LR-Auth: Towards Practical Implementation of Implicit User Authentication on Earbuds

CHANGSHUO HU, Singapore Management University, Singapore
XIAO MA, Singapore Management University, Singapore
XINGER HUANG, Shandong University, China
YIRAN SHEN, Shandong University, China
DONG MA*, Singapore Management University, Singapore

The increasing use of earbuds in applications like immersive entertainment and health monitoring necessitates effective implicit user authentication systems to preserve the privacy of sensitive data and provide personalized experiences. Existing approaches, which leverage physiological cues (e.g., jawbone structure) and behavioral cues (e.g., gait), face challenges such as limited usability, high delay and energy overhead, and significant computational demands, rendering them impractical for resource-constrained earbuds. To address these issues, we present LR-Auth, a lightweight, user-friendly implicit authentication system designed for various earbud usage scenarios. LR-Auth utilizes the modulation of sound frequencies by the user's unique occluded ear canal, generating user-specific templates through linear correlations between two audio streams instead of complex machine-learning models. Our prototype, evaluated with 30 subjects under diverse conditions, demonstrates over 99% balanced accuracy with five 100 ms audio segments, even in noisy environments and during music playback. LR-Auth significantly reduces system overhead, achieving a 20 × to 404 × decrease in latency and a 24 × to 410 × decrease in energy consumption compared to existing methods. These results highlight LR-Auth's potential for accurate, robust, and efficient user authentication on resource-constrained earbuds.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

Additional Key Words and Phrases: User Authentication, Earables, Audio Processing

ACM Reference Format:
Changshuo Hu, Xiao Ma, Xinger Huang, Yiran Shen, and Dong Ma. 2024. LR-Auth: Towards Practical Implementation of Implicit User Authentication on Earbuds. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4, Article 155 (December 2024), 27 pages. https://doi.org/10.1145/3699793

1 INTRODUCTION

Earbuds, as peripherals of mobile devices, are increasingly gaining popularity. According to a forecast by Transparency Market Research Inc. [3], the global earbuds market is expected to grow at a compound annual growth rate of 10.4% from 2023 to 2031, potentially reaching a market size of $43.9 billion by 2031. Recently, earbuds have evolved beyond mere audio playback devices into human-centered computing platforms. Multiple

*Corresponding author

Authors' Contact Information: Changshuo Hu, cs.hu.2023@phdcs.smu.edu.sg, Singapore Management University, Singapore; Xiao Ma, xiaoma.2022@phdcs.smu.edu.sg, Singapore Management University, Singapore; Xinger Huang, xinger.huang@mail.sdu.edu.cn, Shandong University, China; Yiran Shen, yiran.shen@sdu.edu.cn, Shandong University, China; Dong Ma, dongma@smu.edu.sg, Singapore Management University, Singapore.

ACM 2474-9567/2024/12-ART155
https://doi.org/10.1145/3699793

manufacturers are now incorporating personalized audio features into their flagship earbuds [1], and some models [5], equipped with independent chips and storage, can function as stand-alone devices [23]. This shift underscores the growing demand for independent authentication capabilities on earbuds. Research efforts on earbud intelligence are expanding, covering applications such as voice recognition [12, 50], health monitoring [8, 33], posture tracking[10, 31, 32], and fitness assistance [42]. However, the widespread adoption of earbuds presents new privacy and security challenges, which have garnered significant attention and research within the community. These devices increasingly access and store sensitive multi-source information, such as users' voices [21, 38], heart rates[9], blood pressure [8, 51], and frequented locations [46]. Furthermore, earbuds have significant promise as tokens that mediate access to online accounts and various devices within the Internet of Things (IoT) environment [7]. Taking these concerns into account, enabling earbuds to authenticate their wearers provides substantial benefits: 1) preventing unauthorized access to sensitive private information and resources, thereby securing user privacy; 2) providing additional services, such as personalized music genre recommendations or customized acoustic modulations to fit each user's hearing sensitivity; and 3) being part of multi-factor authentication systems and keeping paired devices like smartphones unlocked. Therefore, developing an earbud-based authentication system is of great practical importance.

Recent research in earbud authentication can be categorized into physiological-based methods such as fingerprints [56], ear canal geometry [15, 25, 30, 55], and bone structures [53], and behavioral-based approaches like gait [22], heartbeat [11, 36], and breathing patterns [27]. Despite significant advancements, existing approaches still exhibit a couple of limitations: 1) the effectiveness of these authentication methods can be highly context-dependent. For instance, microphone-based methods are greatly affected by ambient noise or internal earbud sounds [11, 24, 25, 27], and IMU-based approaches suffer from the significant interference when the users are engaged in intensive activities [53]. Additionally, variations in heartbeat or breathing patterns during vigorous exercise can affect their reliability for authentication; 2) enrolling new legitimate users often requires a lengthy and tedious process, deteriorating user experience. For example, user data needs to be collected by performing repetitive gestures (e.g., 700 gestures [53]) or remaining in a certain state for a long time (e.g, 2 hours [11]), for authentication model training; 3) to initiate authentication, specific user actions, such as walking [22] or sliding on the face [53], are required, which increases user effort and limits the application scenarios, especially in the continuous authentication setting; 4) the majority of existing methods leverage machine learning (ML) or deep learning (DL) models for authentication, which can be computationally expensive and not feasible for implementation on earbuds with limited memory, computation, and energy resources; 5) some methods actively generate probing signals (e.g., ultrasound chirps [40, 55]) or add certain additional sounds to the music [40], which can create interference or deteriorate the quality of music, as pointed out in [35]; 6) some approaches require the integration of extra sensors such as PPG sensor [15, 36], in the earbuds, which increases both the hardware cost and software complexity.

To this end, we aim to develop an implicit authentication system for earbuds that is unobtrusive, lightweight, user-friendly, and adaptable to multiple scenarios. To ensure unobtrusiveness, we selected ear canal shape as the biometric identifier, eliminating the need for users to actively perform specific actions for authentication. To achieve lightweight computation, we exploited the principle that the occluded ear canal can linearly modulate sound at different frequencies, and the variations of ear canal geometry among individuals create distinctive linear modulations (termed as user-specific templates). Calculating such linear correlation is computationally lightweight, and moreover, instead of using ML and DL, we simply computed the similarity score with the template for authentication. To reduce user burden during enrollment, we judiciously devised a synthesized audio stimulus that can enable precise user template generation within only one second. To support various real-world scenarios, we discovered two user-specific templates associated with ear canal geometry through theoretical analysis. These templates were strategically incorporated for authentication in three typical earbud usage scenarios: 1) wearing earbuds in noisy environments without music playback, 2) earbuds playing music in

quiet environments, and 3) earbuds playing music in noisy environments. By integrating these techniques, we introduce LR-Auth, a novel earbud-based implicit user authentication system. We implemented LR-Auth using an earbud prototype and conducted extensive data collection involving 30 subjects under various conditions. Our analysis demonstrated excellent and robust authentication performance of LR-Auth. Moreover, a study on latency and energy consumption revealed significant computational efficiency gains of LR-Auth compared to existing approaches.

The contributions of this work can be enumerated as:

- We have uncovered two user-specific templates adept at capturing the unique modulation of ear canal geometry on acoustic signals. These templates can be efficiently derived using linear correlation and applied to typical earbud usage scenarios for user authentication.
- We presented LR-Auth, a lightweight and unobtrusive implicit user authentication system for earbuds. It comprises a suite of carefully crafted techniques aimed at minimizing user burden during enrollment while ensuring accurate and robust authentication.
- We implemented LR-Auth with an earbud prototype and collected data with 30 subjects. Extensive experiment results demonstrated the superior performance (over 99% balanced accuracy using 500 ms signal) of LR-Auth across a range of real-world conditions.
- We conducted a latency and energy consumption study of LR-Auth on a microcontroller platform. The results show that LR-Auth achieves a significant reduction in latency, ranging from 20× to 404×, and a remarkable reduction in energy consumption, ranging from 24× to 410×, compared to existing baselines.

2 RELATED WORK

2.1 Earable Authentication

Table 1 presents a comprehensive overview of earbud authentication systems in the literature, where we compare them in various dimensions such as the sensing modality, authentication rationale, user's effort during enrollment and inference, authentication performance, system overhead, and resilience to practical factors. Next, we group these works by the sensor used for authentication, and introduce the core ideas and differences between them.

The microphone, including both in-ear and out-ear types, is the most extensively explored sensor for earbud authentication. It can be used to capture various human behaviors or biometrics for authentication. For example, EarGate [22] measures the bone-conducted sound generated by foot strikes to capture the user's gait. Teethpass [57], ToothSonic [54], and VoiceInEar [24] record the sounds produced by teeth clicking and speech to capture the unique teeth, vocal, and jaw structure, respectively. However, these works require the user to actively perform certain actions for authentication, which introduces additional overhead and may not be suitable for certain scenarios such as a quiet office room. In contrast, HeartPrint [11] and BreathSign [27] measure the unique heartbeat and breathing pattern of a user respectively, while EarEcho [25], EarDynamic [55], and Hu et al.[30] leverages the unique ear canal geometry for authentication. These approaches allow for non-invasive and implicit authentication without requiring the user's active involvement.

As another common sensor often found in existing commercial earbuds, the inertial measurement unit (IMU) has been leveraged for user authentication. Due to its operational principle, all IMU-based authentication systems require users to perform certain actions to be authenticated, such as hand gestures [53] and speech [37]. Specifically, BudsAuth [53] captures user-specific tissue compositions through finger slides on the face, while MandiPass [37] and Jawthenticate [47] capture the unique mandible structure and jaw structures respectively via vibrations generated during speech. PPG sensor, capable of measuring blood flows in vessels, has also been integrated into customized earbud prototypes for authentication. For instance, EarPPG [15] utilizes an in-ear PPG sensor to capture the ear canal deformations caused by human speech, while EarPass [36] measures the blood pulses to derive the corresponding heartbeat pattern to distinguish different users.

Table 1. A comprehensive comparison of existing earbuds authentication systems.

Paper	Sensor	Biometrics	Enrollment effort	Inference effort	Backbone	Energy consumption	Performance	Noise-proof	Music-proof
EarGate [22]	Microphone	Gait	30 seconds	Yes	SVM	Medium	92.5% (BAC)	✓	✓
HeartPrint [11]	Microphone	Heartbeat pattern	2 hours	No	CNN	High	1.6% (FAR) 1.8% (FRR)		
EarSlide [56]	Microphone	Fingerprint	200 gestures	Yes	SiameseNN	High	96.86% (ACC)		
TeethPass [57]	Microphone	Teeth structure	100 samples	Yes	SiameseNN	High	96.8% (ACC)		
ToothSonic [54]	Microphone	Teeth structure	450 samples	Yes	DNN	High	92.9% (ACC)		
BreathSign [27]	Microphone	Breathing pattern	40 samples	No	Triplet network	High	95.22% (ACC)		
VoiceInEar [24]	Microphone	Voice print	75 seconds	Yes	CycleGan	High	3.64% (EER)		
EarEcho [25]	Microphone	Ear canal geometry	400 seconds	No	SVM	Medium	94.52% (BAC)		✓
Hu et al., 2023 [30]	Microphone	Ear canal geometry	96 seconds	No	Cosine similarity	Low	4.84% (EER)	✓	
EarDynamic [55]	Microphone IMU	Ear canal geometry	15 sentences	No	Ensemble learning	Medium	93.04% (ACC)		
BudsAuth [53]	IMU	Tissue composition	700 gestures	Yes	DAL-CNN	High	5% (EER)	✓	✓
MandiPass [37]	IMU	Mandible structure	400 samples	Yes	CNN	High	1.28% (EER)		
Jawthenticate [47]	IMU	Jaw structure	50 samples	Yes	SVM	Medium	92% (BAC)		✓
EarPass [36]	PPG	Heartbeat pattern	15 minutes	No	SVM	Medium	98.7% (ACC)	✓	✓
EarPPG [15]	PPG	Ear canal geometry	180 phrases	Yes	ReGRU	High	94.2% (ACC)	✓	✓
LR-Auth (this paper)	**Microphone**	**Ear canal geometry**	**1 second**	**No**	**Cosine similarity**	**Low**	**99.8% (BAC)**	✓	✓

With variations in sensing modality and authentication rationale, these works require different levels of user's effort during the enrollment and inference stage. For example, even for the same modality - microphone, the data required during enrollment ranges from 30 seconds [22] to 2 hours [11]. Moreover, we observe that to achieve good authentication performance, most works leverage machine learning or deep learning models to verify users, resulting in excessive memory and energy overhead for resource-constrained earbuds [27, 37, 56]. Additionally, many existing systems will fail when the earbud is playing music (its primary function) [11, 56] or in a noisy environment [25, 54, 57]. In contrast, LR-Auth utilizes the unique modulation of sound frequencies by the user's ear canal and captures the linear correlation between two audio streams, instead of relying on complex machine-learning models, as the user's characteristic. This enables quick and efficient template generation in just one second, significantly reducing user efforts during enrollment. Based on this principle, LR-Auth allows for non-invasive and implicit authentication without requiring active user participation and can continuously authenticate the user throughout the entire wearing period. Moreover, LR-Auth achieves excellent authentication performance and robustness even in challenging conditions such as music playback and noisy environments, overcoming the limitations of previous works that often fail in these scenarios.

2.2 In-ear Microphone Sensing

Many off-the-shelf wireless earbuds have been integrated with in-ear microphones for the purpose of active noise cancellation (ANC). In academia, researchers have been exploring various sensing applications using the

in-ear microphone, which can be categorized into health monitoring [9, 16, 19, 29, 33, 41] and human-computing interaction (HCI) [39, 44]. For health monitoring, Christofferson et al.[16] utilized both the in-ear and out-ear microphones to identify sounds associated with poor sleep quality, such as snoring, teeth grinding, and restless movements. Martin et al.[41] leveraged the in-ear microphone to passively record the heartbeat sounds and derive heart rate in stationary cases, whereas Butkow et al. [9] extended the monitoring scenarios to various active conditions such as walking and running using a deep learning-based approach. In contrast, Fan et al.[19] employed a speaker to emit ultrasound signals to probe minute deformations of the ear canal due to heartbeat, and the reflected signals are captured by the in-ear microphone to derive heart rate. Similarly, Jin et al.[33] utilized an ultrasound-based speaker-microphone transceiver setting to monitor ear health, including the detection of ruptured eardrums, earwax buildup, and otitis media.

Interaction on wireless earbuds is an active research topic, as conventional HCI techniques (such as touch input) are not suitable for earbuds due to their unique wearing position (i.e., around the ear, thus not visible to the user). Leveraging the onboard in-ear microphone for interaction does not require any hardware modification, therefore different approaches have been investigated in the literature. Specifically, Dong et al.[39] employed in-ear microphones to detect bone-conducted vibrations incurred by finger taps on the face, which re-purposes the whole human face as an interaction interface. Prakash et al.[44] developed a system that detects teeth actions like tapping and sliding using the in-ear microphone for interaction. Using the speaker-microphone transceiver setting, Jin et al.[34] introduced EarCommand, a hands-free silent speech interface that detects ear canal deformation associated with articulator movements. To the best of our knowledge, LR-Auth is the first to spatially harness both in-ear and out-ear microphones to characterize the user's ear canal shape for implicit authentication. It offers advantages like no human effort, energy efficiency, and consistent performance across various scenarios.

3 PRELIMINARY

3.1 Sound Modulation by Ear Canal

Some prior studies have demonstrated that the geometry of an individual's ear canal is unique and can serve as a means of human authentication [49, 52]. Specifically, when the ear canal is occluded by an in-ear earbud, a cavity forms between the eardrum and the earbud. This cavity possesses a property known as acoustic impedance, which characterizes the cavity's response to acoustic pressure [48]. Essentially, acoustic impedance enhances low-frequency components and attenuates high-frequency components of sound in a linear manner, with varying ratios of enhancement or attenuation across different frequencies. Mathematically, it can be represented as:

$$R_f = \alpha_f S_f, \tag{1}$$

where S_f and R_f denote the frequency component f of the source sound S and ear canal modulated sound R, respectively. α_f refers to the modulation ratio at frequency f for a given acoustic impedance. The differences in ear canal geometry among individuals result in variations in acoustic impedance and, consequently, in the augmentation/suppression ratios across different people.

To capture the uniqueness, previous works mainly leverage the in-ear microphone sensor to record the sound transmitted by the earbud's speaker and extract various features for authentication [25, 55]. However, this approach is limited to situations where the earbud is actively playing audio. In our research, we aim to broaden the scope of authentication scenarios to more real-world conditions. Furthermore, considering the stringent energy and memory constraints of earbuds, we explore alternative signal-processing techniques for more energy-saving and memory-efficient authentication.

3.2 Linear Correlation with the Presence of Earbud Sounds

Most off-the-shelf wireless earbuds are equipped with two microphones: an out-ear microphone used for human speech acquisition and an in-ear microphone that captures in-ear sounds for active noise cancellation. When users wear earbuds and engage in daily activities such as listening to music or making phone calls, the digital signals D (such as music or conversations) from the accompanying phone are initially transformed into the analog signals through the Digital-to-Analog Converter (DAC). Subsequently, these analog signals are amplified by an audio amplifier or automatic gain control (AGC) in a frequency-selective manner and finally converted into acoustic waves by the speaker's diaphragm [35]. Considering an acoustic component D_f at frequency f, the converted component can be written as $D_f \cdot h_f$, where h_f represents the AGC amplification ratio at frequency f, which is determined by the hardware. Finally, the corresponding component captured by the in-ear microphone R_f consists of both the direct path signal from the speaker and the reflected signal from the ear canal cavity [18], as follows.

$$R_f = \beta_f \cdot (D_f \cdot h_f) + \alpha_f \cdot (D_f \cdot h_f). \tag{2}$$

Here, β_f represents the attenuation ratio of sounds (at frequency f) in the direct path, determined by the layout of the speaker and microphone. α_f denotes the modulation ratio attributed to the acoustic impedance of the ear canal cavity, as described in Equation (1).

Previous studies [18, 25] employ a complex signal processing procedure to extract the user-specific modulation ratio α_f for authentication. This procedure typically involves two major steps: 1) estimating the hardware-specific AGC amplification ratios (i.e., h_f) and 2) measuring the transfer function (i.e., β_f) of the direct path in an open space. However, we observed that these steps can be completely eliminated. Specifically, Equation (2) can be rewritten as:

$$R_f/D_f = \beta_f \cdot h_f + \alpha_f \cdot h_f. \tag{3}$$

For a given frequency f, by simply calculating the ratio between recorded signal R_f and the digital signal D_f (which is known to the earbud system), we can still obtain a user-specific ratio ($\beta_f \cdot h_f + \alpha_f \cdot h_f$), because β_f and h_f are fixed constants with a given earbud hardware.

To validate this observation, we conducted experiments with four subjects. In specific, using the earbud prototype developed in Section 5.1, we played a single sine wave at four different frequencies (300 Hz, 400 Hz, 500 Hz, and 600 Hz) with varying volume over time, and recorded the in-ear sound for each subject. Then, we split the digital signal and the recorded in-ear signal into one-second windows and applied the Fast Fourier Transform (FFT) to extract the amplitude and phase information for each window. Figure 1 (a) and (b) present the scatter plots of the amplitude and phase between the digital signal and in-ear signal, where we can observe that 1) at certain frequencies, both the amplitude and phase scatter plots present a linear correlation, and the correlation ratio can be simply derived with linear regression; 2) the ratio at different frequencies are different, determined by the AGC gain, direct path transfer function, and the geometry of the ear canal cavity. Then, Figure 1 (c) and (d) compare the correlation ratio of the four subjects at the frequency of 300Hz for amplitude and phase respectively, from which we observe that the ratios are different among different subjects. Consequently, by combining the ratios at all frequencies, we can obtain a unique correlation curve for each subject and use it for authentication.

3.3 Linear Correlation with the Presence of External Noises

Another common usage scenario of earbuds is operating in active noise cancellation mode without playing music. This often occurs when the user is studying or working in a noisy venue and desires an isolated and quiet environment. In this scenario, there is no strong sound from the speaker[1], so we can't use the linear correlation between the digital signal and the in-ear signal for user authentication. Based on the observation that both the

[1]Note that the anti-noise sound played by the speaker for active noise cancellation is usually weak and encompasses limited frequency components.

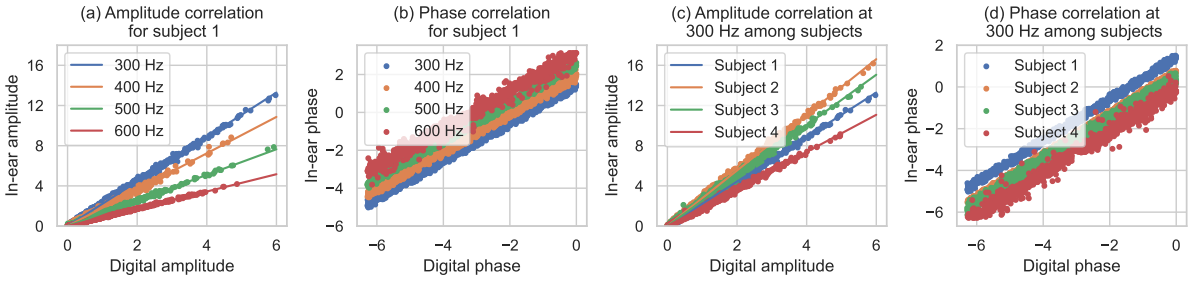

Fig. 1. With the presence of earbud sounds, the scatter plot of (a) amplitude and (b) phase between the digital signal and in-ear signal at different frequencies for subject 1; the scatter plot of (c) amplitude and (d) phase between the digital signal and in-ear signal among different subjects at the frequency of 300 Hz.

out-ear and in-ear microphones can still capture external noise, we attempt to explore whether there exists a linear correlation between these two signals, which could help distinguish different users.

Theoretically, the out-ear microphone initially captures any external noise before it propagates to the in-ear microphone. Considering a noise component N_f of frequency f at the out-ear microphone, the corresponding signal reaching the in-ear microphone R_f consists of two parts: the direct path signal that is attenuated by the earbud case, and the reflected signal that is further modulated by the ear canal cavity. Thus, R_f can be mathematically expressed as:

$$R_f = \gamma_f \cdot N_f + \alpha_f \cdot (\gamma_f \cdot N_f), \tag{4}$$

where γ_f denotes the attenuation ratio in the direct path, α_f denotes the modulation ratio as described in 3.2. The direct path and reflected path can be represented as $\gamma_f \cdot N_f$ and $\alpha_f \cdot (\gamma_f \cdot N_f)$, respectively. Consequently, Equation (4) can be rewritten as:

$$R_f / N_f = \gamma_f + \alpha_f \cdot \gamma_f. \tag{5}$$

where R_f / N_f represents the amplitude ratio between in-ear and out-ear microphone signals. As γ_f is determined by the layout and distance between the in-ear and out-ear microphones on the earbud and fixed for a given earbud [28] and α_f is influenced by the user-specific acoustic impedance, which varies with individual ear canal cavity shapes, $(\gamma_f + \alpha_f \cdot \gamma_f)$ becomes a user-specific ratio and can thus be effectively utilized for authentication.

To validate this hypothesis, we collected data from four subjects. Specifically, we played a single sine wave at four different frequencies (300 Hz, 400 Hz, 500 Hz, and 600 Hz) with varying volume over time using a smartphone near the earbud and recorded both the in-ear and out-ear microphone signals. Then, we applied the same signal processing techniques as used in Section 3.2 to the recorded signals to extract the amplitude and phase information. Figure 2 presents the scatter plots of the amplitude and phase between the in-ear and out-ear signals among different frequencies ((a) and (b)) and different subjects ((c) and (d)). We can clearly draw similar conclusions as from Figure 1, demonstrating a linear correlation between the in-ear and out-ear signals.

4 SYSTEM DESIGN

The two linear correlations derived in Section 3.2 and Section 3.3 not only offer a lightweight method for extracting user-specific features but also provide authentication rationales applicable to various earbud usage scenarios. On the one hand, rather than employing complex signal processing pipelines and extracting hand-crafted features [11, 22], our approach only requires the application of FFT and obtaining the linear ratio between two signals. On the other hand, these rationales enable the extraction of two user-specific templates (a template is defined as a series of linear ratios between two audio signals across different frequencies): **Template 1**

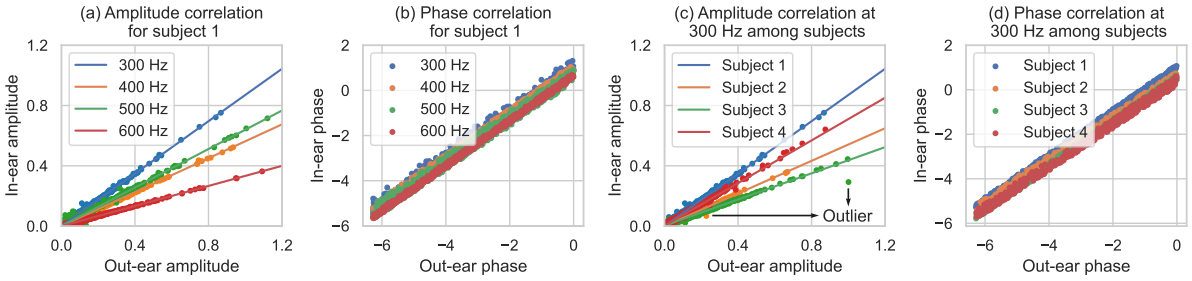

Fig. 2. With the presence of external noises, the scatter plot of (a) amplitude and (b) phase between the digital signal and in-ear signal at different frequencies for subject 1; the scatter plot of (c) amplitude and (d) phase between the digital signal and in-ear signal among different subjects at the frequency of 300 Hz.

incorporates the correlation between out-ear and in-ear signals (Equation (5)), while **Template 2** captures the correlation between the digital and in-ear signals (Equation (3)). The two templates can be utilized in three typical earbud usage scenarios: 1) **Scenario 1** - when the speaker is off in a noisy environment, 2) **Scenario 2** - when the speaker is on in a quiet environment, and 3) **Scenario 3** - when the speaker is on in a noisy environment[2]. The in-ear and out-ear microphones not only capture relevant audio signals for authentication but also serve as indicators of different scenarios. Next, we first present the attack model considered in this paper, followed by an overview of LR-Auth, before delving into the detailed designs for the enrollment stage and authentication stage.

4.1 Attack Model

We consider two major types of attacks when the adversaries possess sufficient knowledge about our system: 1) random attack, in which attackers wear the victim's earbuds and attempt to deceive our system by modulating sound through their ear canals; and 2) replay attack, in which attackers obtain audio recordings from the victim's earbuds, such as out-ear or in-ear signals, and replay these signals to fool our authentication system. However, since LR-Auth leverages the amplitude correlation between two simultaneously recorded audio streams for authentication and is independent of the audio content, any sound replayed by attackers is modulated by the wearer's ear canal. Thus, the access attempt will be accepted only when the wearer is a legitimate user. Therefore, we primarily focus on random attacks in the remainder of this paper.

4.2 System Overview

Figure 3 illustrates the flowchart of LR-Auth, which encompasses the enrollment phase and authentication phase.

- **Enrollment phase:** a user needs to enroll in two templates before using the authentication system. First, we synthesize an audio stimulus containing varying energy at different frequencies over time, which can ensure accurate acquisition of the linear ratios at different frequencies. To enroll in Template 1, the user wears the earbuds and plays the synthesized audio stimulus with a smartphone nearby, during which the in-ear and out-ear signals are recorded concurrently. For Template 2, the stimulus is played through the earbud's speaker, and only the in-ear audio is recorded[3]. For each pair of signals (i.e., out-ear & in-ear, digital & in-ear), we apply a signal processing pipeline, including downsampling, framing, FFT, outlier removal, and linear regression, to obtain the corresponding template.

[2]We do not consider the case where the speaker is off in a quiet environment, as it is uncommon in practice.
[3]The digital stimulus signal can be obtained from the playback system and is synchronized with the recorded in-ear signal

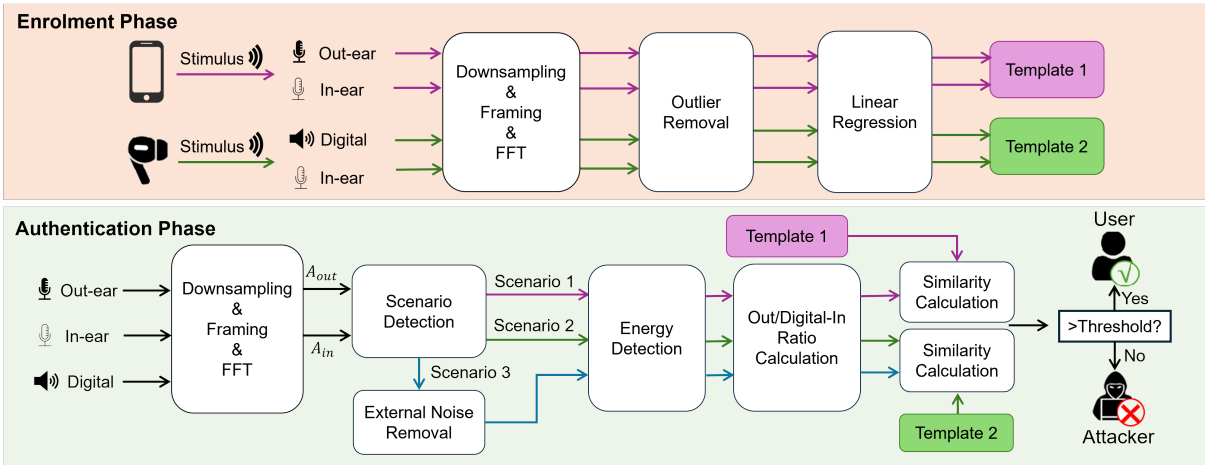

Fig. 3. The flowchart for the enrollment phase (upper) and authentication phase (lower).

- **Authentication phase:** in the authentication process, we record the in-ear and out-ear signals to detect different scenarios. Specifically, we first downsample the signals and calculate their energy using FFT. The energy levels of the two microphones can be mapped to the three scenarios discussed above. Then, for each scenario, we select the corresponding signal pairs to calculate the linear ratios and compare their similarity with the corresponding template. A user is considered legitimate if the similarity score exceeds a pre-defined threshold, and vice versa.

4.3 Enrollment Phase

4.3.1 Stimulus Synthesis. In Section 3, our investigation reveals that the amplitudes of the two signals (out-ear & in-ear or digital & in-ear) exhibit a linear correlation at each frequency, and the correlation coefficient can be derived using linear regression. To accurately characterize the coefficient, the scatter plots of amplitude pairs (e.g., Figure 2) should span a wider range, instead of being centered around similar amplitudes. To achieve this, we design an audio strength modulation scheme and use it to generate a strength-varying stimulus signal.

To obtain an audio stimulus suitable for template generation, we employ a multi-step process as follows: 1) to ensure comprehensive frequency coverage, we generate basic audio frames by combining multiple sine waves of different frequencies, each with unit amplitude. The sampling rate is set at 44.1 kHz, with frequencies ranging from 100 Hz to 4000 Hz in 5 Hz increments, (i.e., 781 sine waves). To avoid harsh periodic artifacts, we assign random phases to the sine waves. Each frame's duration is set to 0.2 seconds to mitigate frequency leakage; 2) to achieve a wider amplitude range, we randomly assign scaling vectors to each basic audio frame, with scaling ratios evenly spanning from 0 to 1 (e.g., [0.2, 0.4, 0.6, 0.8, 1.0] and [0.1, 0.3, 0.5, 0.7, 0.9]). We then create one-second audio segments by concatenating five basic frames together; 3) to evaluate the impact of stimulus length on template generation, we construct the entire stimulus using multiple one-second audio segments, with a total duration of up to 64 seconds. Each one-second segment consists of five basic audio frames that vary only in amplitude while maintaining the same randomly assigned phase. Figure 4 (a) displays a one-second audio segment composed of five basic audio frames with different amplitudes, while Figure 4 (b) illustrates the variation of the scaling vector over time. The synthesized audio stimulus ensures wide coverage of signal amplitudes even at shorter lengths, with amplitude resolution gradually increasing as the signal lengthens.

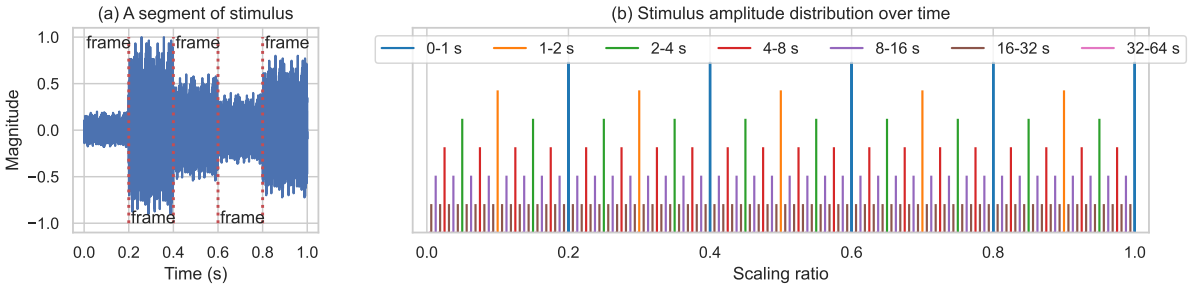

Fig. 4. (a) An audio segment modulated with different amplitude scaling ratios, and (b) the distribution of the scaling ratios over time.

4.3.2 Downsampling, Framing, and FFT. The microphone signals are recorded at a 44.1 kHz sampling rate. However, typical human voice and music mainly contain information below 4 kHz. Thus, based on the Nyquist sampling theory, we downsampled both the synthesized signal and recorded signals to 8 kHz. Subsequently, we segment the signals into frames with a length of 100 ms. Each frame then undergoes a 512-point FFT, resulting in signal amplitudes at 257 frequency components. Given that the synthesized stimulus only contains frequencies from 100 Hz to 4000 Hz, we exclude components with frequencies below 100 Hz, yielding an amplitude vector of length 250. With multiple frames from each of the two signal pairs (i.e., out-ear & in-ear, digital & in-ear), we can obtain multiple amplitude pairs as shown in Figure 1 (a).

4.3.3 Outlier Removal. Due to the inherent noise in hardware circuits and potential interference from user movements, outliers may appear in the recorded signals (e.g., Figure 2 (c)), thereby affecting the performance of linear regression for template generation, making it deviate from the ground truth value. To mitigate this issue, we first perform a linear regression on all {in-ear, out-ear} amplitude pairs in each frequency band to approximate a tentative linear curve. We then compute the vertical distance of each point (i.e., amplitude pair) to this fitted curve. These distances are standardized into z-scores and points with z-scores exceeding a pre-defined threshold (empirically set to 2), which indicates they lack a linear relationship with other points, are discarded. This threshold is chosen to optimize the removal of outliers while retaining sufficient data, eliminating approximately 5% of the total points based on our statistical analysis.

4.3.4 Template Generation. After outlier removal, we perform linear regression on the remaining amplitude pairs at each frequency component and concatenate the regression results from all 250 frequency components to generate the templates. The slope of the fitted curve represents the ratio defined in Equation (3) or Equation (5). The amplitude pairs derived from the out-ear signal and in-ear signal constitute Template 1, used for Scenario 1, while the pairs originating from the digital signal and in-ear signal constitute Template 2, used for Scenarios 2 and 3. After enrollment, each user yields two templates, serving as their unique identifiers, which are stored on the device for subsequent authentication.

4.3.5 Feasibility Assessment. To assess the effectiveness of the derived templates in capturing individual variances in ear canal geometry for user authentication, we conducted a study with 4 participants who underwent the enrollment phase. First, we examined whether the templates derived from different segments of the stimulus signal are consistent for the same subject. As shown in Figure 5 (a) and (b), the patterns of the two templates derived from different stimuli are nearly identical, indicating the high intra-subject similarity of the templates. Second, we assessed whether the templates derived for different subjects are distinguishable. As illustrated in

Figure 5 (c) and (d), both templates exhibit distinct patterns among subjects, suggesting substantial inter-subject dissimilarity. Thus, we can assert the feasibility of utilizing these templates for user authentication.

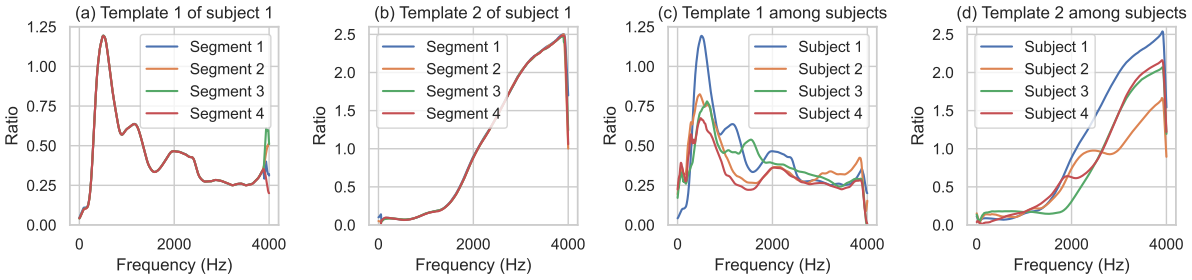

Fig. 5. (a) Template 1 and (b) Template 2 derived from different stimuli for the same subject and (c) Template 1 and (d) Template 2 derived for different subjects.

4.4 Authentication Phase

Considering diverse use cases of earbuds in daily life, our authentication system is designed to accommodate three common scenarios, outlined as follows:

- *Scenario 1:* the earbud is not playing music/speech while the user is in a noisy environment, during which the user can be authenticated using Template 1. For instance, the user is studying in a crowded cafe with ANC activated to suppress surrounding noise.
- *Scenario 2:* the earbud is playing music/speech when the user is in a quiet environment, during which the user can be authenticated using Template 2. For example, the user is listening to music in a library.
- *Scenario 3:* the earbud is playing music/speech and the user is in a noisy environment, during which the user can be authenticated using either Template 1 or Template 2 after additional signal processing. For instance, the user is listening to music while strolling along a bustling street.

4.4.1 Scenario Detection. We re-employ the in-ear and out-ear microphone to detect the current scenario, which is achieved by assessing the amplitude of the recorded signals. Specifically, the recorded signals are first downsampled to 8 kHz and decomposed into 100 ms frames. Then, we apply a 512-point FFT to the in-ear and out-ear frames and compute the average amplitudes (A_{in} and A_{out} for in-ear and out-ear signals, respectively) across all frequency components[4]. If the predefined amplitude threshold for in-ear and out-ear signals are given as Th_{in} and Th_{out}, we can determine the current scenario as follows: (1) $A_{in} < Th_{in}$ & $A_{out} > Th_{out}$ indicates Scenario 1; (2) $A_{in} > Th_{in}$ & $A_{out} < Th_{out}$ indicates Scenario 2; and (3) $A_{in} > Th_{in}$ & $A_{out} > Th_{out}$ indicates Scenario 3. Next, the system proceeds to the authentication for each scenario.

4.4.2 Authentication for Scenario 1.

- *Energy Detection:* During enrollment, we utilized a specially synthesized audio stimulus to ensure that the recorded signals encompass energy across all frequencies. This guarantees successful derivation of linear correlation coefficients at each frequency band to construct the templates. However, real-world sounds, such as music, speech, and environmental noises, may not exhibit energy at all frequencies. As a result, the acquired linear correlation ratios may be inaccurate and fail to align with the template.

[4]Note that instead of using time-domain approaches to compute the energy of the audio signals, we directly apply FFT to acquire the energy information in the frequency domain. This design has the potential for energy saving as the FFT results can be reused later to derive the signal ratios for authentication.

Hence, we examine the amplitude of each frequency component of the in-ear signal after applying FFT. If the amplitude falls below a threshold (set empirically as twice the amplitude of the baseline noise), we deem the frequency component absent in the current frame and exclude the corresponding ratio in the template for similarity calculation. Conversely, if the amplitude is high, the frequency component is deemed valid for authentication. Authentication is initiated only when the number of valid frequency components exceeds 125 (i.e., half the length of the template); otherwise, the current frame is considered invalid and discarded.

- *Amplitude Ratio Calculation and Weighting:* For the valid frequency components, we calculate the ratio between the amplitude of out-ear frame and the in-ear frame, as presented in Equation (5). These ratios form a vector and will be compared with the stored Template 1 later.
  From Figure 5 (c) and (d), we can observe that not every frequency exhibits significant variations among subjects. Specifically, in Template 1, users show minimal differences in the frequency range from 2500 to 3500, while substantial disparities are apparent from 500 Hz to 2000 Hz. To optimize the authentication performance, we assign higher weights to frequencies with greater discrepancies and lower weights to frequencies with higher similarities. To quantify these discrepancies, we calculate the standard deviation of each frequency component across 30 subjects and normalize them to a range between 0 to 1. These normalized standard deviations serve as the weights to be multiplied with the ratio of each frequency component during similarity calculation.
- *Similarity Calculation and Authentication:* We calculate the weighted cosine similarity between the valid ratios obtained after energy detection and the corresponding ratios in Template 1. If the similarity is higher than a certain threshold, the current wearer is considered a legitimate user, and vice versa. Notably, this threshold ranges between -1 and 1 and can be adjusted based on user requirements. Decreasing the threshold can enhance convenience by lowering the false rejection rate (FRR) while increasing the threshold can enhance security by reducing the false acceptance rate (FAR).

4.4.3 Authentication for Scenario 2. The authentication process for Scenario 2 is similar to that for Scenario 1. The key distinctions lie in 1) instead of calculating the amplitude ratios between out-ear and in-ear frames, we compute the ratios between the digital frames from the earbud and the in-ear frames; and 2) the derived amplitude ratios should be compared with Template 2 rather than Template 1. To maintain clarity, we will refrain from reiterating the detailed process.

4.4.4 Authentication for Scenario 3. Due to the presence of both external noises and earbud sounds, the in-ear and out-ear signals recorded in this scenario become more complex. Specifically, the in-ear signal captures a mixture of external noises and earbud sounds (i.e., the superimposition of the in-ear signals from Scenario 1 and Scenario 2, denoted as $R_{in} = R_{noise} + R_{speaker}$), while the out-ear signal comprises external noises and attenuated sounds played inside the earbuds. As the speaker faces inwards and the volume of the earbud sounds is typically low, the impact of the earbud sounds on the out-ear microphone is negligible. To validate this, we examined the out-ear signals recorded in Scenario 2 and observed an extremely low energy level. Thus, we consider the out-ear signal only contains the external noises.

To authenticate the wearer, we first obtain the corresponding in-ear noise signal (i.e., R_{noise}) by dividing the out-ear signal by Template 1[5]. Then, we subtract the in-ear noise (i.e., R_{noise}) from the in-ear signal (i.e., R_{in}), yielding the in-ear component introduced by the earbud sounds (i.e., $R_{speaker}$). As such, we convert Scenario 3 into Scenario 2 (i.e., a linear correlation between the digital signal and $R_{speaker}$) and utilize Template 2 for

[5]Note that to acquire the time domain signal, we need to account for both amplitude and phase. While amplitude can be converted with Template 1, the phase correlation is unavailable. Since out-ear and in-ear phases also present a linear correlation as depicted in Figure 2 (b), we derive another phase template to convert the out-ear phase in Scenario 3. The time domain R_{noise} is then obtained by combining the converted amplitudes and phases.

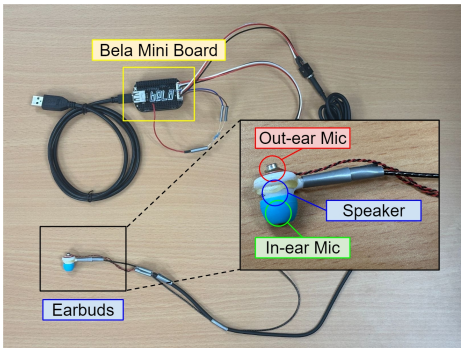

Fig. 6. The developed earbuds prototype and accompanying data recording device.

authentication. Alternatively, we can convert Scenario 3 to Scenario 1 and authenticate with Template 1. We have conducted experiments to compare the two options in Section 7.3.

5 PROTOTYPING AND DATA COLLECTION

5.1 Prototype

Our system requires a speaker and both the in-ear and out-ear microphone, components commonly found in many commercial wireless earbuds, such as Apple AirPods [1] and Huawei Freebuds [4]. However, the in-ear microphone is primarily designed for active noise cancellation, where the algorithm is executed on the onboard audio chip for minimized latency. Thus, manufacturers usually do not release the API for accessing the raw in-ear signal. To address this issue, we designed and implemented an earbud prototype to evaluate LR-Auth. As shown in Figure 6, we modified a pair of commercial earbuds with speakers (Fransun E20) by integrating two analog microphones (CMC-4015-40L100 from CUI Devices). One microphone (in-ear) was positioned at the front of the earbud speaker, facing towards the ear canal, and the other microphone (out-ear) was located outside the earbud shell, facing outward. Both microphones were connected to the audio inputs of a Bela Mini [2] development board through a 3.5mm audio jack. The speakers were connected to the audio outputs of the same board using another 3.5 mm audio jack. The Bela Mini board is equipped with an integrated development environment (IDE) for audio playback and recording control, ensuring the played and recorded audios are synchronized. The microphone sampling rate was set to 44.1kHz. We also provide silicon ear tips in three different sizes to accommodate various ear canal dimensions, thereby maintaining a good sealing quality of the earbud as well as enhancing participants' comfort during data collection.

5.2 Data Collection

After obtaining approval from the Institutional Review Board (IRB), we recruited a total of 30 participants, consisting of 17 males and 13 females, for large-scale data collection. The age range of the participants was between 20 and 32 years old, with a mean age of 22.63 and a standard deviation of 3.19. All participants possessed normal hearing and had no history of hearing function impairments. The participants were instructed to wear the earbuds prototype and maintain a good sealing quality, in a seated position. The data collection process consists of four stages as detailed below.

Stage 1 - Templates: each subject needs to generate two templates that represent the two correlations described in Equation (3) and Equation (5). For Template 1, we played the synthesized audio for 64 seconds using a smartphone (iPhone 14) positioned 40 cm away from the earbud, while recording both in-ear and out-ear microphones using

the Bela Mini Board. For Template 2, the same synthesized audio was played through the earbud's speaker, during which only the in-ear microphone was recorded.

Stage 2 - Scenario 1: this stage simulates an environment where only ambient noise is present. We selected four distinct types of external noise commonly encountered in daily life: traffic noise, public transportation noise, ambient cafe sounds, and conversational chatter. Each type of noise was played for one minute at three volume levels: low (45 dB), medium (55 dB), and high (65 dB), using the same smartphone settings. Both the in-ear and out-ear microphones were recorded simultaneously. Thus, this stage collected a total of $4 \times 3 \times 1$=12 minutes of data for each participant.

Stage 3 - Scenario 2: this stage simulates a quiet environment where the sounds are solely emitted by the earbud. We selected four types of auditory content typically presented in daily earbud usages, including male songs, female songs, human speech, and instrumental music. Each audio content was played for one minute at three different volume levels: low, medium, and high. Here, the volumes for the earbud's speaker were determined based on individual preference, corresponding to the volume for listening to music in a completely quiet environment (low), a normal working environment (medium), and a noisy environment (high). In total, this stage collected $4 \times 3 \times 1$=12 minutes of data for each participant.

Stage 4 - Scenario 3: this stage attempts to replicate an environment in which users are exposed to ambient noises while listening to sounds through earbuds. To reduce data collection overhead, we combined traffic noise and public transportation noise into a single "traffic-type" noise, and merged cafe sounds with conversational chatter into a "person-type" noise. For earbud sound types discussed in stage 2, we selected male songs and female songs. Regarding the volume, we only set two volume levels (medium and high) for each sound type. Thus, we crafted four (2×2) sound type combinations for the smartphone and earbuds, each played at four volume combinations (2×2). With a one-minute audio length, this stage collected a total of $4 \times 4 \times 1$=16 minutes of data for each participant.

With 30 subjects, we collected a total of (2+12+12+16)×30=1,260 minutes=21 hours of data.

6 EVALUATION

6.1 Evaluation Metrics

To evaluate the effectiveness of LR-Auth, we consider the following metrics with the given true positive (TP), false negative (FN), false positive (FP), and true negative (TN):

- **False Acceptance Rate (FAR):** this metric quantifies the likelihood of the system mistakenly identifying a non-legitimate user as legitimate. It is a critical measure for assessing the security of an authentication system. FAR is computed as $FAR = \frac{FP}{FP+TN}$.
- **False Rejection Rate (FRR):** this metric indicates the probability that the system incorrectly rejects a legitimate user. It is crucial for measuring the system's usability and user satisfaction. FRR is computed as $FAR = \frac{FN}{FN+TP}$.
- **Equal Error Rate (EER):** this metric represents the point at which the FAR is equal to the FRR. Specifically, it is the threshold where the system's likelihood of incorrectly accepting a non-legitimate user matches its likelihood of incorrectly rejecting a legitimate user. A lower EER indicates stronger security and better usability of the authentication system.
- **Balanced Accuracy (BAC):** this metric provides an overall measure of the system's performance. It is the average of the true positive rate (TPR) and the true negative rate (TNR), calculated as $BAC = \frac{TPR+TNR}{2}$. TPR assesses the system's ability to correctly recognize legitimate users, while TNR evaluates its effectiveness in preventing against unauthorized users.

6.2 Overall Performance

To evaluate the performance of LR-Auth, we iteratively designate one of the 30 subjects as legitimate users, while the remaining 29 subjects serve as attackers. One second out of the recorded synthesized audio during enrollment (i.e., Stage 1 in data collection) is used to derive the two user-specific templates. All the data collected during Stages 2-4 are segmented into 100 ms frames and used to assess the authentication performance in the three scenarios. Figure 7 illustrates the average FAR, FRR, and EER under different thresholds across all the subjects, for all three scenarios respectively. We can observe that with a single 100 ms frame, LR-Auth can achieve 3.22%, 5.97%, and 7.10% EER for the three scenarios respectively, demonstrating its superior performance in balancing the usability and security of the authentication system. Specifically, Scenario 2 and Scenario 3 achieve higher EER as both of them are authenticated based on Template 2, which has lower inter-subject differences. In Figure 8, we plot the BAC of each subject for all three scenarios, where we can see that the performance is consistent across different subjects.

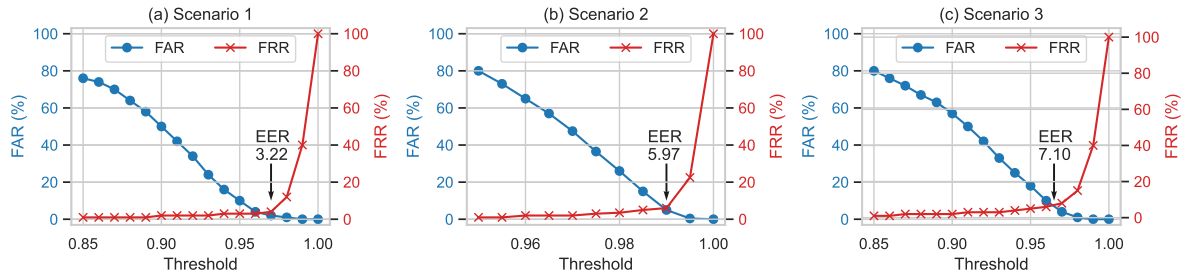

Fig. 7. FAR, FRR, and EER of (a) Scenario 1, (b) Scenario 2, and (c) Scenario 3.

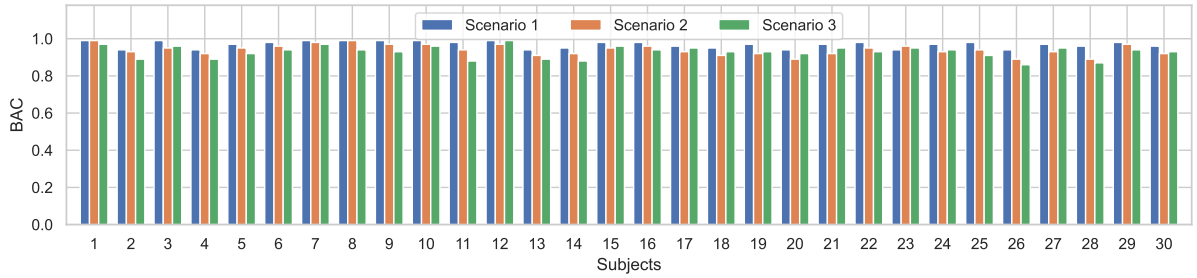

Fig. 8. Individual authentication performance of the three scenarios.

6.3 Impact of Stimulus Length

In the enrollment phase, LR-Auth requires the user to play the synthesized stimulus for deriving the two templates. To assess the user's burden during enrollment, we generate the two templates using different stimulus lengths and compare the authentication performance. Figure 9 (a) and (b) plot multiple copies of Template 1 and Template 2 generated using different stimulus lengths for one subject, where we can observe that the patterns of these templates are nearly identical. Figure 9 (c) further shows the corresponding average BAC for the seven stimulus lengths, in which we can see that the BACs are almost equal. These results imply that stimulus length has

negligible impact on the authentication performance and LR-Auth can obtain reliable and high-quality user templates with merely one second enrollment data. Such low overhead is attributed to the judicious design of the synthesized stimulus, where we assign amplitude scaling ratios in full range (i.e., 0-1) within each one-second segment. Thus, stimulus length only affects the resolution of amplitude pairs for linear regression, yet has minimal impact on the quality of linear regression (i.e., the slope of the fitted curve), as illustrated in Figure 9 (d).

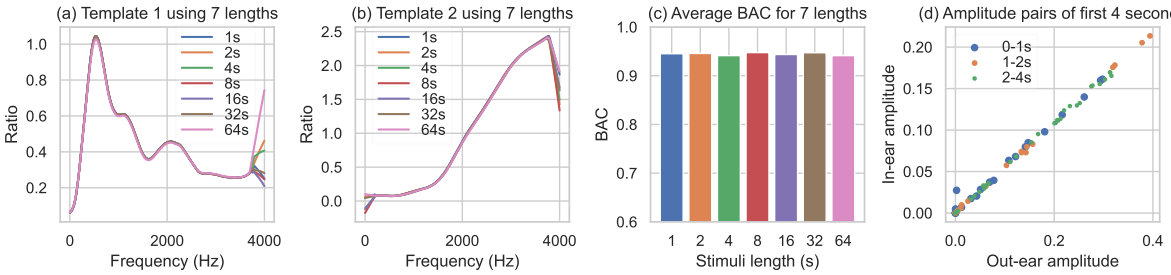

Fig. 9. (a) Template 1 and (b) Template 2 generated using stimuli of different lengths, (c) authentication performance when using different stimulus lengths for enrollment, and (d) the {out-ear, in-ear} amplitude pairs with different stimulus lengths.

6.4 Impact of the Type and Volume of External Noises

Scenario 1 relies on the external noises for authentication. However, users may wear the earbuds in various real-world environments with different noise characteristics, such as frequency and volume. To evaluate the robustness of LR-Auth against real-world noises, as presented in Section 5.2, we collected four types of environmental noise, including traffic noise (noise from the roadside), public transportation noise (noise within bus or subway), ambient cafe sounds (background noise with music and speech), and conversational chatter (human talking in a crowd), each at three different volume levels - 45, 55, and 65 dB. Figure 10 (a) displays the balanced accuracy for the four noise types, with combined volume levels. We can observe that all noise types achieve at least 95% BAC and the variance among noise types is minor, indicating that LR-Auth is resilient to different noise types. This resilience is particularly attributed to the proposed energy detection scheme. Specifically, it is unlikely that a real-world noise contains energy at all frequency bands. Thus, the ratios derived from zero or low energy frequency bands can not capture the ear canal modulation, leading to incorrect linear correlation. Instead, our energy detection scheme filters out such frequency bands, and the similarity comparison with the template is only carried out when the number of valid frequency bands exceeds half of the total frequency bands. Figure 10 (b) illustrates the ratio of frames being retained for authentication (i.e., with sufficient energy on more than half of the frequency bands) for each noise type. We can see that the authentication rate is higher in complex noisy environments with diverse frequencies (e.g., roadside traffic noise) while being lower for pure human speech with limited frequency variations (e.g, conversational chatter).

Figure 10 (c) and (d) depict the average balanced accuracy and the corresponding frame retention ratios at the three volume levels, respectively, with combined noise types. We can see that LR-Auth achieves around 96% BAC at 55 dB and 65 dB, while only 86% BAC is obtained at 45 dB. This discrepancy arises because the strength of the external volume at 45 dB is insufficient to derive a reliable amplitude ratio between the out-ear and in-ear signals. Additionally, Figure 10 (d) demonstrates that around 70% of the frames captured at 45 dB are filtered out, while most of the frames are retained for higher volume levels. This experiment reveals that Scenario 1 requires a relatively loud environment of at least 55 dB, which is very common in real-life scenarios (e.g., roadsides and restaurants)[26].

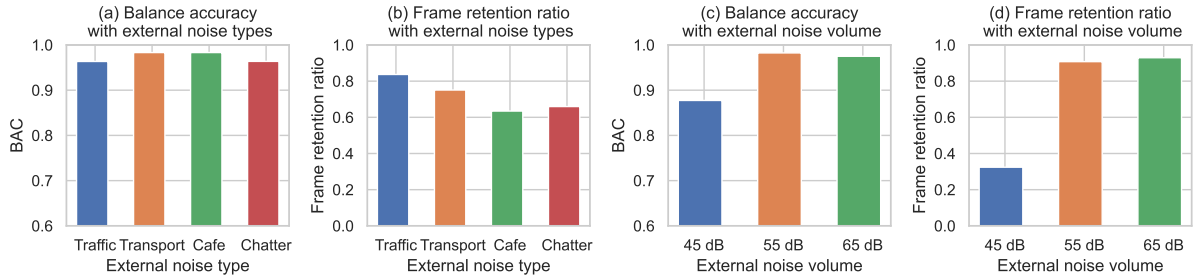

Fig. 10. (a) Balanced accuracy and (b) frame retention ratio for different external noise types, and (c) balanced accuracy and (d) frame retention ratio for different noise volumes.

6.5 Impact of the Type and Volume of Earbuds Sounds

Since the primary function of earbuds is for music playback or voice calls, it is crucial to ensure that LR-Auth remains robust when different types of sound are playing inside the earbud. Thus, we evaluate the BAC for Scenario 2 when the earbud is playing four types of sound (male songs, female songs, human speech, and instrumental music) at three different volume levels (low, medium, and high). From Figure 11 (a), we can observe that sounds combining both music and speech (e.g, male song, female song) yield better performance compared to sound with less-diverse frequency components (e.g., speech and instrument), similar to the impact of external noise type discussed in Section 6.4. Figure 11 (b) also presents a similar pattern, where more frames from sounds with few frequencies are excluded for valid authentication. Another reason for the low frame retention rate is that there exist silent periods in songs or pauses during speech, within which the frames are discarded directly.

Figure 11 (c) shows the BAC under varying earbud volume levels, where we see that the authentication performance is consistent regardless of the volume. This is because the in-ear microphone is positioned close to the speaker and facing inwards the ear canal. Even at a low volume level, the energy is sufficiently strong to derive a reliable amplitude ratio for accurate authentication. However, as illustrated in Figure 11 (d), there are actually more frames being discarded (i.e., slightly lower frame retention ratio) when the sound has a low volume. This suggests that the energy detection mechanism can effectively address issues related to low signal energy and ensure that valid frames used for authentication have high quality.

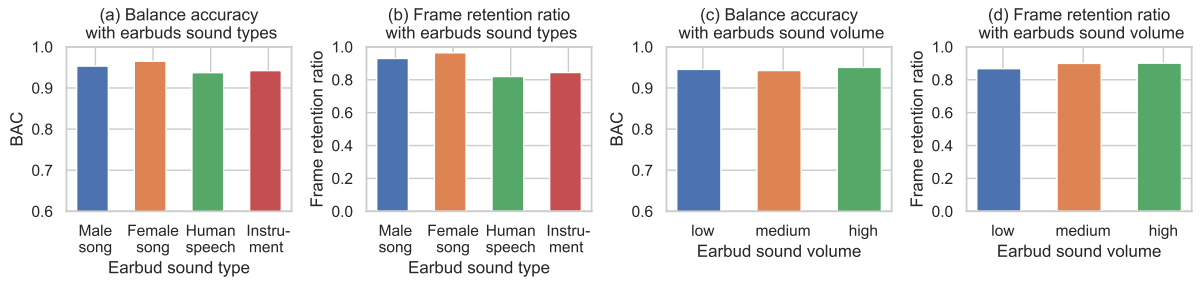

Fig. 11. (a) Balanced accuracy and (b) frame retention ratio for different earbud sound types, and (c) balanced accuracy and (d) frame retention ratio for different earbud volumes.

6.6 Impact of Multiple Wearing

Putting on and taking off earbuds are common actions in daily usage that can cause slight changes in the depth or angle of insertion of earbuds. To evaluate the robustness of LR-Auth against the slight changes during putting on and taking off, we conducted an experiment with 30 participants. The first 20 wore the earbuds consistently without adjustment during data collection, while the remaining 10 removed and rewore the earbuds during each noise or song segment, potentially altering the insertion depth and angle. The results indicate that the average BAC for those who did not re-wear the earbuds are 96.97%, 94.21%, and 92.67%, compared to 96.40%, 93.67%, and 93.37% for those who re-wore them. Moreover, previous research [25, 55] on ear canal-based authentication supports that multi-wear behavior does not significantly impact the system performance. Therefore, users can wear the earbuds normally during authentication, and slight changes in wearing position will not affect LR-Auth's accuracy.

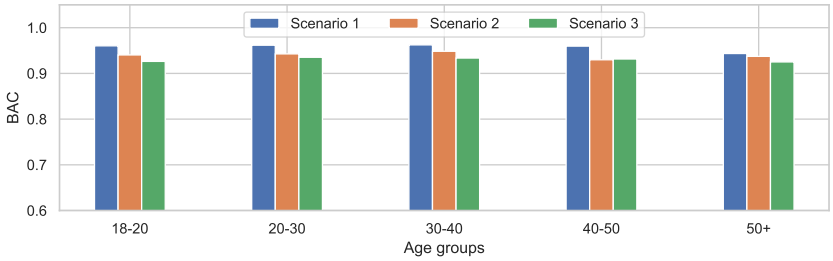

Fig. 12. Average BAC across different age groups.

6.7 Impact of User Age

The structure and skin properties of the human ear canal may gradually change with age. Previous studies [20, 45] claimed that Ear-Canal Reflectance decreases with age, but the significant effect was observed only at 1 kHz, which has minimal impact on the template of our system. To investigate the applicability of our system across a broader age range, we included additional participants spanning various age groups in our experiments. Specifically, we recruited participants aged 18-20, 30-40, 40-50, and over 50, with two individuals in each group. Figure 12 presents the average BAC for these age groups across three different scenarios. Our results demonstrate that age does not significantly impact the performance of LR-Auth, indicating the robustness of our system across diverse age demographics.

6.8 Impact of Earbud Layouts

We modified commercial earbuds by integrating both in-ear and out-ear microphones to develop a prototype for LR-Auth. The proposed system relies on templates derived from audio streams captured by these microphones. These audio streams can be influenced by factors such as the user's ear canal shape and the properties of the earbud, including microphone placement, response to speaker vibrations, and multipath reflections. Investigating whether the proposed system remains effective with different earbud layouts will be crucial for determining its feasibility across various devices and demonstrating its potential for integration into commercial earbuds.

We 3D-printed an earbud shell designed to resemble ordinary commercial models and experimented with various microphone placement combinations. These layouts were based on commercial earbuds such as the Apple AirPods Pro 2 [1], Sony WF-1000XM5 [6], and Huawei FreeBuds Pro 3 [4]. As illustrated in Figure 13, the in-ear microphone can be positioned either in front of (position 1) or behind (position 2) the speaker, while the

out-ear microphone can be located at the bottom (position 1) or the middle (position 2) of the stem. As shown Table 2, we defined four layouts based on these different microphone placement combinations and tested whether these layouts could effectively capture the templates. Additionally, we evaluated whether these templates from different users exhibited sufficient distinctiveness for authentication.

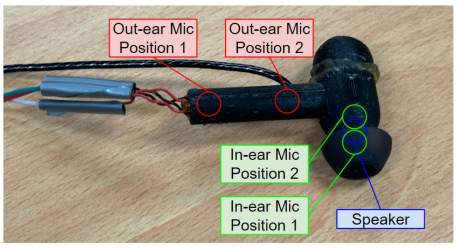

Fig. 13. Prototype with Different microphone placements.

Table 2. Earbud layouts for different microphone positions.

	Out-ear mic position 1	Out-ear mic position 2
In-ear mic position 1	Layout 1	Layout 2
In-ear mic position 2	Layout 3	Layout 4

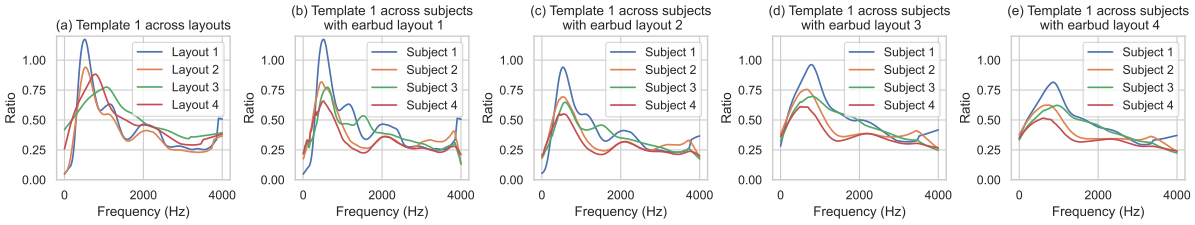

Fig. 14. (a) Template 1 with four earbud layouts for the same subject; Template 1 across four subjects with earbud layouts: (b) layout 1, (c) layout 2, (d) layout 3, and (e) layout 4.

Figure 14 (a) shows Template 1, derived from in-ear and out-ear signals, with four earbud layouts for the same subject, while Figure 14 (b), 14 (c), 14 (d), and 14 (e) display Template 1 with these layouts across different subjects. From these figures, we can observe that 1) the template exhibits different patterns across various layouts, indicating that while the ear canal shape is a unique biometric, the template generated by LR-Auth also captures the layout of the device and thus varies with different devices. This necessitates re-generating templates when changing devices, which is acceptable given that our enrolment process is very lightweight; 2) although all four layouts can generate the templates, some configurations show lower inter-subject distinctness, which is detrimental to accurately identifying different users. For instance, placing the in-ear microphone in front of the speaker ((Layouts 1 and 2) yields better templates compared to placing it behind the speaker (Layouts 3 and 4). However, the in-ear microphone is mainly used for active noise cancellation, which is supposed to be placed in front of the speaker. The position of the out-ear microphone on the stem, whether at the bottom or middle, shows similar performance. These results demonstrate that the layout can affect LR-Auth's performance to some extent.

6.9 Longitudinal Test

To evaluate the durability of LR-Auth over time, we conducted a three-week longitudinal study involving four subjects. Specifically, we scheduled a data collection session every three days, resulting in a total of seven sessions. In each session, participants first underwent the enrollment stage to derive the two templates on the specific day. Then, we collected two minutes of data for each scenario (traffic noise for Scenario 1, male song for Scenario 2, and a combination of traffic noise and male song for Scenario 3) at a medium volume for authentication.

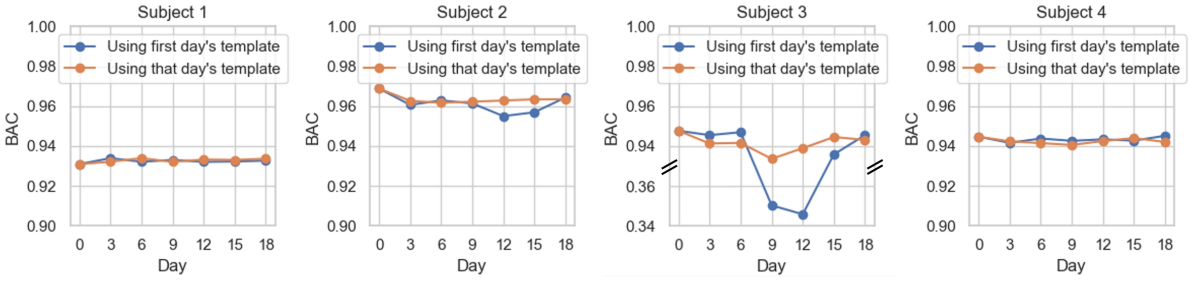

Fig. 15. Authentication performance among four subjects ((a)-(d)) when using different templates.

Figure 15 depicts the BAC of the four subjects, where the blue curve represents the BAC with templates derived on the first day, while the orange curve denotes the accuracy when the templates are derived on the same day as the authentication. From subjects 1, 2, and 4, we can observe that the accuracy remains consistent between the two cases, indicating that the templates captured by LR-Auth are stable and valid over a prolonged period. Notably, we observe a significant drop in BAC for subject 3 on day 9 and day 12 when authenticated with the templates derived from the first day. The reason is that this subject experienced ear canal inflammation during this period, which slightly altered the shape of the ear canal, thereby leading to a mismatch between the amplitude ratios and pre-registered templates. Because using earbuds during an ear infection may exacerbate the condition, we recommend that users authenticate with the other healthy ear (if available) or use alternative authentication methods during periods of inflammation. Interestingly, once the inflammation subsided on day 15, the ear canal returned to its original shape and the authentication performance resumed accordingly. On the other hand, the BAC remains consistent across all subjects when using templates derived on the same day as authentication, even during the inflammation period for subject 3.

6.10 In-the-wild Test

To assess the performance of LR-Auth under more realistic conditions, including uncontrolled external noise and daily user activities, we conducted an in-the-wild study with five participants. First, the participants underwent the enrollment stage for the derivation of the two templates in a quiet environment. Then, the participants proceeded to a roadside environment with a noise level of approximately 65 dB. Subsequently, we collected data in two sessions: one with earbuds turned off (i.e., Scenario 1) and the other with earbuds turned on for music playing (i.e., Scenario 3). During each session, the participants performed a series of activities, including standing, walking, running, speaking, and eating snacks, each for two minutes.

Figure 16 (a) and (b) illustrate the average BAC of various activities within each minute without and with earbud sounds, respectively. We observe that standing achieves the highest accuracy, with walking and running experiencing around 0.4% and 8% accuracy drop, respectively. The main reason is that the earbuds may become loosely fitted during intensive activities, leading to a shape change in the cavity formed by the earbud and ear canal. A significant accuracy drop is seen for speaking and eating, which can be attributed to three factors: 1) the sounds produced internally by speaking and eating impact both in-ear and out-ear signals through bone conduction, significantly disrupting the linear relationship; 2) speaking and chewing cause movements of the jaw, which can alter the shape of the ear canal [17, 43], deviating from the templates; 3) the speech and eating sounds may affect the amplitude of both microphones, misleading the scenario detector to use the wrong template. There are some potential solutions to address these issues. Specifically, for running, we will demonstrate how the accuracy can be enhanced to around 97% through post-processing mentioned in Section 7.2. Regarding speaking,

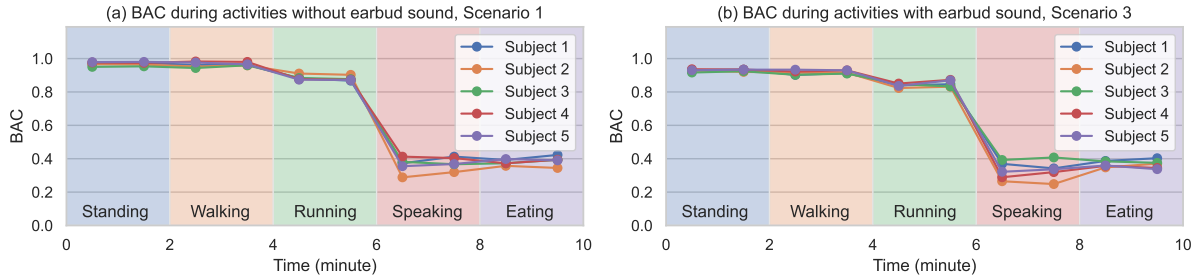

Fig. 16. Authentication performance during different activities (a) without earbud sound and (b) with earbud sound.

an additional template that considers the bone conduction path of human speech should be derived. As for chewing, given its less common occurrence during daily earbud usage scenarios, it can be safely disregarded.

6.11 System Performance

LR-Auth capitalizes on the lightweight linear correlation between two audio signals for authentication and streamlines unnecessary signal processing procedures, promising a reduction in system overhead. To demonstrate this, we conducted latency and power consumption measurements to compare our approach with four baselines EarEcho [25], EarGate [22], HeartPrint [11], and MandiPass [37]). These measurements were carried out on a microcontroller unit (MCU) STM32F767ZI with 512 KB memory and 2 MB flash. We replicated the methodology for each baseline and decomposed it into two stages: signal processing (e.g., filtering and feature extraction) and authentication (e.g., ML or DL inference). Each stage was executed for 1000 runs and the average results are presented in Table 3. Next, we first provide a brief overview of the authentication pipeline for each method and analyze the results.

- *EarEcho [25]*: an authentication result can be produced every second with two audio streams. It segments one-second audio into 150 ms frames with a 100 ms overlap, resulting in 20 frames for each stream. Each frame undergoes a 2048-point FFT to extract the amplitude features which are subsequently fed to a pre-trained binary SVM model for classification.
- *EarGate [22]*: the authentication is also conducted using one-second audio yet only one stream is needed. In its lightweight version, it only extracts 40 Mel-frequency cepstral coefficients (MFCC) and feeds them to a pre-trained binary SVM model for classification.
- *HeartPrint [11]*: the authentication is based on two 130 ms audio signals from the left and right ears. Each signal is further divided into 32 8 ms frames, from which three sets of features are extracted for each frame including 32 MFCC, 32 linear prediction coefficients (LPC), and 32 Euclidean distances between the left and right ear. In total, a $3 \times 32 \times 32$ feature matrix is extracted. These features are fed to a pre-trained CNN model with 4 layers and a total of 488,200 parameters for classification.
- *MandiPass [37]*: the authentication relies on a 200 ms six-axis IMU signal. After lowpass filtering and normalization, it feeds the processed signal with a dimension of a 6×60 to a pre-trained CNN model with 7 layers and a total of 592 million parameters for classification. Since the CNN model is 5 MB, exceeding the MCU's flash capacity (2 MB), we compressed the model using STM32Cube.AI to 1.4 MB for measurement.

From Table 3, we can observe that LR-Auth can complete an authentication event in 3.9 ms and consumes only 2.5 mJ energy, translating to a 20× - 404× reduction in latency and a 24× - 410× reduction in energy consumption, compared to the four baselines. Specifically, these reductions can be attributed to several factors: 1) compared to EarEcho and EarGate which require one-second audio data and more complex features for

Table 3. Latency and energy consumption comparison between LR-Auth and four baselines.

Methods	Operation	EarEcho	EarGate	HeartPrint	MandiPass	**LR-Auth**
Latency (ms)	Signal processing	160.2	342.8	1571.2	3.2	**2.2**
	Authentication	21.2	20.9	5.1	76.0	**1.7**
	Overall	181.4	363.7	1576.3	79.2	**3.9**
Energy consumption (mJ)	Signal processing	100.9	218.9	1021.3	2.0	**1.4**
	Authentication	13.8	13.6	3.9	57.8	**1.1**
	Overall	114.7	232.5	1025.2	59.8	**2.5**

authentication, LR-Auth only needs 100 ms data and leverages linear coefficient after FFT for authentication; 2) in contrast to HeartPrint, which extracts a multi-dimensional and complex feature matrix (e.g., LPC), our signal processing pipeline is significantly faster and lighter; and (3) although MandiPass utilizes a lightweight feature set from IMU, its CNN-based authentication model incurs heavy overhead. Moreover, linear correlation does not exist in IMU signals, making it impossible to replace MandiPass's authentication model with our lightweight method. We believe that the remarkably low system overhead of LR-Auth makes it readily suitable for on-device implementation in resource-constrained wireless earbuds.

7 DISCUSSION

7.1 Ablation Study on Template Weighting

In Section 4.4.2, we introduced a weighting technique designed to emphasize frequency components with higher inter-subject disparity by assigning higher weights to these frequencies (and vice versa) when comparing similarities. We first assess the extent to which this weighting technique enhances authentication performance, and then we evaluate whether this improvement can be generalized to unseen subjects.

Table 4 presents the BAC without and with the application of the weighting technique across three scenarios in the ablation study column. The overall weighting vector, derived from the templates of all 30 participants, was applied to each subject. The results demonstrate that introducing weighting improved the average BAC by approximately 0.61%, 2.96%, and 2.17% for the three scenarios, respectively. This suggests that the proposed weighting technique is effective in enhancing performance, particularly for Template 2, which exhibits lower inter-subject similarity and is used in Scenarios 2 and 3.

The generalization validation column in Table 4 shows the BAC without and with the application of both the overall and validation weighting vectors. We conducted a three-fold validation experiment with 30 participants, splitting them into 20 for development and 10 for validation. The overall weighting vectors were derived from the development set, while the validation weighting vectors were derived from the validation set. We then calculated the average BAC using the validation set. The results indicate that the average BAC improvements based on overall weighting vectors are 0.63%, 2.82%, and 1.91%, while the improvements based on validation weighting vectors are 0.96%, 3.26%, and 2.02%, respectively. These results demonstrate that even subjects not involved in the generation of the weighting vectors can benefit from this technique, indicating that the weighting vectors have the potential to generalize across a broader range of subjects.

7.2 Further Performance Improvement with Majority Vote

Unlike existing earbud authentication systems [53, 55, 56] that require nearly one second of data to capture the unique biometric of a user, LR-Auth can perform an authentication event with only one 100 ms frame. However, in practical scenarios, such frequent authentication may not always be necessary. This presents an opportunity

Table 4. Authentication performance without and with weighting.

	Ablation Study			Generalization Validation		
BAC	Scenario 1	Scenario 2	Scenario 3	Scenario 1	Scenario 2	Scenario 3
Without weighting	96.17%	91.07%	90.73%	96.30%	92.41%	92.23%
Overall weighting	96.78%	94.03%	92.90%	97.07%	95.23%	94.14%
Validation weighting	NA	NA	NA	97.24%	95.67%	94.25%

to enhance authentication performance by combining multiple adjacent frames. To explore this, we applied a majority voting technique to the authentication results obtained from N (an odd number) adjacent frames. Table 5 illustrates the comparison of the BAC with N set to 1, 3, and 5 frames, which demonstrates that majority voting can significantly improve authentication performance. Specifically, in Scenario 1, the BAC can be raised to over 99% with three frames (300 ms). In Scenario 2 and Scenario 3, the BAC reaches 98% with three frames and is further enhanced to 99% with five frames (500 ms). Consequently, within a reasonable latency for an authentication event (e.g., 500 ms), LR-Auth can achieve nearly optimal accuracy, outperforming existing methods [25, 30]. Moreover, majority voting can be exploited to improve the authentication performance in challenging scenarios. For instance, we applied it to the running sessions described in the in-the-wild test (Section 6.10). The results indicate that the BAC of running in both two scenarios can be improved to over 96% with a five-frame authentication, significantly lifting the authentication performance.

Table 5. Authentication performance with majority voting.

BAC	Scenario 1	Scenario 2	Scenario 3	Running (Scenario 1)	Running (Scenario 3)
One frame	96.78%	94.03%	92.90%	87.47%	84.92%
Three frames	99.71%	98.97%	98.56%	95.67%	93.37%
Five frames	99.97%	99.80%	99.68%	98.37%	96.98%

7.3 Comparison of Music-based and Noise-based Approaches for Scenario 3

Scenario 3 involves both external noise and earbud sound, making it impossible to directly authenticate with any one of the two templates. To tackle this issue, as presented in Section 4.4.4, we first utilize Template 1 to transform the external noise into its in-ear counterpart, which is then subtracted from the original in-ear signal to isolate the in-ear earbud sound. As a result, the digital sound and isolated in-ear earbud sound can be used for authentication with Template 2 (which refers to a music-based approach, converting Scenario 3 to Scenario 2). Alternatively, another approach is to convert the digital signal using Template 2 and authenticate with Template 1 after subtraction (refers to noise-based approach, converting Scenario 3 to Scenario 1). Thus, we conducted experiments on the data collected for Scenario 3 to compare the efficacy of the two approaches.

Table 6 presents the BAC of music-based and noise-based methods on data with different volume level combinations. The results indicate that the two approaches achieve similar performance, with music-based method slightly outperforming the noise-based approach. Under various volume settings, the music-based method performs better with high-volume music, while the noise-based method excels in high-volume noise environments. Thus, to achieve optimal performance, it would be possible to select one of the two approaches based on the relative volumes of external noise and earbud sound.

Table 6. Authentication performance of music-based and noise-based approaches.

BAC	Medium noise, medium music	Medium noise, high music	High noise, medium music	High noise, high music	Average
Music-based (Scenario 3 to Scenario 2)	92.36%	93.59%	92.44%	93.21%	92.90%
Noise-based (Scenario 3 to Scenario 1)	92.18%	90.13%	92.95%	92.56%	91.96%

7.4 Potential Performance Improvement with Both Earbuds

Currently, LR-Auth utilizes only the left earbud to authenticate users by analyzing the unique characteristics of the left ear canal shape through audio signals. However, as noted in previous studies [49, 58], the geometry of the left and right ear canals differs significantly. Therefore, incorporating both earbuds could potentially enhance the performance and robustness of LR-Auth. To achieve this, users would need to generate four templates (two for each ear) during the enrollment stage. Since the synthesized stimulus played by the smartphone and earbud speakers can be captured by the microphones on both the left and right earbuds simultaneously, deriving the two additional templates will not incur extra user overhead.

7.5 User Template Generation and Adaptation

While LR-Auth already boasts a very low enrollment overhead, there is still an opportunity to completely eliminate users' enrollment burden. Similar to the adaptive learning mechanisms in iPhone's Face ID [14], LR-Auth can passively collect in-ear and out-ear microphone data while the user is engaged in various daily scenarios. Over time, the system can monitor the similarity between each authentication attempt and the existing template. When deviations are detected, such as those caused by gradual physiological changes like aging, the system can automatically update the template by integrating new data points using techniques like weighted averaging or incremental linear regression[13]. This update process can occur seamlessly in the background, without notifying or disturbing the user.

7.6 Limitations of LR-Auth

As mentioned in Section 6.9 and Section 6.10, LR-Auth cannot effectively authenticate users under certain extreme conditions, such as when a user suffers from ear inflammation or while speaking and eating. Ear inflammation significantly alters the shape of the ear canal in short-term, and jaw movements during speaking and eating also change the ear canal shape dynamically. These activities impact the in-ear and out-ear audio streams nonlinearly, disrupting the linear correlation and causing the scenario detector to use the wrong template. Additionally, LR-Auth cannot authenticate users in a quiet environment without music playback due to insufficient audio signal energy to calculate the linear correlation. However, this is considered a less common scenario in practical earbuds usage and users are recommended to utilize alternative authentication approaches in such cases. While LR-Auth is demonstrated to be performed well with 30 participants, its performance in more diverse demographics and large-scale populations needs further exploration. This is because ear canal shapes can appear similarly at a coarse level, thus more fine-grained templates may be necessary to ensure robust authentication in larger deployments.

8 CONCLUSION

This paper presents LR-Auth, a novel lightweight and non-invasive user authentication framework that capitalizes on the unique linear correlations between two audio streams (i.e., out-ear & in-ear, digital & in-ear) from the earbud. LR-Auth comprises a fast enrollment phase to derive the two user-specific templates, and a lightweight and

non-invasive user authentication phase. Particularly, the two templates are judiciously employed to authenticate users in three common earbud usage scenarios. With data collected from 30 subjects, we showcase LR-Auth's exceptional authentication performance across various real-world conditions. Moreover, our latency and energy measurement study suggests that LR-Auth achieves a remarkable reduction in system overhead compared to state-of-the-art approaches, paving the way for real-world implementation on resource-constrained earbuds.

REFERENCES

[1] Online. AirPods Pro (2nd generation). https://www.apple.com/sg/airpods-pro/.

[2] Online. Bela Mini board. https://learn.bela.io/products/bela-boards/bela-mini/.

[3] Online. Earbuds Market, Transparency Market Research, Inc. https://finance.yahoo.com/news/earbuds-market-expected-expand-compound-122300780.html.

[4] Online. Huawei Freebuds 5i. https://consumer.huawei.com/en/headphones/freebuds5i/.

[5] Online. OpenSwim Waterproof Swimming Headphone. https://shokz.com/products/openswim.

[6] Online. Sony WF-1000XM5. https://www.sony.com.sg/headphones/products/wf-1000xm5.

[7] Brett Barros and Lauren Wunderlich. 2024. Authentication State Preservation by Peripherals to Enable Personal Data Access from Paired Primary Devices. (2024).

[8] Nam Bui, Nhat Pham, Jessica Jacqueline Barnitz, Zhanan Zou, Phuc Nguyen, Hoang Truong, Taeho Kim, Nicholas Farrow, Anh Nguyen, Jianliang Xiao, et al. 2019. ebp: A wearable system for frequent and comfortable blood pressure monitoring from user's ear. In *The 25th annual international conference on mobile computing and networking*. 1–17.

[9] Kayla-Jade Butkow, Ting Dang, Andrea Ferlini, Dong Ma, and Cecilia Mascolo. 2023. heart: Motion-resilient heart rate monitoring with in-ear microphones. In *2023 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 200–209.

[10] Gaoshuai Cao, Kuang Yuan, Jie Xiong, Panlong Yang, Yubo Yan, Hao Zhou, and Xiang-Yang Li. 2020. Earphonetrack: involving earphones into the ecosystem of acoustic motion tracking. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 95–108.

[11] Yetong Cao, Chao Cai, Fan Li, Zhe Chen, and Jun Luo. 2023. HeartPrint: Passive Heart Sounds Authentication Exploiting In-Ear Microphones. In *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*. IEEE, 1–10.

[12] Yetong Cao, Fan Li, Huijie Chen, Xiaochen Liu, Shengchun Zhai, Song Yang, and Yu Wang. 2023. Live Speech Recognition via Earphone Motion Sensors. *IEEE Transactions on Mobile Computing* (2023).

[13] Jagmohan Chauhan, Young D Kwon, Pan Hui, and Cecilia Mascolo. 2020. Contauth: Continual learning framework for behavioral-based user authentication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–23.

[14] Ivan Cherapau, Ildar Muslukhov, Nalin Asanka, and Konstantin Beznosov. 2015. On the Impact of Touch {ID} on {iPhone} Passcodes. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*. 257–276.

[15] Seokmin Choi, Junghwan Yim, Yincheng Jin, Yang Gao, Jiyang Li, and Zhanpeng Jin. 2023. EarPPG: Securing Your Identity with Your Ears. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 835–849.

[16] Kenneth Christofferson, Xuyang Chen, Zeyu Wang, Alex Mariakakis, and Yuntao Wang. 2022. Sleep Sound Classification Using ANC-Enabled Earbuds. In *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. 397–402. https://doi.org/10.1109/PerComWorkshops53856.2022.9767394

[17] Aidin Delnavaz and Jérémie Voix. 2013. Energy harvesting for in-ear devices using ear canal dynamic motion. *IEEE Transactions on Industrial Electronics* 61, 1 (2013), 583–590.

[18] Berken Utku Demirel, Ting Dang, Khaldoon Al-Naimi, Fahim Kawsar, and Alessandro Montanari. 2024. Unobtrusive Air Leakage Estimation for Earables with In-ear Microphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 4 (2024), 1–29.

[19] Xiaoran Fan, David Pearl, Richard Howard, Longfei Shangguan, and Trausti Thormundsson. 2023. APG: Audioplethysmography for Cardiac Monitoring in Hearables. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*. 1–15.

[20] M Patrick Feeney and Chris A Sanford. 2004. Age effects in the human middle ear: Wideband acoustical measures. *The Journal of the Acoustical Society of America* 116, 6 (2004), 3546–3558.

[21] Huan Feng, Kassem Fawaz, and Kang G Shin. 2017. Continuous authentication for voice assistants. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*. 343–355.

[22] Andrea Ferlini, Dong Ma, Robert Harle, and Cecilia Mascolo. 2021. EarGate: gait-based user identification with in-ear microphones. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 337–349.

[23] Andrea Ferlini, Dong Ma, Lorena Qendro, and Cecilia Mascolo. 2022. Mobile health with head-worn devices: Challenges and opportunities. *IEEE Pervasive Computing* 21, 3 (2022), 52–60.

[24] Yang Gao, Yincheng Jin, Jagmohan Chauhan, Seokmin Choi, Jiyang Li, and Zhanpeng Jin. 2021. Voice in ear: Spoofing-resistant and passphrase-independent body sound authentication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*

5, 1 (2021), 1–25.
[25] Yang Gao, Wei Wang, Vir V Phoha, Wei Sun, and Zhanpeng Jin. 2019. EarEcho: Using ear canal echo for wearable authentication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–24.
[26] Monica S Hammer, Tracy K Swinburn, and Richard L Neitzel. 2014. Environmental noise pollution in the United States: developing an effective public health response. *Environmental health perspectives* 122, 2 (2014), 115–119.
[27] Feiyu Han, Panlong Yang, Shaojie Yan, Haohua Du, and Yuanhao Feng. 2023. Breathsign: Transparent and continuous in-ear authentication using bone-conducted breathing biometrics. In *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*. IEEE, 1–10.
[28] Cyril M Harris. 1966. Absorption of sound in air versus humidity and temperature. *The Journal of the Acoustical Society of America* 40, 1 (1966), 148–159.
[29] Changshuo Hu, Thivya Kandappu, Yang Liu, Cecilia Mascolo, and Dong Ma. 2024. BreathPro: Monitoring Breathing Mode during Running with Earables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2 (2024), 1–25.
[30] Changshuo Hu, Xiao Ma, Dong Ma, and Ting Dang. 2023. Lightweight and Non-Invasive User Authentication on Earables. In *Proceedings of the 24th International Workshop on Mobile Computing Systems and Applications*. 36–41.
[31] Jingyang Hu, Hongbo Jiang, Daibo Liu, Zhu Xiao, Qibo Zhang, Jiangchuan Liu, and Schahram Dustdar. 2023. Combining IMU With Acoustics for Head Motion Tracking Leveraging Wireless Earphone. *IEEE Transactions on Mobile Computing* (2023).
[32] Nan Jiang, Terence Sim, and Jun Han. 2022. EarWalk: towards walking posture identification using earables. In *Proceedings of the 23rd Annual International Workshop on Mobile Computing Systems and Applications*. 35–40.
[33] Yincheng Jin, Yang Gao, Xiaotao Guo, Jun Wen, Zhengxiong Li, and Zhanpeng Jin. 2022. Earhealth: an earphone-based acoustic otoscope for detection of multiple ear diseases in daily life. In *Proceedings of the 20th annual international conference on mobile systems, applications and services*. 397–408.
[34] Yincheng Jin, Yang Gao, Xuhai Xu, Seokmin Choi, Jiyang Li, Feng Liu, Zhengxiong Li, and Zhanpeng Jin. 2022. EarCommand: " Hearing" Your Silent Speech Commands In Ear. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–28.
[35] Dong Li, Shirui Cao, Sunghoon Ivan Lee, and Jie Xiong. 2022. Experience: practical problems for acoustic sensing. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*. 381–390.
[36] Jiao Li, Yang Liu, Zhenjiang Li, and Jin Zhang. 2023. EarPass: Continuous User Authentication with In-ear PPG. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing*. 327–332.
[37] Jianwei Liu, Wenfan Song, Leming Shen, Jinsong Han, Xian Xu, and Kui Ren. 2021. Mandipass: Secure and usable user authentication via earphone imu. In *2021 IEEE 41st International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 674–684.
[38] Dong Ma, Ting Dang, Ming Ding, and Rajesh Balan. 2024. ClearSpeech: Improving Voice Quality of Earbuds Using Both In-Ear and Out-Ear Microphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 4 (2024), 1–25.
[39] Dong Ma, Andrea Ferlini, and Cecilia Mascolo. 2021. Oesense: employing occlusion effect for in-ear human sensing. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*. 175–187.
[40] Shivangi Mahto, Takayuki Arakawa, and Takafumi Koshinak. 2018. Ear acoustic biometrics using inaudible signals and its application to continuous user authentication. In *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 1407–1411.
[41] Alexis Martin and Jérémie Voix. 2017. In-ear audio wearable: Measurement of heart and breathing rates for health and safety monitoring. *IEEE Transactions on Biomedical Engineering* 65, 6 (2017), 1256–1263.
[42] Anouk Nijs, Peter J Beek, and Melvyn Roerdink. 2021. Reliability and validity of running cadence and stance time derived from instrumented wireless earbuds. *Sensors* 21, 23 (2021), 7995.
[43] Chester Pirzanski and Brenda Berge. 2005. Ear canal dynamics: Facts versus perception. *The Hearing Journal* 58, 10 (2005), 50–52.
[44] Jay Prakash, Zhijian Yang, Yu-Lin Wei, Haitham Hassanieh, and Romit Roy Choudhury. 2020. EarSense: earphones as a teeth activity sensor. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–13.
[45] John J Rosowski, Hideko H Nakajima, Mohamad A Hamade, Lorice Mahfoud, Gabrielle R Merchant, Christopher F Halpin, and Saumil N Merchant. 2012. Ear-canal reflectance, umbo velocity, and tympanometry in normal-hearing adults. *Ear and hearing* 33, 1 (2012), 19–34.
[46] Darius Satongar, Chris Pike, Yiu W Lam, and Anthony I Tew. 2015. The influence of headphones on the localization of external loudspeaker sources. *Journal of the Audio Engineering Society* 63, 10 (2015).
[47] Tanmay Srivastava, Shijia Pan, Phuc Nguyen, and Shubham Jain. 2023. Jawthenticate: Microphone-free Speech-based Authentication using Jaw Motion and Facial Vibrations. (2023).
[48] Stefan Stenfelt and Sabine Reinfeldt. 2007. A model of the occlusion effect with bone-conducted stimulation. *International journal of audiology* 46, 10 (2007), 595–608.
[49] Michael R Stinson and BW Lawton. 1989. Specification of the geometry of the human ear canal for the prediction of sound-pressure level distribution. *The Journal of the Acoustical Society of America* 85, 6 (1989), 2492–2503.
[50] Xue Sun, Jie Xiong, Chao Feng, Haoyu Li, Yuli Wu, Dingyi Fang, and Xiaojiang Chen. 2024. EarSSR: Silent Speech Recognition via Earphones. *IEEE Transactions on Mobile Computing* (2024).

[51] Hoang Truong, Alessandro Montanari, and Fahim Kawsar. 2022. Non-invasive blood pressure monitoring with multi-modal in-ear sensing. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6–10.

[52] Susan E Voss and Jont B Allen. 1994. Measurement of acoustic impedance and reflectance in the human ear canal. *The Journal of the Acoustical Society of America* 95, 1 (1994), 372–384.

[53] Yong Wang, Tianyu Yang, Chunxiao Wang, Feng Li, Pengfei Hu, and Yiran Shen. 2024. BudsAuth: Towards Gesture-Wise Continuous User Authentication Through Earbuds Vibration Sensing. *IEEE Internet of Things Journal* (2024).

[54] Zi Wang, Yili Ren, Yingying Chen, and Jie Yang. 2022. Toothsonic: Earable authentication via acoustic toothprint. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–24.

[55] Zi Wang, Sheng Tan, Linghan Zhang, Yili Ren, Zhi Wang, and Jie Yang. 2021. Eardynamic: An ear canal deformation based continuous user authentication using in-ear wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–27.

[56] Zi Wang, Yilin Wang, and Jie Yang. 2024. EarSlide: a Secure Ear Wearables Biometric Authentication Based on Acoustic Fingerprint. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 1 (2024), 1–29.

[57] Yadong Xie, Fan Li, Yue Wu, Huijie Chen, Zhiyuan Zhao, and Yu Wang. 2022. Teethpass: Dental occlusion-based user authentication via in-ear acoustic sensing. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 1789–1798.

[58] Jen-Fang Yu, Kun-Che Lee, Ren-Hung Wang, Yen-Sheng Chen, Chun-Chieh Fan, Ying-Chin Peng, Tsung-Hsien Tu, Ching-I Chen, and Kuei-Yi Lin. 2015. Anthropometry of external auditory canal by non-contactable measurement. *Applied ergonomics* 50 (2015), 50–55.