stairs, and taking an elevator) with and without baseline drift elimination. In Figure 14, drift elimination significantly improves performance across all scenarios. Without it, FAR and FRR rise, particularly in the "climbing stairs" and "using an elevator" scenarios. In "climbing stairs," FAR and FRR increase to 17.54% and 18.13%, respectively, but with drift elimination, they drop to 6.31% and 6.13%, improving by over 11%. In the "taking an elevator" scenario, FAR and FRR decrease from 14.69% and 16.21% to 4.22% and 4.15%. Even in the "walking" scenario, which involves minimal altitude variation, drift elimination reduces FAR by 0.6%, likely due to EMD's ability to isolate high-frequency sensor noise.

3) Speaking Rhythm: We analyze the effect of rhythm-based data augmentation on the model's robustness to different speaking rhythms. As shown in Figure 15, without augmentation, FAR and FRR fluctuate significantly under slow and fast rhythms. FAR reaches 7.55% and 8.14% for slow and fast rhythms, compared to 1.62% for regular rhythm. Similarly, FRR increases to 9.57% and 9.88%, while it is only 1.33% for regular rhythm, showing a higher rejection of legitimate users. With rhythm-based augmentation, FAR decreases to 1.87%, 1.56%, and 2.04% for slow, regular, and fast rhythms, respectively, and FRR drops to 1.55%, 1.31%, and 1.86%. These results show that augmenting speech data by varying can improve the adaptability of BaroIDNet.

4) Voice Command Length: Intuitively, longer voice commands offer stronger protection, and decreases false rejections. In Figure 16, we report the average, maximum, and minimum values of FAR and FRR for BaroAuth under different voice command lengths, along with the ablation study results for two critical system modules: the Cross Attention module and Inception Blocks. The experimental results confirm that as the length of voice commands increases, both FAR and FRR decrease, supporting our hypothesis. Specifically, when the length of voice commands increases from 2-3 words to 4-5 words, the average FAR and FRR decrease from 3.40% and 3.31% to 2.24% and 2.21%, respectively. Extending the command length further to 6-7 words reduces the average FAR and FRR to 1.61% and 1.52%, respectively, achieving the best observed performance. In addition, the ablation study reveals that removing the Cross Attention module and Inception Blocks results in a slight increase in both FAR and FRR, underscoring the importance of these modules in enhancing system performance.

5) Temporal Stability: Authentication persistence is a key metric for evaluating the usability of continuous authentication systems. In this experiment, we investigate whether the performance of BaroAuth will be affected over time. To do so, we test BaroAuth, trained on the Trace A dataset, using the

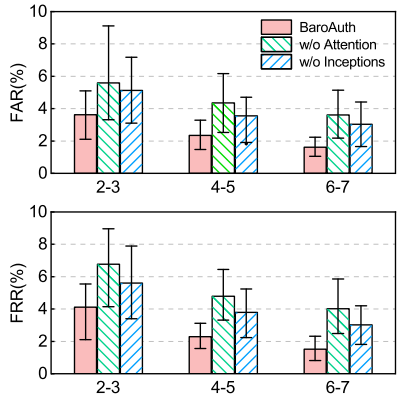


Fig. 17: Long term performance.

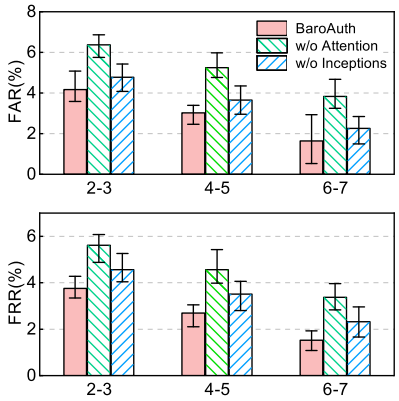


Fig. 18: FAR and FRR under zero-knowledge attack

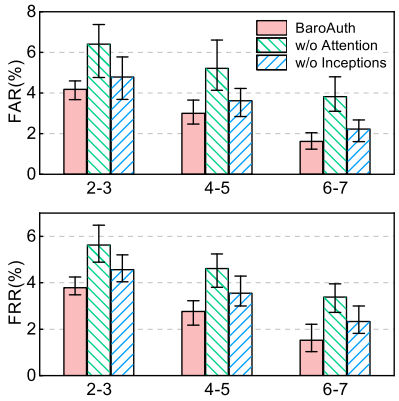


Fig. 19: FAR and FRR under knowledge-based attack

Trace C dataset. As shown in Figure 17, BaroAuth maintains stable authentication performance across all commands, with only minor fluctuations in FAR and FRR. For instance, for commands with 2–3 words, the FRR increases by only about 0.8%. For commands with 4–5 words, both FAR and FRR remain below 2.5%. For commands with 6–7 words, FAR and FRR exhibit negligible changes, demonstrating exceptional robustness. Overall, the results indicate that BaroAuth sustains stable authentication performance over prolonged usage, highlighting its strong persistence and practicality in continuous authentication tasks.

D. Attack Resistance

We examine the security of BaroAuth according to the threat model presented in Section III-A by using Trace B.

1) Zero-Knowledge Attacks: Figure 18 illustrates the performance of BaroAuth under general zero-knowledge attacks, demonstrating strong robustness against such scenarios. As shown in the table, compared to normal usage, for short commands (2-3 words and 4-5 words), the FAR increases by approximately 0.77% and 0.81%, while the FRR increases by about 0.45% and 0.49%, respectively. Although there is a slight increase, the increments remain below 1%. For long commands (6-7 words), changes in FAR and FRR are minimal, remaining almost identical to normal usage. These results indicate that even for short commands, the system exhibits limited error rate increases, while longer commands further enhance resistance to zero-knowledge attacks, ensuring stable and reliable overall performance.

2) Knowledge-based Attacks: As shown in Figure 19, despite the increased complexity of knowledge-based attacks, BaroAuth continues to demonstrate strong robustness. For short commands (2-3 words), the FAR and FRR increase by approximately 0.8% and 0.47%, respectively. For medium-length commands (4-5 words), the FAR and FRR increase by about 0.77% and 0.55%. Notably, for long commands (6-7 words), the increases in FAR and FRR are negligible, remaining almost identical to normal usage. These findings indicate that as the length of voice commands increases, the system's resistance to attacks improves significantly, effectively handling more complex attack scenarios.

TABLE III: Performance with and without Attention Modules and under Varying Numbers of Inception Blocks.

Metrics	Attention Module			Inception Blocks			
	Time	Frequency	Time+Freq.	1	3	6	9
FAR	2.23%	3.75%	**1.57%**	2.12%	**1.57%**	1.57%	1.58%
FRR	2.65%	3.16%	**1.31%**	2.13%	**1.31%**	1.30%	1.30%

E. Ablation Study

1) Effectiveness of Initial Convolution: As described in the Overall Performance section, BaroIDNet uses Conv1D to extract features from concatenated SPSs and Conv2D to process time-frequency representations, capturing complementary features between binaural signals and enhancing both time-domain and frequency-domain expressiveness. Table II shows that compared to single-ear channels, BaroIDNet reduces FAR and FRR by about 2% in binaural channel. This demonstrates that convolution not only extracts subtle differences and complementary patterns in binaural signals but also integrates time-frequency information, improving feature discrimination and overall system performance.

2) Effectiveness of Attention Module: We evaluate the impact of the cross-attention mechanism on the fusion of SPS time-domain and frequency-domain features. As shown in Table III, without cross-attention, FAR and FRR are 2.23% and 2.65% for time-domain features and 3.75% and 3.16% for frequency-domain features. This indicates that frequency-domain features are more noise-sensitive and perform worse in isolation. With cross-attention, the model captures the correlation between time-domain and frequency-domain features, leveraging their complementary information. FAR and FRR drop to 1.57% and 1.31%, the best results. This demonstrates that cross-attention dynamically focuses on key regions and effectively integrates temporal dynamics with frequency distributions, significantly improving accuracy and robustness.

3) Effectiveness of Inception Blocks: As shown in Table III, using only one Inception block results in 2.12% FAR and 2.13% FR, respectively, indicating limited single-scale feature extraction. With $three$ Inception blocks, FAR and FRR drop to 1.57% and 1.31%. However, the number of Inception

blocks increase to six and $nine$, performance gains plateau, with FAR and FRR stabilizing around 1.57%-1.58% and 1.30%-1.31%. This suggests that feature extraction saturates, and additional Inception blocks increase redundancy without meaningful improvements. Three Inception blocks balance feature extraction and computational complexity, achieving optimal performance.

VIII. Conclusion

In this paper, we propose BaroAuth, a novel biometric authentication system that utilizes SPSs captured by MEMS barometers in earbuds to extract dynamic and static physiological features, enabling efficient, robust, and accurate user authentication. By leveraging unique in-ear pressure signals, BaroAuth adapts to diverse mobile scenarios, withstands environmental noise, and resists both zero-knowledge and knowledge-based attacks. Extensive experiments demonstrate its effectiveness and robustness. However, BaroAuth has limitations, such as high FRR when the user shakes their head violently during speech. Future work will address this by using head motion data to compensate for ear pressure signal changes. Additionally, we plan to evaluate BaroAuth on various mobile devices and expand the user base to enhance generalization and applicability.

Acknowledgment

This research was supported in part by the National Natural Science Foundation of China (Grant No. 62472083, 62432008), the Natural Science Foundation of Shanghai (Grant No. 22ZR1400200), and the AI-Enhanced Research Program of Shanghai Municipal Education Commission (Grant No. SMEC-AI-DHUZ-01).

References

[1] Business Research Insights, "Earphones and headphones market size, share, growth, and industry analysis by type (in-ear and over-ear), by application (music & entertainment, sports & fitness, and gaming & virtual reality), regional insights, and forecast to 2031." https://www.businessresearchinsights.com/market-reports/earphones-and-headphones-market-101171, 2024.

[2] G. Chen, Y. Zhang, Z. Zhao, and F. Song, "{QFA2SR}:{Query-Free} adversarial transfer attacks to speaker recognition systems," in *USENIX Security Symposium*, 2023.

[3] H. Wen, S. Chang, L. Zhou, W. Liu, and H. Zhu, "Opticloak: Blinding vision-based autonomous driving systems through adversarial optical projection," *IEEE Internet of Things Journal*, vol. 11, no. 17, pp. 28931–28944, 2024.

[4] Y. Ge, P. Chen, Q. Wang, L. Zhao, N. Mou, P. Jiang, C. Wang, Q. Li, and C. Shen, "More simplicity for trainers, more opportunity for attackers:{Black-Box} attacks on speaker recognition systems by inferring feature extractor," in *USENIX Security Symposium*, 2024.

[5] J. Liu, W. Song, L. Shen, J. Han, X. Xu, and K. Ren, "Mandipass: Secure and usable user authentication via earphone imu," in *IEEE International Conference on Distributed Computing Systems, ICDCS*, 2021.

[6] Y. Zou, J. Weng, H. Lei, D. Wang, V. C. Leung, and K. Wu, "Earprint: Earphone-based implicit user authentication with behavioural and physiological acoustics," *IEEE Internet of Things Journal*, 2024.

[7] S. Choi, J. Yim, Y. Jin, Y. Gao, J. Li, and Z. Jin, "Earppg: Securing your identity with your ears," in *International Conference on Intelligent User Interfaces*, 2023.

[8] S. Chang, L. Zhou, W. Liu, H. Zhu, X. Hu, and L. Yang, "Combating voice spoofing attacks on wearables via speech movement sequences," *IEEE Transactions on Dependable and Secure Computing, TDSC*, vol. 22, no. 1, pp. 819–832, 2024.

[9] A. Ferlini, D. Ma, R. Harle, and C. Mascolo, "Eargate: gait-based user identification with in-ear microphones," in *ACM Annual International Conference on Mobile Computing and Networking, MobiCom*, 2021.

[10] Y. Xie, F. Li, Y. Wu, H. Chen, Z. Zhao, and Y. Wang, "Teethpass: Dental occlusion-based user authentication via in-ear acoustic sensing," in *IEEE Conference on Computer Communications, INFOCOM*, 2022.

[11] Z. Wang, Y. Wang, and J. Yang, "Earslide: a secure ear wearables biometric authentication based on acoustic fingerprint," *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 1, pp. 1–29, 2024.

[12] Y. Wang, T. Yang, C. Wang, F. Li, P. Hu, and Y. Shen, "Budsauth: Towards gesture-wise continuous user authentication through earbuds vibration sensing," *IEEE Internet of Things Journal*, 2024.

[13] F. Han, P. Yang, S. Yan, H. Du, and Y. Feng, "Breathsign: Transparent and continuous in-ear authentication using bone-conducted breathing biometrics," in *IEEE Conference on Computer Communications, INFOCOM*, 2023.

[14] Y. Cao, C. Cai, F. Li, Z. Chen, and J. Luo, "Heartprint: Passive heart sounds authentication exploiting in-ear microphones," in *IEEE Conference on Computer Communications, INFOCOM*, 2023.

[15] J. Li, Y. Liu, Z. Li, and J. Zhang, "Earpass: Continuous user authentication with in-ear ppg," in *ACM International Symposium on Wearable Computing*, pp. 327–332, 2023.

[16] C. Hu, X. Ma, X. Huang, Y. Shen, and D. Ma, "Lr-auth: Towards practical implementation of implicit user authentication on earbuds," *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 4, pp. 1–27, 2024.

[17] Z. Wang, S. Tan, L. Zhang, Y. Ren, Z. Wang, and J. Yang, "Eardynamic: An ear canal deformation based continuous user authentication using in-ear wearables," *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 1, pp. 1–27, 2021.

[18] A. Delnavaz and J. Voix, "Energy harvesting for in-ear devices using ear canal dynamic motion," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 1, pp. 583–590, 2013.

[19] C. Pirzanski and B. Berge, "Ear canal dynamics: Facts versus perception," *The Hearing Journal*, vol. 58, no. 10, pp. 50–52, 2005.

[20] Z. Wang, Y. Ren, Y. Chen, and J. Yang, "Toothsonic: Earable authentication via acoustic toothprint," *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 2, pp. 1–24, 2022.

[21] M. J. Grenness, J. Osborn, and W. L. Weller, "Mapping ear canal movement using area-based surface matching," *The Journal of the Acoustical Society of America*, vol. 111, no. 2, pp. 960–971, 2002.

[22] M. R. Stinson and S. M. Khanna, "Sound propagation in the ear canal and coupling to the eardrum, with measurements on model systems," *The Journal of the Acoustical Society of America*, vol. 85, no. 6, pp. 2481–2491, 1989.

[23] S. E. Voss, N. J. Horton, R. R. Woodbury, and K. N. Sheffield, "Sources of variability in reflectance measurements on normal cadaver ears," *Ear and Hearing*, vol. 29, no. 4, pp. 651–665, 2008.

[24] B. Atal and L. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 3, pp. 201–212, 1976.

[25] E. H. Norden, S. Zheng, R. L. Steven, C. W. Manli, H. S. Hsing, Z. Quanan, Y. Nai-Chyuan, C. T. Chi, and H. L. Henry, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences*, vol. 454, no. 1971, pp. 903–995, 1998.

[26] I. Cohen, Y. Huang, J. Chen, J. Benesty, J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," *Noise Reduction in Speech Processing*, pp. 1–4, 2009.

[27] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell system technical Journal*, vol. 45, no. 9, pp. 1493–1509, 1966.

[28] H. Ismail Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. F. Schmidt, J. Weber, G. I. Webb, L. Idoumghar, P.-A. Muller, and F. Petitjean, "Inceptiontime: Finding alexnet for time series classification," *Data Mining and Knowledge Discovery*, vol. 34, no. 6, pp. 1936–1962, 2020.

[29] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.

[30] T. Röddiger, T. King, D. R. Roodt, C. Clarke, and M. Beigl, "Openearable: Open hardware earable sensing platform," in *ACM International Symposium on Wearable Computers*, 2022.