HeadSonic: Usable Bone Conduction Earphone Authentication via Head-Conducted Sounds

Zhixiang He, Jing Chen, *Senior Member, IEEE*, Kun He, Yangyang Gu, Qiyi Deng, Zijian Zhang, *Senior Member, IEEE*, Ruiying Du, Qingchuan Zhao, and Cong Wu

***Abstract*—Earables (ear wearables) are rapidly emerging as a new platform encompassing a diverse of personal applications, prompting the development of authentication schemes to protect user privacy. Existing earable authentication methods are all specifically designed for air-conduction earphones, which are not suited for bone conduction earphones (BCEs) that rely on bone conduction mechanisms. In this paper, we propose HeadSonic, a usable BCE authentication system based on the unique head-conducted sounds, which can be acquired when the user wears the BCE device. Specifically, the system emits a millisecond-level sound to initiate the authentication session. The signal captured by the BCE microphone is propagated through the user's head, which is unique in density, geometry, and bone-tissue ratio. It operates implicitly, while maintaining robustness across different behaviors. Extensive experiments involving 60 subjects demonstrate that HeadSonic achieves a commendable balanced accuracy of 96.59%, proving its efficacy and resilience against replay and synthesis attacks. Our dataset and source codes are available at https://anonymous.4open.science/r/HeadSonic-1CE4.**

***Index Terms*—Wearable authentication, biometrics, acoustic sensing.**

Received 27 November 2024; revised 10 March 2025; accepted 11 March 2025. Date of publication 13 March 2025; date of current version 6 August 2025. This work was supported in part by the State Key Lab of Intelligent Transportation System under Grant 2024-B004, in part by the National Natural Science Foundation of China under Grant 62172303 and Grant 62472323, in part by the Key R&D Program of Hubei Province under Grant 2024BAB018, in part by the Wuhan Scientific and Technical Achievements Project under Grant 2024030803010172, and in part by the Key R&D Program of Shandong Province under Grant 2022CXPT055. Recommended for acceptance by L. Guo. *(Corresponding author: Jing Chen.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board of Wuhan University under Application No. WHU-NS-IRB2023003.

Zhixiang He and Kun He are with the Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China, and also with the State Key Lab of Intelligent Transportation System, Beijing 100191, China (e-mail: zhixianghe@whu.edu.cn; hekun@whu.edu.cn).

Jing Chen, Yangyang Gu, Qiyi Deng, and Ruiying Du are with the School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China (e-mail: chenjing@whu.edu.cn; guyangyang@whu.edu.cn; qiyideng@whu.edu.cn; duraying@whu.edu.cn).

Zijian Zhang is with the School of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing 100811, China (e-mail: zhangzijian@bit.edu.cn).

Qingchuan Zhao is with the Department of Computer Science, City University of Hong Kong,, Hong Kong (e-mail: cs.qczhao@cityu.edu.hk).

Cong Wu is with the Department of Electrical and Electronic Engineering, University of Hong Kong,, Hong Kong (e-mail: congwu@hku.hk).

Digital Object Identifier 10.1109/TMC.2025.3551272

I. Introduction

EARABLES, with their rich around-the-head sensing capabilities, are rapidly emerging as a new platform for a broad range of personal applications [1], [2]. Particularly, bone conduction earphones (BCEs) stand out, finding increasing adoption in the military, sports, and hearing aid fields [3]. The widespread use of BCEs raises new security concerns, as they hold the potential as other earables, enabling inference of user's health metrics [4], [5], private activities [6], [7], [8], and serving as access tokens for smart devices [1], [9]. The international data corporation reports that the BCE market hits $876.3 million in 2023, with an anticipated compound annual growth rate of 21.8% [3]. Therefore, secure authentication is crucial for BCEs to prevent unauthorized access to sensitive data and services.

Existing earable-based authentication methods fall into passive and active sensing two categories. Passive sensing methods capture signals from the ear canal passively, such as sounds from teeth occlusion [10], [11], finger-face sliding [12], footstep [13], heartbeat [14], and breathing [9]. Active sensing methods emit signals into the ear canal to characterize its static geometry [15], [16] or dynamic deformation [17] actively. However, all these studies rely on capturing target biometrics from the ear canal with in-ear microphones, which BCEs do not have. Therefore, existing earable-based methods can not be directly applied to BCEs. As depicted in Fig. 1, sound waves of a BCE travel through the skull rather than the ear canal to reach the cochlea, facilitating auditory perception. This unique signal transmitting mechanism necessitates a specialized authentication solution.

In this work, we propose HeadSonic, a usable BCE authentication system based on characterizing head biometrics in the acoustic domain. Specifically, a user who puts on a BCE device typically adjusts the clip to ensure a snug fit. During authentication, the proposed system emits a devised millisecond-level acoustic signal through the BCE speaker to actively sense the user's head and collect the responses using the microphone. HeadSonic is based on the fact that individuals have unique head density, geometry, and bone-tissue ratio [10], leading to distinct signal attenuation and reflection during head-conduction. Being implicit, HeadSonic does not impose any additional demands on users.

We face several challenges to realize HeadSonic. First, HeadSonic measures the acoustic-domain head biometrics using BCE's unique head contact architecture, an approach that has not been studied. How to delineate the sensing mechanism

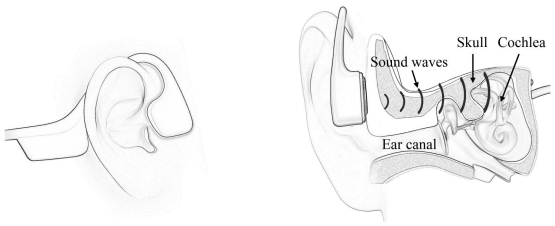


Fig. 1. Wearing position and sound transmission mechanism of BCEs.

and characterize the head's physiological traits embedded in the received signal is a challenge. Second, human motion can distort the signal. For example, when a person is walking, the relative position between a BCE and the head drifts due to inertia, resulting in a different speaker-mic channel for sound wave propagation, which introduces variability of the received signal and may compromise the authentication performance. Therefore, how to eliminate the influence of different behaviors to achieve behavior-irrelevant authentication is a challenge.

To tackle the first challenge, we analyze the signal's multiple transmission paths and reveal the head's impact on its propagation (Section II-B). Building on the analysis, we design novel representations that characterize human head conduction (Section III-D1) and geometry (Section III-D2) characteristics, which have not been explored in prior authentication studies. To address the second challenge, we design a feature reconstruction model based on a Behavior-Adaptation Neural Network (BANN) to obtain behavior-irrelevant features (Section III-D3). The BANN model comprises a CNN-based feature reconstructor, a Transformer-based behavior discriminator, and a GRU-based user classifier. It takes the extracted features as input for reconstruction, and then feeds the reconstructed features into the behavioral discriminator and the user classifier to eliminate behavioral relevance with adversarial learning and increase user distinctiveness, respectively. The reconstructed features are then used to determine the user's identity (Section III-E).

We implement HeadSonic by using 4 kinds of BCEs. We recruit 60 participants (46 males and 14 females) and ask them to put on BCEs for authentication in diverse scenarios. We simulate 5 types of attacks to test the anti-attack ability of our system. The results demonstrate that HeadSonic is accurate across different scenarios and can resist various spoofing attacks. Our contributions can be summarized as follows.

- We propose HeadSonic, a usable BCE authentication system, leveraging BCE's unique head-contact architecture to capture head biometrics in the acoustic domain. HeadSonic can operate implicitly without burdening users.
- By analyzing the signal transmission path, we design a novel biometric representation that captures the uniqueness of the user's head conduction and head geometry characteristics.
- We design BANN, a model tailored for our scenario to enable behavior-irrelevant feature reconstruction with adversarial learning, integrating a feature reconstructor, a behavior discriminator, and a user classifier.
- We conduct extensive experiments with four brands of BCEs and verify HeadSonic's effectiveness under different scenarios. The results demonstrate that it achieves a commendable balanced accuracy of 96.59% while preventing various spoofing attacks.

II. Preliminary

A. Threat Model

The goal of an adversary is to cheat HeadSonic to bypass the authentication. According to the specific professional knowledge and technical capabilities that an adversary could possess, the following attacks are considered:

Zero-effort attack. The adversary simply places the victim's BCE on a desk or holds it by hand before activating the authentication process, hoping the behavior could induce similar impacts on the signal and break the authentication system.

Impersonation attack: The adversary has observed the victim's authentication process, so he/she tries to pass the authentication by wearing the BCE in person, hoping similar head biometrics could be presented.

Replay attack: The commodity BCEs typically have a sound leakage phenomenon, where the speaker sound can be heard at a close distance from the user. We assume that the adversary has obtained the authentication sound by placing a hidden microphone near the victim to eavesdrop on the leaked audio. The audio file is later amplified and replayed to the target BCE with a speaker.

Synthesis attack: The adversary attempts to synthesize the victim's authentication sound by creating a model based on previously eavesdropped signals [18]. The synthesized sounds are then replayed to the target BCE via a speaker. Compared to the replay attack, the synthesis attack aggregates information from multiple samples to generate attack data. We consider the implementation of both traditional signal analysis and powerful deep learning synthesis methods.

Denial-of-service attack: The adversary plays dedicated noises near the victim with a hidden speaker. The noise may overwhelm the authentication signal and keep the victim rejected by the system [19].

B. Acoustic Signal Propagation

This study proposes capturing head biometrics in the acoustic domain to authenticate BCE users. We observe that when users wear a BCE, they typically adjust the clip to ensure a snug fit. The device and human head can be considered as a rigid body [20], with small deformation and a stable distance between any two given points. Within this system, an acoustic channel is established between the BCE speaker and the microphone, through which the propagated sound waves undergo reflection and absorption. Fig. 2 illustrates how the acoustic signal interacts with the human head. In particular, in the frequency domain, the relationship between the speaker signal $S(f)$ and the microphone signal $R(f)$ could be formulated as

$$R(f) = H(f)S(f) + N(f), \tag{1}$$

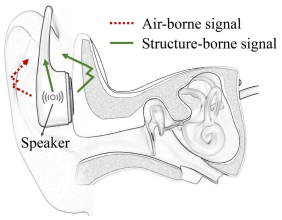


Fig. 2. Sound propagation model of wearing a BCE.

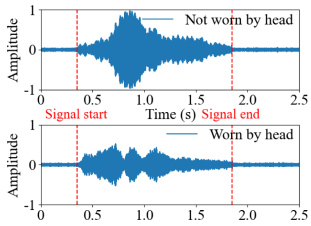


Fig. 3. The impact of human head to the BCE sound.

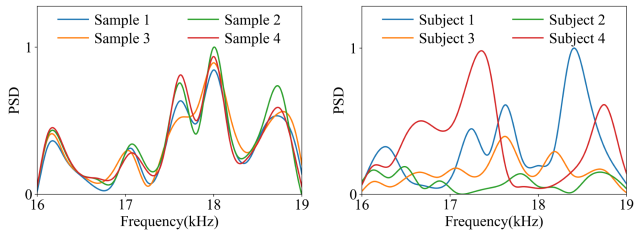
where $H(f)$ represents the channel response of speaker-microphone, and $N(f)$ denotes the combined ambient noises. $H(f)$ could be further decomposed into two parts, i.e., the channel response associated with the rigid body $H_{rig}(f)$, and environment reflection H_{env}. Therefore, the microphone signal could be further expressed as

$$R(f) = H_{rig}(f)S(f) + H_{env}(f)S(f) + N(f). \quad (2)$$

Environmental reflections, particularly the ultrasonic components, suffer significant attenuation as they travel through the air. As a result, $H_{rig}(f)S(f)$, which contains both direct-path and head-conducted components, could dominate the received signal, though exposed to ambient noises [21]. The direct-path signal remains unchanged as the structure of a BCE is fixed. The characteristics of the head-conducted signal are influenced by the head's biometrics. For example, unique head density and bone-tissue ratio can impact signal attenuation, while distinct head geometry leads to varied signal pathways, which all contribute to the cross-user distinctions in head-conducted signals. Given the theoretical analysis of signal propagation, we next conduct a feasibility study to verify the uniqueness of head-conducted sounds.

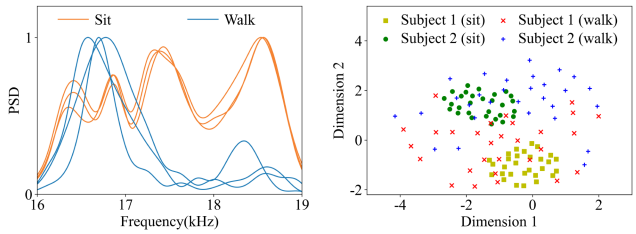
C. Feasibility Study

To investigate the feasibility of utilizing head-conducted sounds for authentication, we conduct several studies. We first explore how the human head affects signals. Specifically, we use the BCE speaker to play a 1.5 s, 16–19 kHz chirp signal. Fig. 3 shows the microphone sound when a BCE is placed on a table and worn by a subject. We observe that when the signal sweeps from 16 kHz to 19 kHz, its amplitudes are degraded or enhanced by the human head with different scales. Overall, the human head suppresses the speaker signals by an average of 2 dB. These observations indicate the frequency-selective nature of the head-conducted response.



(a) (b)

Fig. 4. Normalized PSD of four samples from the same subject (a) and different subjects (b).



(a) (b)

Fig. 5. Normalized PSDs (a) and t-SNE visualization (b) for illustrating intra-class variability under sitting and walking.

Next, we investigate if head-conducted sounds can be used to distinguish users. Four subjects participate in this study, with data gathered as they sit on a chair wearing the BCE. Fig. 4(a) and (b) display the normalized Power Spectrum Densities (PSDs) of received chirp signals from the same and different subjects, respectively. It is clear that the PSDs are consistent for the same subject but vary significantly throughout different subjects. The results demonstrate the feasibility of using head-conducted sounds for user authentication. Moreover, the PSDs exhibit user discrimination at different frequencies, encouraging us to use a wide range of frequency components for describing user head biometrics with fine granularity.

D. Towards Behavior-Irrelevant

Different user behaviors could affect the authentication. For example, when a person is in motion, the relative position between the BCE and the head drifts due to inertia. This leads to changes in the speaker-microphone channel and further introduces variations in the measured biometrics.

We conduct an experiment to test the influence of user behavior. Specifically, we ask two subjects to wear a BCE in static (sitting on the chair) and dynamic (walking) two scenarios. The BCE speaker plays a 1.5 s, 16–19 kHz chirp signal, while the microphone collects corresponding responses. Fig. 5(a) presents the PSD profiles of two behaviors from one subject, where the signals exhibit different patterns. We further use t-distributed Stochastic Neighbor Embedding (t-SNE) projection [22] to visualize the sample PSDs in two-dimension space, as depicted in Fig. 5(b). We observe that compared to sitting, walking introduces more intra-class variability, resulting in blurred classification boundaries among samples from different subjects.

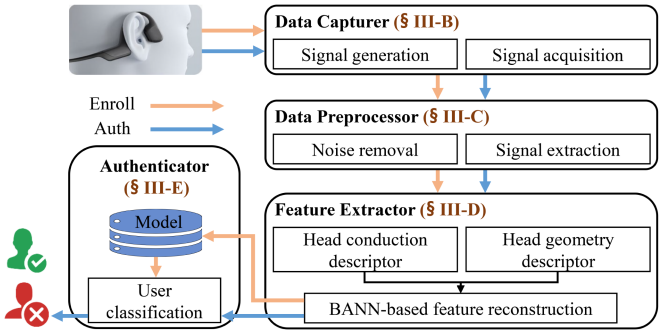


Fig. 6. Workflow of HeadSonic.

Therefore, to build a usable BCE authentication scheme, we should minimize the reliance on behavior biometrics, and realize a behavior-irrelevant authentication system.

III. System Design

In this section, we introduce the system architecture and design details for HeadSonic.

A. System Overview

HeadSonic operates in enrollment and authentication two phases. During enrollment, the system creates a user profile based on the data collected from the BCE microphone. During authentication, the user profile serves as a template for matching incoming microphone data.

As shown in Fig. 6, HeadSonic consists of four modules: data capturer, data preprocessor, feature extractor, and authenticator. The data capturer generates a probe signal, transmitting it via the BCE speaker and capturing the response with the BCE microphone. The data preprocessor removes noise with a high-pass filter and performs signal extraction with cross-correlations to derive the target signal shaped by the human head. The feature extractor derives novel biometric features from the signal by combining the uniqueness of head conduction and head geometry. Subsequently, a BANN model is developed to reconstruct these features into behavior-irrelevant representations. Finally, these reconstructed representations are fed into the authenticator for user differentiation.

B. Data Capturer

Leveraging speakers to emit probe signals and microphones to capture acoustic responses is a common practice in many smart devices [21], [23], [24]. We consider the following key design criteria for the probe signal: i) could capture biometric information for effective user differentiation; ii) complex enough for adding synthesis attack difficulty [18]; and iii) have a millisecond-level period for short waiting time. These requirements necessitate careful design for the signal.

To meet the first requirement, we utilize the chirp signals (upward sweep) [19], [21] to capture the frequency selectivity contributed by the physiological characteristics of the human head. We meet the second requirement by adopting the frequency-division multiplexing approach [25], i.e., embedding multiple chirps within a time slot to enhance its frequency complexity. Specifically, the probe signal $s(t)$ could be expressed as

$$s(t) = \sum_{i=1}^{m} A_i \sin\left(2\pi f_i(t)t + \phi_i\right), \tag{3}$$

where A_i, $f_i(t)$, and ϕ_i are the amplitude, frequency, and phase of the i_{th} chirp signal. m is the number of chirps employed.

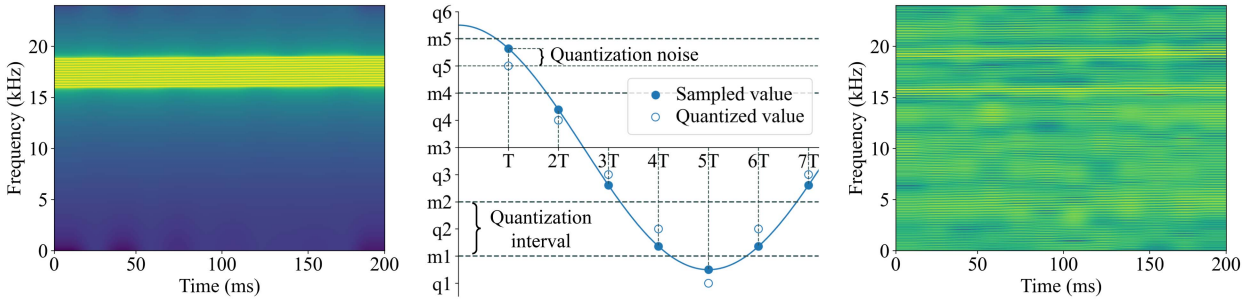
In our case, each chirp signal spans T milliseconds in duration and B Hz in bandwidth within the range $f_{\min}$ to $f_{\max}$. The signal $s(t)$ is depicted in Fig. 7(a). Analog signals must undergo sampling and quantization, further converted into digital signals for speaker playback. Fig. 7(b) illustrates the analog signal sampling and quantization process, where $q_1, \ldots, q_6$ represent the quantization levels, and $m_1, \ldots, m_6$ denote the endpoints of quantization interval. Sampling discretizes continuous signals, while quantization maps sampled values within quantization intervals to nearby quantization levels. The difference between the sampled and quantized values is known as quantization noise [26]. We observe that after $s(t)$ digitization, quantization noise introduces a variety of new frequencies across the entire spectrogram, as shown in Fig 7(c), further enhancing its complexity. The digitized signal $\widetilde{s}(t)$ is played through the BCE speaker and simultaneously recorded with the BCE microphone. The playing takes T milliseconds, while the recording takes $1.5T$ milliseconds to compensate for the sensor's synchronization deviation. The microphone sampling rate is denoted as f_s. We get a signal with $1.5 \times (T/1000) \times f_s$ points as data capturer output, as shown in Fig. 8(a). The microphone signal $r(t)$ is passed through the data preprocessor.

C. Data Preprocessor

Data preprocessing involves noise removal and signal extraction. First, we use a high-pass filter with a cutoff frequency of f_c to remove common noise, such as human voice and urban environmental noise (generated predominantly by traffic) [27]. Note that we do not utilize a bandpass filter with passband from $f_{\min}$ to $f_{\max}$ as useful sensing information also exists outside this band. Next, we isolate the desired signal in the microphone data. Specifically, we employ the speaker signal $\widetilde{s}(t)$ as a reference signal and iteratively shift the microphone signal $r(t)$, calculating its cross-correlation [28] with the reference signal, as shown in Fig. 8(b). The delay corresponding to the maximum cross-correlation between two signals can be calculated as:

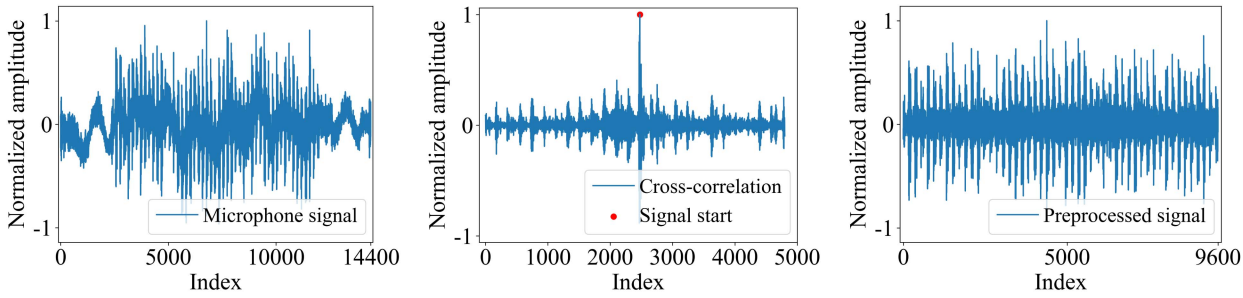
$$delay = \arg\max_{n} \sum_{t=0}^{N-n-1} r(t+n)\widetilde{s}(t), \tag{4}$$

where N is the length of $r(t)$ and n is the number of shifted samples. This delay is utilized to align the two signals, determining the starting point of the desired signal $\widetilde{r}(t)$. The signal after data preprocessing is shown in Fig. 8(c).



(a) Analog signal (b) Sampling and quantization process (c) Digital signal

Fig. 7. Illustration of signal digitization. Analog signal (a), analog signal sampling and quantization process (b), and corresponding digital signal (played by speaker) after quantization (c).



(a) Microphone signal (b) Signal extraction process (c) Preprocessed signal

Fig. 8. The signal received by a BCE microphone (a), the cross-correlation between the microphone signal and the reference signal (b), and the signal after data preprocessing (c).

D. Feature Extractor

Our goal is to develop a set of person-distinguishable representations from the preprocessed signal. In this section, we present approaches for feature extraction to characterize the conduction and geometry properties of a user's head.

1) Level-1: Head Conduction Descriptor: When a BCE generates an acoustic wave with its built-in speaker, part of the wave conducts through the head, and is ultimately captured by the microphone. The signal's conduction is affected by the head's inherent biometrics, such as density and bone-tissue ratio, which vary from person to person due to the unique composition (e.g., water, lipids, and fat-free solids) of each individual's head [29]. As a result, the received signal exhibits distinct attenuation across different frequencies [30]. We thus use the signal's frequency response to characterize these unique head biometrics, termed head conduction descriptor.

We use Linear Prediction Cepstrum Coefficients (LPCC), Spectral Centroid (SC), and Spectral Spread (SS) to describe the frequency response of the signal. LPCC is widely used for acoustic modeling in audio-related applications [31], [32]. Compared to MFCC which focuses on low-frequency details, LPCC excels at capturing subtle variations across different frequencies by deriving the cepstral coefficients. The spectral centroid represents the spectrum's center of mass, while the spectral spread measures its dispersion. We extract k-dimension LPCC features to describe the signal's spectral characteristics. For SC and SS, we segment the data into k frames and calculate them within each frame. In all, the head conduction descriptor can be represented as $\{F_{LPCC}(1), \ldots, F_{LPCC}(k), F_{SC}(1), \ldots, F_{SC}(k), F_{SS}(1), \ldots, F_{SS}(k)\}$.

2) Level-2: Head Geometry Descriptor: The acoustic wave traveling through the head does not follow a single path; instead, before reaching the microphone, it undergoes multipath reflection due to the complex geometry of each head—a property that has demonstrated uniqueness across individuals [21]. We thus use the signal's multipath characteristics to characterize the unique head geometry traits, termed head geometry descriptor.

We use Transfer Function (TF), Cross Correlation (CC), and Auto Correlation (AC) to describe the signal's multipath characteristics. Transfer function measures the impact of multipath propagation on the signal and is commonly used in channel modeling [15], [16]. Cross correlation reveals time delays between transmitted and received signals caused by multipath arrivals, while auto correlation measures repetitive patterns within the received signal. Using Welch's method [33], we estimate TF over k frequency bands. For CC and AC, we divide both the transmitted and received signals into R fragments and calculate correlations respectively. For each fragment, we first identify the alignment coefficient as "center", then take $(k-1)/2$ coefficients before and after it. We average these coefficients across fragments to describe the overall multipath pattern. This yields head geometry descriptor as $\{F_{TF}(1), \ldots, F_{TF}(k), F_{CC}(1), \ldots, F_{CC}(k), F_{AC}(1), \ldots, F_{AC}(k)\}$.

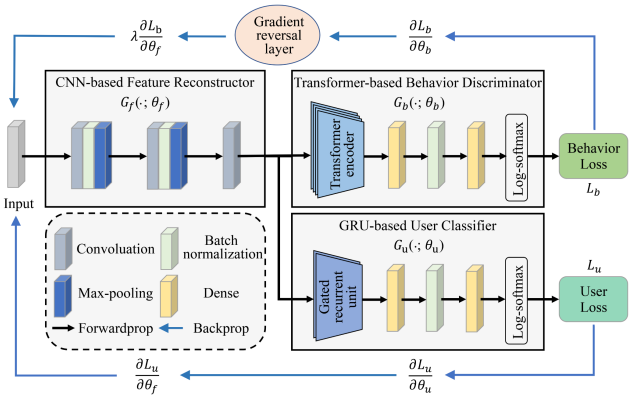


Fig. 9. Illustrating of BANN architecture.

3) BANN-Based Feature Reconstruction: In Section II-D, we explore the influence of user behavior, revealing that behavioral traits can increase intra-class variability and potentially diminish authentication performance. To tackle this, we design a BANN model that reconstructs the extracted features into behavior-irrelevant representations. The BANN model, as depicted in Fig. 9, comprises a feature reconstructor, a behavior discriminator, and a user classifier. Although each sub-network serves a distinct function, the BANN relies on their effective collaboration to achieve behavior adaptation. Specifically, the feature reconstructor takes hybrid features (a combination of two descriptors) as input, aiming to reconstruct them into behavior-irrelevant representations. Meanwhile, the behavior discriminator takes the reconstructed representations as input, outputting N-class probabilities that correspond to different behavioral patterns. The two sub-networks co-evolve during training until the behavior discriminator can not distinguish different behaviors. To enable this, we insert a Gradient Reversal Layer (GRL) [34] between the feature reconstructor and behavior discriminator, which multiplies the gradient by a negative coefficient λ during backpropagation. This unique layer drives the feature reconstructor to eliminate behavioral traits during optimization. Nevertheless, the adversarial training process may also compromise intrinsic physiological traits. Therefore, we further introduce a user classifier (outputting M-class probabilities) to ensure the reconstructed representation retains its ability to differentiate users. Next, we detail the design of each sub-network.

Feature reconstructor: The feature reconstructor transforms extracted features into behavior-irrelevant representations. A CNN forms its foundation due to its proficiency in abstracting features [24], [35]. It begins with a convolutional layer with 3×1 kernels to learn large-scale features, followed by a batch normalization layer and a max pooling layer with kernel size 3×1. Subsequently, another two convolutional layers with 3×1 kernels are utilized to learn small-scale features, with a batch normalization layer and a max pooling layer with kernel size 2×1 applied after the first layer. Following the last convolutional layer, a dropout layer with a probability of 0.5 is applied to prevent overfitting. ReLU is used as the active function after each max pooling layer. The output, with dimension 13×10, serves as input for the subsequent behavior discriminator and user classifier.

Behavior discriminator: Since the features are extracted from the signals induced by body behaviors, specific behavior characteristics are inevitably embedded in these features. We develop a behavior discriminator using a Transformer encoder [36] with multi heads, which jointly attend to behavioral information from different representation sub-spaces. Specifically, the behavior suppressor consists of a transformer encoder and two dense layers. The transformer encoder has 5 heads and a hidden dimension of 512. Following the first dense layer, batch normalization and ReLU activation functions are applied. A log-softmax activation function is used after the last dense layer, converting the output into logarithmic probabilities corresponding to different behavior categories.

User classifier: The reconstructed features have different importance to distinguish users. Therefore, we employ the Gated Recurrent Unit (GRU) [37] to construct a user classifier, utilizing its gating mechanisms to select and preserve useful information for individual differentiation. Specifically, the user classifier consists of two stacked GRUs, a batch normalization layer, and two dense layers. Based on the two GRUs, the information of the user head biometric is compressed at the last time step. The subsequent dimensionality reduction process is similar to the behavior suppressor, as shown in Fig. 9. Finally, a log-softmax activation function maps the output to logarithmic probabilities corresponding to M user classes.

BANN training: We train the BANN with data collected from 30 subjects (i.e., regarded as default users), each contributing 100 samples across four behaviors: sitting on a chair, rotating the head, rotating the body, and walking. We assume this default user data, solely used for model generation, has been preloaded into the device upon shipment. During training, the feature reconstructor inputs hybrid features, while the behavior discriminator and user classifier output behavior (4-class) and user (30-class) probabilities, respectively. The multi-user classification ensures a generalizable extraction capability in the feature extractor. We conduct a grid search to find the best hyper-parameter combination and use Adam optimizer for parameters optimization [38]. Negative Log Likelihood (NLL) losses from the behavior discriminator and user classifier are added for backpropagation. The BANN model needs to be trained only once. After that, the behavior discriminator and user classifier are discarded, and the feature reconstructor is applied to new users directly.

Feature visualization: We compare the extracted features, including head conduction descriptor, head geometry descriptor, hybrid features, and reconstructed features. As shown in Fig. 10, t-SNE projection [22] is utilized to visualize the feature space in two dimensions. The data are gathered in the walking scenario, with each subject contributing 30 samples. Overall, both the head conduction descriptor and head geometry descriptor demonstrate user differentiation, where different clusters exhibit distinct centroids. Combining them together clarifies the cluster

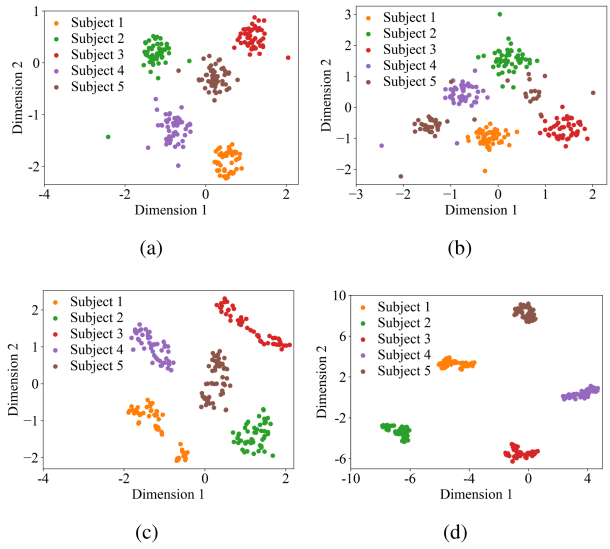


Fig. 10. t-SNE visualization of head conduction descriptor (a), head geometry descriptor (b), hybrid features (c), and reconstructed features (d).

boundaries. After feature reconstructing, with behavior inconsistency mitigated, the samples exhibit smaller intra-class distances and larger inter-class distances. This validates the effectiveness of the extracted features for analyzing head-conducted signals.

E. Authenticator

In a real-world authentication setting, the attackers' samples are unavailable during training. Therefore, we profile the legitimate user with four methods, including i) Local Outlier Factor (LOF), ii) Isolation Forest (IF), iii) One-Class Support Vector Machine (OC-SVM), and iv) Binary Support Vector Machine (Bi-SVM). LOF [39] is a density-based method that identifies outliers by comparing the local density deviation between the test sample and its n neighbors. A sample with a substantially lower density than its neighbors will be regarded as an outlier. IF [40] is an anomaly detection method based on ensemble learning, which identifies outliers by recursively splitting data to construct m $iTrees$. A sample with a shorter path length in the trees is seen as an outlier. OC-SVM [41] is a distance-based classifier that works by mapping samples into high-dimension feature space with the kernel function. A hyper-sphere is optimized for including the training samples with minimal volume. A sample outside the hyper-sphere is regarded as the outlier. Bi-SVM [42] constructs a hard margin to separate two classes by maximizing the distance between support vectors and the decision boundary. Different from one-class classification, it requires two-class data for training. To fit the authentication scenario, we train Bi-SVM with data from legitimate and default users, aiming to create a generalizable boundary that effectively distinguishes the legitimate user.

IV. Implementation

A. Experimental Setup

1) Design Details: For data capturer, each chirp signal has a duration of $T = 200$ ms and a bandwidth of $B = 200$ Hz. Specifically, 12 chirps uniformly distributed within $f_{\min} = 16$ kHz to $f_{\max} = 19$ kHz are employed, with frequencies ranging as follows: 16–16.2 kHz, 16.25–16.45 kHz, . . . , 18.75–18.95 kHz. The sampling frequency f_s is 48 kHz. For data preprocessor, the cut-off frequency of the high-pass filter is $f_c = 4$ kHz. The reason is that the human voices range 0.3–3.4 kHz [43], and urban environmental noises (e.g., background music, loud radios, vehicle alarms, and roadway traffic) span 1–4 kHz [27]. For feature extractor, we set $k = 13$ and $R = 50$. The hybrid feature dimension of the two descriptors is $6 \times 13 = 78$, which is fed into the feature reconstruction model, producing reconstructed presentations with the dimensionality of 130. The negative coefficient λ gradually decreases from 0 to -1 with the training epoch increase. For authenticator, we consider $n = 5$ neighbors in LOF and $m = 100$ $iTrees$ in IF. For OC-SVM, we use the radial basis function as the kernel function, and the optimal parameters γ and ν are 0.7 and 0.01. For Bi-SVM, we employ the radial basis function as well and use $C = 100$ and $\gamma = 0.01$.

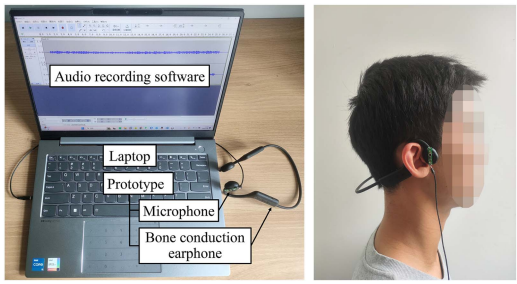
2) Dealing With Long-Term Biometric Changes: Like other biometrics, the head biometric may exhibit variations over time, resulting in the classification model poorly identifying newly acquired samples. Many strategies have been proposed to tackle this problem [44], [45], [46], with the core idea being to consistently update the classification model with new samples. We adopt this approach, and the key steps are as follows: 1) The system maintains a training dataset after initial user enrollment. 2) When a new authentication sample arrives, it is marked as legitimate if verified successfully. 3) The system updates the dataset with this legitimate sample in a first-in, first-out manner. 4) The classification model is retrained on the updated dataset every few days or adjusted based on the user's authentication frequency.

3) HeadSonic Prototype: We experiment with four different BCEs, including Enkor EB103, Langsdom BS17, Yuans X7, and Shokz OpenRun. The Enkor EB103 is used in all scenarios. The acoustic signal is played by the BCE speaker, and 70% volume is used to reduce power consumption and disturbance. We build our data collection circuit based on a MEMS microphone chip, as shown in Fig. 11. The microphone chip is connected to a Laptop via a 3.5 mm audio interface for data transmission. All data captured at 48 kHz is anonymized for offline processing.

B. Data Collection

After receiving the IRB approval from our institute, we start data collection in September 2023. We employ 60 subjects (46 males and 14 females) aged from 19 to 30. These subjects are recruited from undergraduate and graduate students in our institute, with each given $5.5 as an incentive for participating.

Before data collection, subjects need to re-wear the device for 5 minutes to get familiar with it. These subjects are divided into



(a) (b)

Fig. 11. System prototype (a) and illustration of a subject wearing the device (b).

two groups, i.e., *group-A* and *group-B*. Data from *group-A* (30 subjects) is used for BANN and Bi-SVM training. Data from *group-B* (30 subjects) is used for system evaluation. We set up the default experimental environment in an office room with $\sim$ 40 dB of ambient noise. Each subject is required to collect one session data (100 samples) for each of the four different behaviors, i.e., sitting on a chair, rotating the head, rotating the body, and walking. This generates *dataset-1* and *dataset-2* with $30 \times 4 \times 100 = 12000$ samples respectively.

We also collect data under different scenarios, such as speaker volumes, wear positions, noise levels, etc. For each variable scenario (e.g., 50% speaker volume), data from five subjects are compiled in one session. We select legitimate samples from *dataset-2* to train the authentication model and conduct testing across different scenarios.

C. Evaluation Metrics

The following metrics are utilized to evaluate HeadSonic. False Acceptance Rate (FAR) is the ratio of non-users gaining access. False Rejection Rate (FRR) is the ratio of legitimate users being denied access. Balanced Accuracy (BAC) is used to evaluate the model's overall performance under unbalanced data [47]. It is defined as the average of the True Acceptance Rate (TAR), which represents the ratio of legitimate users gaining access, and the True Rejection Rate (TRR), which represents the ratio of non-users being denied access. Frequency Count of Scores (FCS) [48] is the frequency count of samples' prediction scores (In Bi-SVM, the prediction score is defined as the difference between the legitimate and illegitimate probabilities).

V. Performance Evaluation

A. Overall Performance

1) Performance of BANN: To examine the security gain brought by the BANN model, we compare the authentication performance of using hybrid features (without BANN) and reconstructed features (with BANN). The Bi-SVM is employed in this evaluation. We treat each subject as the legitimate user, using 80% of their samples for training, and the remaining 20%, along with samples from other subjects, for testing. Fig. 12 illustrates the BAC comparison among 30 subjects. Overall, when using BANN alone, our system achieves an average BAC of 86.80%. Adding the BANN model significantly improves the authentication performance, resulting in an average BAC of 96.59%. We also observe a performance gain for each subject with the application of BANN, with subject 26 achieving the largest BAC increase of 31.81%. The results demonstrate BANN's efficacy in enhancing authentication reliability through behavior-irrelevant feature reconstruction.

2) Performance of Different Classifiers: We compare the performance of different classifiers, each trained with 80% of the legitimate user samples. Fig. 13 presents the results. Overall, the average BACs for LOF, IF, OC-SVM, and Bi-SVM across four behaviors are 93.94%, 90.70%, 92.99%, and 96.63%, respectively. Notably, Bi-SVM outperforms the other three classifiers across different behaviors, which may due to the involvement of default users strengthen the classification boundary. To ensure comparability across different experiments, we consistently use Bi-SVM and train the classifier with 80% of the legitimate user data in all subsequent experiments, unless specified otherwise.

B. Impact of Various Factors

1) Different Devices: We evaluate our system on four BCEs: Enkor EB103 ($19.25), Langsdom BS17 ($16.43), Yuans X7 ($31.57), and Shokz OpenRun ($94.54). They have different speaker configurations, surface materials, body shapes, and prices. Fig. 14 illustrates the BAC achieved by the four devices. We find our system performs well across these devices. Overall, our system achieves an average BAC of 96.81%, 96.34%, 97.42%, and 97.15% on Enkor EB103, Langsdom BS17, Yuans X7, and Shokz OpenRun, respectively. Moreover, for each of the four behaviors, the four devices achieve high BACs. The results demonstrate that our system can scale among different BCE devices.

2) Human Speech: To evaluate the impact of human speech, we ask subjects to read different textual materials, including individual words, sentences, and paragraphs, which are excerpted from the novel "The Old Man and The Sea". During reading, subjects are required to keep an interval of 2 s between adjacent elements to create diverse reading rhythms. Moreover, we consider speech from males and females separately, as their voices have different fundamental frequencies (90–155 Hz for males and 165–255Hz for females) [43]. As shown in Fig. 15, the BACs for males are 96.33%, 95.16%, 94.78%, and 94.15% during periods of silence, word reading, sentence reading, and paragraph reading, respectively. For females, the corresponding BACs are 96.72%, 94.36%, 94.97%, and 94.61%. Compared to the silent scenario, speaking results in an average BAC decrease of 1.63% for males and 2.07% for females. Additionally, a t-test analysis [49] indicates that gender had no significant impact on BAC, with a p-value of 0.74, exceeding the threshold of 0.05.

3) Training Data Size: To investigate the impact of training data size, we vary the number of legitimate samples used to train the classification model from 20 to 320. The result is shown in Fig. 16. We can observe that as the size of training data increases from 20 to 320, the BAC first increases and then remains stable. In particular, when the training data size is 120

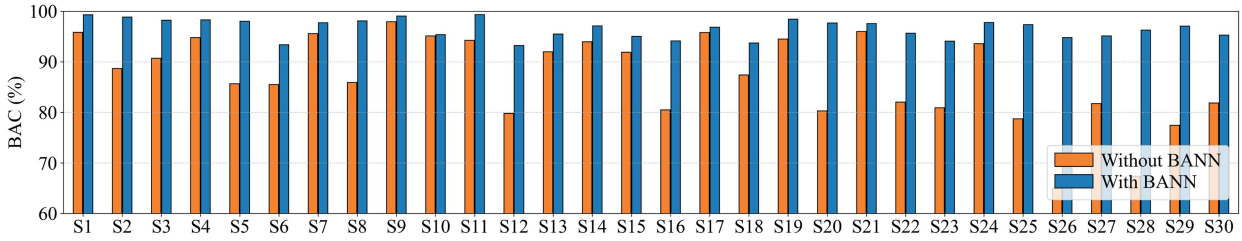


Fig. 12. Performance comparison of without/with BANN.

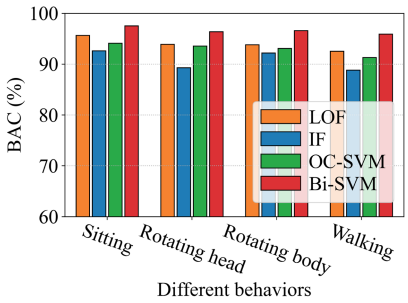


Fig. 13. BAC of different classifiers.

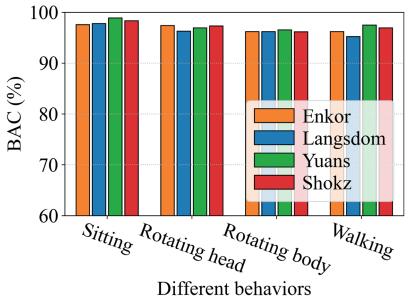


Fig. 14. Impact of different devices.

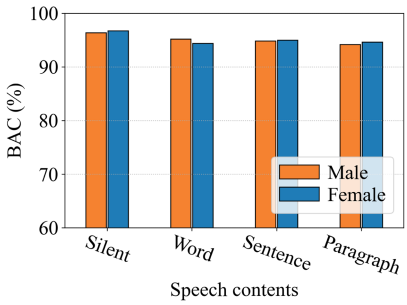


Fig. 15. Impact of speech contents.

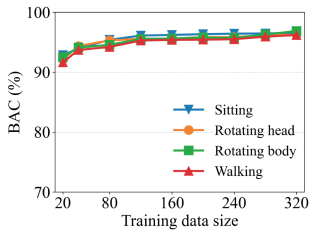


Fig. 16. Impact of training data size.

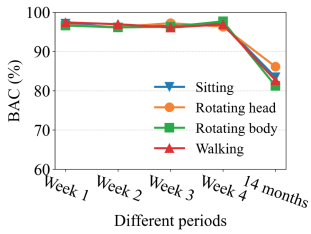


Fig. 17. Performance of different weeks.

Fig. 18. Illustration of BCE wearing positions.

samples, HeadSonic can achieve a BAC of 96.12%, 95.59%, 95.57%, and 95.29% under sitting, rotating head, rotating body, and walking, respectively.

4) Consistency Over Time: To evaluate the consistency of HeadSonic over different periods, we consider short-term and long-term two conditions. Five subjects participate in this evaluation. We use their samples from *dataset-2* as the first week data, with the same quantity gathered in each subsequent period. In short-term condition, we conduct a four-week study, i.e., train our model with data from week 1 and test it with data from weeks 1, 2, 3, and 4. In the long-term condition, we test the model trained in week 1 using data collected 14 months later. Fig. 17 illustrates the BACs of four behaviors at different periods. We do not observe an obvious BAC decrease within the four weeks, and the average BACs are 96.98%, 96.38%, 96.57%, and 97.01% in weeks 1, 2, 3, and 4, respectively. After 14 months, the average BAC decrease to 83.39%. Using the method in Section IV-A2, we retrain each user-specific model with 80% samples collected after 14 months. As a result, the average BACs increase to 97.12%, demonstrating the efficacy of our approach in tackling long-term biometric change.

5) Wearing Positions: Apart from movement, simple putting on and taking off BCEs in daily usage also leads to wear position change. We consider three different wear positions, as shown in Fig. 18. We collect one session data (100 samples) for each scenario. The first is the normal multi-wear scenario, where

TABLE I
IMPACT OF DIFFERENT WEARING POSITIONS

Wearing Position	Mean BAC (%)	Std BAC (%)
Multi-wear	96.17	1.84
15° clockwise	95.49	2.39
30° clockwise	95.23	3.54

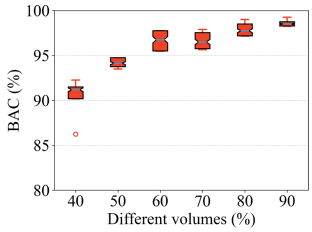


Fig. 19. Impact of different volumes.

subjects re-wear the BCE every 20 samples to introduce wear position variation. In the other two scenarios, the BCE is rotated 15° and 30° clockwise intentionally from its normal position. We train user-specific classifiers on *dataset-2*, which is gathered in a fixed wear position, and test them in three scenarios. Table I shows that HeadSonic achieves BACs of 96.17%, 95.49%, and 95.23% in three scenarios. In the multi-wear scenario, a high BAC is achieved as human ear contour constraining BCE position variation. In the other two scenarios, we observe a slight BAC decrease. However, most subjects report feeling strange with these rotated positions as the BCE does not fit snugly around the ear contours, suggesting their infrequent occurrence in daily usage.

6) Unseen Behaviors: Our authentication model is trained on four pre-defined behaviors. We now test its ability to distinguish users under additional, unseen behaviors, including standing up and sitting down, jumping, and running. Note that "standing up and sitting down" is a behavior that contains two actions. Five subjects participate in the evaluation, with data compiled in two sessions for each behavior. Overall, the BACs are 94.69%, 92.78%, and 92.47% for standing up and sitting down, jumping, and running, respectively. We find that the system retains a certain ability to differentiate users across unseen behaviors. There are two possible reasons. First, the ear contour constrains BCE drift during movement, preserving physiological traits across activities. Second, the feature reconstructor helps eliminate earphone drift inconsistencies, retaining key physiological traits in the reconstructed features.

7) Speaker Volumes: HeadSonic utilizes the BCE speaker to play the probe signals. The speaker volume affects the received sounds' signal-to-noise ratio, thus impacting authentication performance. We assess the system performance at different speaker volumes, as shown in Fig. 19. We observe that as the volume increases from 40% to 90%, the BAC first increases and then stabilizes. Specifically, the BACs are 90.27%, 94.19%, 96.65%, 96.66%, 97.91%, and 98.61% when the speaker volume is set to 40%, 50%, 60%, 70%, 80%, and 90%, respectively. The results confirm our system can work well across a wide range of volumes.

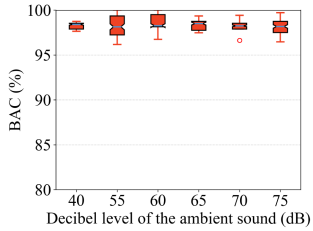


Fig. 20. Impact of ambient noises.

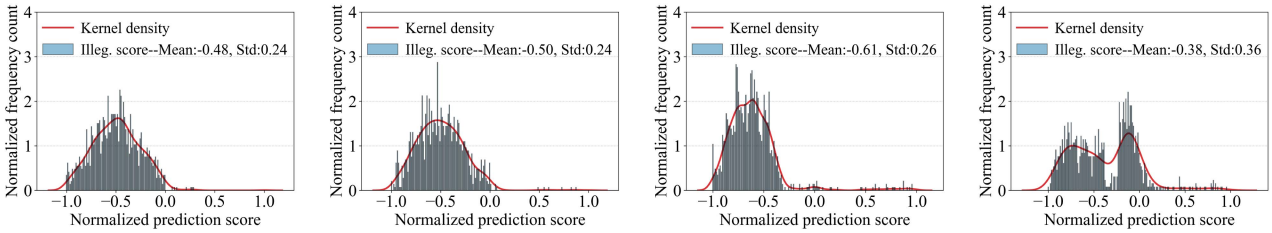
8) Ambient Noises: The above results are obtained in the regular office scenario with a 40 dB noise level, we next examine the impact of different types of ambient noises. We consider daily ambient noises from various settings: a café at 55 dB, a working air conditioner at 60 dB, casual conversations at 65 dB, a shopping mall at 70 dB, and a train station at 75 dB. Subjects are registered in an office environment and tested in these noisy scenarios. Fig. 20 displays the system performance under these scenarios. In particular, the BACs are 96.85%, 97.18%, 96.65%, 96.57%, 96.73%, and 96.52% under noise levels 40 dB, 55 dB, 60 dB, 65 dB, 70 dB, and 75 dB, respectively. The results show that our system can work well under regular ambient noises. The reason is that the ambient noise typically occupies the low-frequency range, which has limited capabilities to corrupt signals in the high frequencies.

C. Evaluation of Attack Resistance

1) Zero-Effort Attack: We evaluate our system under zero-effort attackers, who try to gain access to the system without presenting head biometrics. We simulate this attack under four BCE placement scenarios: i) hold by hand, ii) on a solid wooden desk, iii) on a medium-density fiberboard, and iv) on a soft rubber pad (e.g., mouse pad). Subjects in *dataset-2* are treated as victims respectively, and we collect one session data under each scenario for attacking. Overall, the FARs for scenarios i) to iv) are 0.79%, 1.14%, 1.35% and 1.67%, respectively. The result reflects the effectiveness of our system in rejecting authentication requests when the BCE is not worn by a human head.

2) Impersonation Attack: We consider two types of impersonation attacks. For *random impersonation*, each subject acts as the victim, while the data from other subjects is used for attacking. *dataset-2* is utilized for evaluating this attack. For *knowledgeable impersonation*, five subjects act as skilled attackers, each observing authentication videos of five victims to imitate their wearing behaviors and collecting data for one session each. Overall, the FARs are 2.17% and 2.40% for random and knowledgeable impersonations, respectively. The results confirm the system's security under impersonation attacks, as human head biometrics (e.g., density, geometry, and bone-tissue ratio) are hard for an adversary to imitate.

3) Replay Attack: We simulate this attack by placing the Redmi K60 near the target BCE for recording during the authentication process. Four different devices are utilized for replaying the recorded audio, including i) Redmi K60 (smartphone), ii) Lenovo Legion R7000 (laptop), iii) Xiaomi Sound Pro (smart speaker), and iv) Amazon Echo Dot (smart speaker). To generate



(a) Synthesizing our signal with transfer function. (b) Synthesizing our signal with least square. (c) Synthesizing our signal with GAN. (d) Synthesizing chirp signal with GAN.

Fig. 21. Normalized FCS and KD of synthesis samples under different modeling methods and probe signals.

sufficient signal power, the replay devices are positioned close to the BCE microphone and operate at the maximum volume. Five subjects act as victims, each having their data recorded in one session. Overall, the FARs are 0.60%, 1.20%, 2.20%, and 2.80% for replay devices i) to iv), respectively. The results demonstrate the effectiveness of our system in preventing replay attacks. The reason is that the eavesdropped audio is the air-borne sound, while the user's head biometrics are primarily embedded in the head-conducted sound (see Section II-B), which is difficult for in-air microphones to capture.

4) Synthesis Attack: We employ three modeling methods, including two signal analysis techniques, i.e., transfer function-based and least square-based methods [18], and a deep learning approach, i.e., Generative Adversarial Network (GAN) [50]. The samples are synthesized with the above side-channel eavesdropping data and injected into the BCE microphone via the Amazon Echo Dot, which demonstrates the highest FARs in replay attacks. Compared to the replay attack, the synthesis attack aggregates information from multiple samples to generate attack data. We report the FAR, FCS and kernel density (KD) of normalized prediction scores under the Gaussian kernel [24] for the synthesis samples.

Transfer function-based synthesis: We use the transfer function to model the system's nonlinear impact on acoustic signals. The transfer function is calculated with speaker signal $\widetilde{s}(t)$ and eavesdropped response using Welch's method [33]. Each subject is considered as the victim in turn, and the transfer function is computed by averaging 5 responses. Therefore, we generate multiple transfer functions for each subject based on different measurements, and then produce synthetic samples using each of them. Fig. 21(a) illustrates the FCS and KD of prediction scores for synthesis samples. We observe that most samples are correctly classified as illegal (with predicted scores below zero), yielding a total FAR of 0.73%.

Least square-based synthesis: We use the least square to directly estimate the signal mapping. Specifically, the method is formulated as $Ax = R$, where A and R are the speaker signal $\widetilde{s}(t)$ and eavesdropped response, and x is the least square solution. Similar to transfer function-based synthesis, each model is calculated by averaging 5 responses. We obtain multiple least square solutions based on different measurements for each subject and generate synthetic samples with each solution. From Fig. 21(b), the majority of attack samples yield prediction scores below 0, leading to an overall FAR of 1.39% for this attack.

GAN-based synthesis: GAN is a machine learning method that engages a game between a generator *G* and a discriminator *D*. *G* aims to generate new data (i.e., synthesis samples) from a noise vector z, while *D* aims to discriminate the generated data and the ground truth (i.e., victim's legal samples). *G* and *D* evolves together during training, until *D* cannot distinguish the data generated by *G* as fake. As a powerful model, GAN is widely used for various data generation tasks [51], [52]. In our scenario, we utilize all eavesdropped responses from each victim to train a GAN and generate a set of attacking samples with the trained model. From Fig. 21(c), most attack samples also have prediction scores below 0, resulting in a total FAR of 2.47% for this attack.

Synthesis attack resistance analysis: To investigate the effect of employing a complex probe signal for defending against synthesis attacks, we compare it with the chirp signal, which is widely used in various authentication tasks [18], [19], [21]. Specifically, a 16-19 kHz, 200 ms chirp is utilized for comparison. Five subjects are recruited, with each required to collect data for one session. We synthesize attack samples using the GAN-based synthesis method with eavesdropped data, which has demonstrated the highest FAR among the three modeling methods. Fig. 21(d) illustrates that employing a simple chirp leads to a higher number of samples with prediction scores exceeding 0 (i.e., misclassified as legitimate), resulting in a total FAR of 8.83%. The results confirm the advantages of our complex probe signal in preventing synthesis attacks.

5) Denial-of-Service Attack: We examine the Denial-of-Service (DoS) attack, which generates noises purposely to block authentication sounds. Two kinds of sounds are utilized as interference: the 17–19 kHz and 4–24 kHz white noises. Note that we do not consider noises below 4 kHz, as they are filtered out during data preprocessing. We choose five sound pressure levels from 20 dB to 60 dB, and the performance is shown in Fig. 22. We observe that our system achieves high TAR with up to 50 dB interference, although the performance decreases slightly from 20 dB to 50 dB. We also find that the system performs better under 16–19 kHz noises, as our signal contains biometric information beyond these frequencies. When the noises are 50 dB at 16–19 kHz and 4–24 kHz, our system achieves TARs of 95.21% and 94.53% respectively. We thus set 50 dB as the threshold for environmental noise detection. If the noise level within 4-24 kHz is higher than it, the user is recommended to use alternative authentication methods.

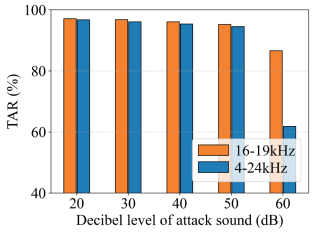


Fig. 22. DoS attack performance.

TABLE II
CONTENT OF QUESTIONNAIRE

No.	Questions
Q1	HeadSonic is easy to use.
Q2	There is no discomfort using HeadSonic.
Q3	HeadSonic is easy to learn.
Q4	HeadSonic does not introduce much cognitive load .
Q5	The login time is short.
Q6	HeadSonic can be used daily.
Q7	I am willing to use HeadSonic on my (future) bone conduction earphones.

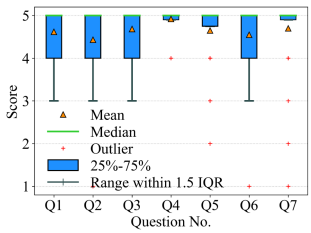


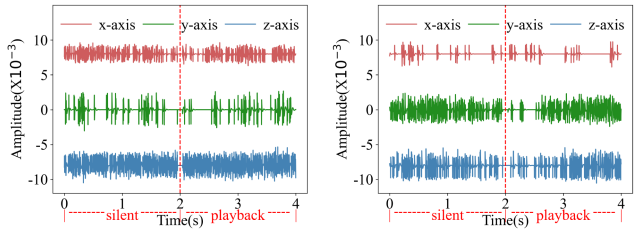
Fig. 23. Questionnaire score.

D. Evaluation of Computational Delay

We investigate the HeadSonic's computational delay, starting from receiving the acoustic data to generating the authentication result. All data is transmitted to a PC equipped with a 3.0 GHz Intel i7-9700 CPU and 16 GB memory. We conduct experiments with a batch size of 8 to measure the average computational time for noise reduction, signal segmentation, level-I and II feature extraction, BANN-based feature reconstruction, and user authentication. The level-I and II feature extraction takes the longest processing time, averaging 46.42 ms. For noise reduction, signal segmentation, feature reconstruction, and user authentication, the average computational time is 0.3 ms, 0.47 ms, 36.37 ms, and 15.17 ms, respectively. Overall, HeadSonic processes a sample and completes user authentication in approximately 98.73 ms, demonstrating the computational efficiency of our system.

E. User Study

The user study aims to assess the usability of our system from subjects' subjective perceptions [58]. After data collection, subjects are asked to complete a questionnaire by responding to 7 questions on a 5-point Likert scale (with 1 = strongly disagree and 5 = strongly agree). The questions are listed in Table II. All 60 subjects respond to the questions, with Fig. 23 illustrating the results. Overall, most subjects express their satisfaction with HeadSonic, as evidenced by a median score of 5 across 7 questions. The average scores for different questions are also high: Q1 ($\mu = 4.62, \sigma = 0.66$), Q2 ($\mu = 4.43, \sigma = 0.92$), Q3 ($\mu = 4.68, \sigma = 0.56$), Q4 ($\mu = 4.92, \sigma = 0.28$), Q5 ($\mu = 4.65, \sigma = 0.68$), Q6 ($\mu = 4.55, \sigma = 0.76$), and Q7 ($\mu = 4.70, \sigma = 0.75$). In summary, HeadSonic is well perceived by users (Q6, Q7), primarily due to its easy to use (Q1), no discomfort (Q2), easy to learn (Q3), low cognitive load (Q4), and short login time (Q5). The results highlight the usability of HeadSonic.



(a) Our Signal. (b) 50Hz-6kHz chirp signal.

Fig. 24. Visualization of cross-head vibrations, with each sub-figure containing 3-axis accelerometer data under silent and playback states.

F. Comparision With Related Work

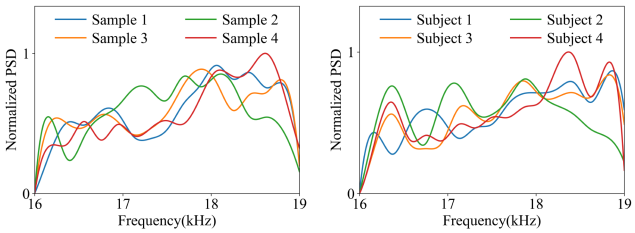
The work most closely related to our research is SkullID [53]. SkullID mounts a surface transducer on the right mastoid process to play probe signals and captures subtle skull-conducted vibrations via piezo-microphones (dedicated hardware) contacted on the left side of the head. In this experiment, we first conduct a case study to investigate whether BCE can capture such cross-head vibrations. Specifically, we mute the left-side speaker of the BCE and use the right-side speaker to emit signals, which are captured by an accelerometer attached to the left side of the BCE shell. Two sounds are employed as the probe signals: i) our frequency-division chirp signals; ii) a chirp from 50 Hz to 6 kHz (employed in SkullID). Each signal is played for 2 seconds at 70% volume. Fig. 24 shows normalized three-axis accelerometer data with and without signal playback. Two key observations emerged. Without playback, the accelerometer readings are small and erratic, dominated by random noise. During playback, signal amplitudes do not change much, suggesting the vibrations are overwhelmed by the noise. This result demonstrates that BCEs struggle to capture vibrations traveling across the whole head. There are two possible reasons. First, in contrast to our scenario, the signal travels a much longer distance, resulting in significant attenuation. Second, the contact area between BCEs and the skin is typically covered with rubber as a buffer, which further attenuates the signal and introduces noises. This means that although the microphone on one side of the BCE can capture sound from the other side, it is picking up air-borne leakage rather than signals transmitted across the head.

We further investigate the feasibility of using such air-borne sound for authentication. Following the configuration in Section II-C, we employ a 1.5 s, 16–19 kHz signal, which is emitted by BCE's right-side speaker and received by its left-side microphone. Fig. 25 displays the normalized PSDs of received signals from the same and different subjects. We observe that the

TABLE III
A COMPARISON OF MOST RELATED AUTHENTICATION METHODS

Method	Biometric	Devices	Sensors	Actions	Performance	Experimental conditions					
						Body behaviors	Ambient noises	Human speech	Wear angles	Speaker volumes	Security analysis
EarEcho [15]	Ear canal geometry	Wired earphone	In-ear speakers and in-ear microphones	None	94.52% BAC	✓	✓	×	✓	×	✓
Earmonitor [16]	Ear canal geometry	Wired earphone	In-ear speakers and in-ear microphones	None	96.40% BAC	✓	✓	✓	✓	×	×
EarDynamic [17]	Ear canal deformation	Wired earphone	In-ear speakers and in-ear microphones	Voice commands	93.04% Acc.	✓	✓	N/A	✓	×	✓
Wang et al. [21]	Skull structure	VR	In-air speaker and in-air microphone	None	98.87% Acc.	×	✓	×	N/A	×	✓
SkullID [53]	Skull structure	Smartglasses	Bone Conduction transducer and piezo-microphone	None	2.94% EER	×	✓	×	×	×	✓
SkullConduct [56]	Skull structure	Smartglasses	In-air speaker and in-air microphone	None	6.94% EER	×	×	×	×	×	×
HeadSonic	Skull structure	Bone conduction earphone	Bone conduction transducer and in-air microphone	None	96.59% BAC	✓	✓	✓	✓	✓	✓

Note that denotes "studied" condition, and × denotes conditions "not considered".



(a) Intra-class study. (b) Inter-class study.

Fig. 25. Normalized PSD of four air-borne samples from he same subject (a) and different subjects (b).

PSD profiles vary significantly for the same subject. We further compute the Person Correlation Coefficients (PCCs) for intra-class and inter-class sample PSDs. The average PCCs are 0.61 and 0.52 for intra-class and inter-class, while in Section II-C, they are 0.98 and 0.57, respectively. The results demonstrate that the method in SkullID is challenging to employ for BCE authentication.

VI. RELATED WORK

In this section, we review two kinds of biometrics-based authentication systems related to HeadSonic.

Earable-based Authentication: Existing earable-based authentication schemes fall into two main categories: passive and active sensing. Passive sensing methods leverage sensors to capture signals emitted by users. For example, TeethPass [10] and ToothSonic [11] use microphones to capture teeth occlusion sound from the ear canal for authentication. EarGate [13] verifies users with the uniqueness of footstep sounds transmitted to the ear canal. EarSlide [12] utilizes sounds generated when users slide their fingers across their faces for authentication. However, these methods [10], [11], [12], [13] still require the user's active participation and are easily affected by the user's behavior inconsistency. Gao et al. [54] identify users by capturing voice from both air and the ear canal. But it requires users to speak and is not suitable in some environments (e.g., library and conference room). HeartPrint [14] and breathSign [9] capture heartbeat and breathing sounds from the ear canal, but these sounds are too faint and easily drowned out by background noise.

Active sensing methods involve emitting probing signals actively from the speakers. For example, EarEcho [15] and Earmonitor [16] emit regular conversations and repetitive chirps into the ear, analyzing the reflections to characterize the ear canal's geometry. EarDynamic [17] transmits an ultrasonic signal into the ear canal to measure its deformation for authentication. But emitting ultrasound into the ear canal often raises health concerns [55]. Apart from the limitations mentioned above, all existing earable-based research relies on capturing target biometrics from the ear canal with in-ear microphones, which BCEs do not have. Therefore, these studies can not be directly applied to BCEs.

Bone Conduction-based Authentication: Bone conduction-based authentication exploits body tissues, especially bones, for signal propagation, enabling captured signals to inherently contain physiological traits. For example, TouchPass [57] characterizes finger touches on a smartphone screen with active vibration sensing. Taprint [30] regards the knuckles on the hand back as a virtual number pad. During input, the tapping vibration propagates from the hand to the wrist, detected by the smartwatch's IMU. Some studies have investigated the potential of employing head biometrics for wearable authentication [21], [53], [56]. For example, Wang et al. [21] propose a VR device authentication system incorporating acoustic sensing. The proposed system actively emits an ultrasonic signal, which passes the hollow space enclosed by the VR device and the face and is finally recorded with the VR microphone. SkullID [53] develops a smartglasses prototype that positions a speaker on the right mastoid process to emit signals, and uses piezo-microphones contacted on the left side of the head to record the audio responses. But it requires dedicated piezo microphones, whereas BCEs typically use in-air microphones for calls. Moreover, from the experiments in Section V-F, we demonstrate that the subtle cross-head vibrations are hard for BCE to capture, and the air-borne sounds are challenging to employ for BCE authentication. SkullConduct [56] leverages a Google Glass speaker to emit white noises behind the ear, which traverse across the head

and are recorded by the in-air microphone in front of the brow. However, due to the speaker being air-gapped from the head, the signal transmission paths in the system are unclear. Moreover, SkullConduct is impractical as its performance significantly degrades under environmental noise [53].

Compared with existing active acoustic sensing-based authentication methods [15], [16], [17], [21], [53], [56], our approach features a different signal transmission mode, resulting in variations in the captured biometric traits. Specifically, most active acoustic sensing approaches rely on air-gap acoustic signals, which are emitted by speakers away from the skin and undergo multipath reflections on the body surface before reaching the microphone. The propagation characteristics cause biometrics related to body shape, such as ear canal and head geometry, to dominate the received signal. In contrast, the BCE speaker directly contacts the head, where the acoustic wave travels through, undergoing damping and refraction before reaching the microphone. This makes biometrics like head density and tissue properties embedded in the received signal. The distinct sound transmission modes (along the body surface and within the body) shape the unique biometric traits analyzed in our approach.

Overall, different from these studies, we develop a novel BCE authentication system by 1) utilizing BCE's distinct head-contact architecture to measure head biometrics at the near-ear position; 2) devising a unique probe signal based on the frequency-division multiplexing approach for adding synthesis attack difficulty; 3) designing a novel biometric representation that characterizes the head conduction and head geometry characteristics; and 4) designing a BANN model tailored for our scenario to enable behavior-irrelevant feature reconstruction. These practical concerns are, by and large, omitted in prior works. Table III outlines the key differences between our approach and existing methods.

VII. Conclusion

This paper proposes HeadSonic, a usable BCE authentication system based on head biometrics. The system interacts with the human head via acoustic sensing, capturing head-conducted sounds for authentication. We devise a unique probe signal employing frequency-division multiplexing to increase synthesis attack difficulty. We extract novel biometric features that characterize the uniqueness of head conduction and head geometry. We also design a BANN model that reconstructs the extracted features into behavior-irrelevant representations. Extensive experiments validate the authentication efficacy of our system in various scenarios, including its effectiveness in preventing various spoofing attacks.

References

[1] R. R. Choudhury, "Earable computing: A new area to think about," in *Proc. Int. Workshop Mobile Comput. Syst. Appl.*, 2021, pp. 147–153.

[2] C. Min, A. Mathur, and F. Kawsar, "Exploring audio and kinetic sensing on earable devices," in *Proc. 4th ACM Workshop Wearable Syst. Appl.*, 2018, pp. 5–10.

[3] Report: Bone conduction headphones market segmentation, 2024. [Online]. Available: https://www.linkedin.com/pulse/bone-conduction-headphones-market-segmentation-zkqxe

[4] N. Bui et al., "eBP: A wearable system for frequent and comfortable blood pressure monitoring from user's EAR," in *Proc. 25th Annu. Int. Conf. Mobile Comput. Netw.*, 2019, pp. 1–17.

[5] W. Xie, Q. Hu, J. Zhang, and Q. Zhang, "Earspiro: Earphone-based spirometry for lung function assessment," in *Proc. ACM Conf. Interactive, Mobile, Wearable Ubiquitous Technol.*, 2023, pp. 1–27.

[6] J. Prakash, Z. Yang, Y.-L. Wei, H. Hassanieh, and R. R. Choudhury, "Earsense: Earphones as a teeth activity sensor," in *Proc. ACM Conf. Mobile Comput. Netw.*, 2020, pp. 1–13.

[7] Y. Jin et al., "Earcommand: " hearing" your silent speech commands in EAR," in *Proc. ACM Conf. Interactive, Mobile, Wearable Ubiquitous Technol.*, 2022, pp. 1–28.

[8] K. Li, R. Zhang, B. Liang, F. Guimbretière, and C. Zhang, "EariO: A low-power acoustic sensing earable for continuously tracking detailed facial movements," in *Proc. ACM Conf. Interactive, Mobile, Wearable Ubiquitous Technol.*, 2022, pp. 1–24.

[9] F. Han, P. Yang, S. Yan, H. Du, and Y. Feng, "Breathsign: Transparent and continuous in-ear authentication using bone-conducted breathing biometrics," in *Proc. IEEE Conf. Comput. Commun.*, 2023, pp. 1–10.

[10] Y. Xie, F. Li, Y. Wu, H. Chen, Z. Zhao, and Y. Wang, "TeethPass: Dental occlusion-based user authentication via in-EAR acoustic sensing," in *Proc. IEEE Conf. Comput. Commun.*, 2022, pp. 1789–1798.

[11] Z. Wang, Y. Ren, Y. Chen, and J. Yang, "ToothSonic: Earable authentication via acoustic toothprint," in *Proc. ACM Conf. Interactive, Mobile, Wearable Ubiquitous Technol.*, 2022, pp. 1–24.

[12] Z. Wang, Y. Wang, and J. Yang, "EarSlide: A secure ear wearables biometric authentication based on acoustic fingerprint," in *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, 2024, pp. 1–29.

[13] A. Ferlini, D. Ma, R. Harle, and C. Mascolo, "EarGate: Gait-based user identification with in-ear microphones," in *Proc. ACM Conf. Mobile Comput. Netw.*, 2021, pp. 337–349.

[14] Y. Cao, C. Cai, F. Li, Z. Chen, and L. Jun, "HeartPrint: Passive heart sounds authentication exploiting in-ear microphones," in *Proc. IEEE Conf. Comput. Commun.*, 2023, pp. 1–10.

[15] Y. Gao, W. Wang, V. V. Phoha, W. Sun, and Z. Jin, "EarEcho: Using ear canal echo for wearable authentication," in *Proc. ACM Conf. Interactive, Mobile, Wearable Ubiquitous Technol.*, 2019, pp. 1–24.

[16] X. Sun et al., "Earmonitor: In-ear motion-resilient acoustic sensing using commodity earphones," in *Proc. ACM Conf. Interactive, Mobile, Wearable Ubiquitous Technol.*, 2023, pp. 1–22.

[17] Z. Wang, S. Tan, L. Zhang, Y. Ren, Z. Wang, and J. Yang, "EarDynamic: An ear canal deformation based continuous user authentication using in-ear wearables," in *Proc. ACM Conf. Interactive, Mobile, Wearable Ubiquitous Technol.*, 2021, pp. 1–27.

[18] J. Li, K. Fawaz, and Y. Kim, "Velody: Nonlinear vibration challenge-response for resilient user authentication," in *Proc. ACM Conf. Comput. Commun. Secur.*, 2019, pp. 1201–1213.

[19] L. Huang and C. Wang, "PCR-auth: Solving authentication puzzle challenge with encoded palm contact response," in *Proc. IEEE Symp. Secur. Privacy*, 2022, pp. 1034–1048.

[20] Wikipedia. rigid body. 2024. [Online]. Available: https://en.wikipedia.org/wiki/Rigid_bodyt

[21] R. Wang, L. Huang, and C. Wang, "Low-effort VR headset user authentication using head-reverberated sounds with replay resistance," in *Proc. IEEE Symp. Secur. Privacy*, 2023, pp. 3450–3465.

[22] L. Van der Maaten and G. Hinton, "Visualizing data using T-SNE," *J. Mach. Learn. Res.*, 2008, pp. 2579–2605.

[23] L. Lu et al., "LipPass: Lip reading-based user authentication on smartphones leveraging acoustic signals," in *Proc. IEEE Conf. Comput. Commun.*, 2018, pp. 1466–1474.

[24] C. Wu, J. Chen, K. He, Z. Zhao, R. Du, and C. Zhang, "EchoHand: High accuracy and presentation attack resistant hand authentication on commodity mobile devices," in *Proc. ACM Conf. Comput. Commun. Secur.*, 2022, pp. 2931–2945.

[25] S. Weinstein and P. Ebert, "Data transmission by frequency-division multiplexing using the discrete fourier transform," *IEEE Trans. Commun. Technol.*, vol. 19, no. 5, pp. 628–634, Oct. 1971.

[26] L. Frenzel, *Principles of Electronic Communication Systems*. New York, NY, USA: McGraw-Hill, Inc, 2007.

[27] Y. Jin et al., "SonicASL: An acoustic-based sign language gesture recognizer using earphones," in *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, 2021, pp. 1–30.

[28] W. Chen et al., "ViObject: Harness passive vibrations for daily object recognition with commodity smartwatches," in *Proc. ACM Conf. Interactive, Mobile, Wearable Ubiquitous Technol.*, 2024, pp. 1–26.

[29] W. E. Siri, "The gross composition of the body," in *Proc. Adv. Biol. Med. Phys.*, 1956, pp. 239–280.
[30] W. Chen et al., "Taprint: Secure text input for commodity smart wristbands," in *Proc. 25th Annu. Int. Conf. Mobile Comput. Netw.*, 2019, pp. 1–16.
[31] M. E. Ahmed, I.-Y. Kwak, J. H. Huh, I. Kim, T. Oh, and H. Kim, "Void: A fast and light voice liveness detection system," in *Proc. USENIX Secur. Symp.*, 2020, pp. 2685–2702.
[32] Y. Meng et al., "Your microphone array retains your identity: A robust voice liveness detection system for smart speakers," in *Proc. USENIX Secur. Symp.*, 2022, pp. 1077–1094.
[33] P. Welch, "The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Trans. Audio Electroacoust.*, vol. 15, no. 2, pp. 70–73, Jun. 1967.
[34] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.
[35] C. Wu, K. He, J. Chen, Z. Zhao, and R. Du, "Liveness is not enough: Enhancing fingerprint authentication with behavioral biometrics to defeat puppet attacks," in *Proc. USENIX Secur. Symp.*, 2020, pp. 2219–2236.
[36] A. Vaswani, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
[37] Y. Ren, Y. Wang, S. Tan, Y. Chen, and J. Yang, "Person re-identification in 3D space: A {WiFi} vision-based approach," in *Proc. USENIX Secur. Symp.*, 2023, pp. 1077–1094.
[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[39] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2000, pp. 93–104.
[40] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. IEEE Int. Conf. Data Mining*, 2008, pp. 413–422.
[41] L. M. Manevitz and M. Yousef, "One-class SVMS for document classification," *J. Mach. Learn. Res.*, vol. 2, pp. 139–154, 2001.
[42] A. Laura and R. Moro, "Support vector machines (SVM) as a technique for solvency analysis," DIW Berlin Discussion Paper, 2008.
[43] wikipedia. voice frequency. 2024. [Online]. Available: https://en.wikipedia.org/wiki/Voice_frequency
[44] H. Zhu, M. Xiao, D. Sherman, and M. Li, "Soundlock: A novel user authentication scheme for VR devices using auditory-pupillary response," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2023, pp. 1–18.
[45] A. Rattani, B. Freni, G. L. Marcialis, and F. Roli, "Template update methods in adaptive biometric systems: A critical review," in *Proc. Adv. Biometrics: Third Int. Conf.*, 2009, pp. 847–856.
[46] F. Roli, L. Didaci, and G. L. Marcialis, "Adaptive biometric systems that can improve with use," in *Advances in Biometrics: Sensors, Algorithms and Systems*, Berlin, Germany: Springer, 2008.
[47] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *Proc. Int. Conf. Pattern Recognit.*, 2010, pp. 3121–3124.
[48] S. Sugrim, C. Liu, M. McLean, and J. Lindqvist, "Robust performance metrics for authentication systems," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2019, pp. 1–15.
[49] T. K. Kim, "T test as a parametric statistic," *Korean J. Anesthesiol.*, vol. 68, no. 6, pp. 540–546, 2015.
[50] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[51] P. Hu et al., "Accear: Accelerometer acoustic eavesdropping with unconstrained vocabulary," in *Proc. IEEE Symp. Secur. Privacy*, 2022, pp. 1757–1773.
[52] P. Hu, Y. Ma, P. S. Santhalingam, P. H. Pathak, and X. Cheng, "Milliear: Millimeter-wave acoustic eavesdropping with unconstrained vocabulary," in *Proc. IEEE Conf. Comput. Commun.*, 2022, pp. 11–20.
[53] H. Shin et al., "Skullid: Through-skull sound conduction based authentication for smartglasses," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2024, pp. 1–19.
[54] Y. Gao, Y. Jin, J. Chauhan, S. Choi, J. Li, and Z. Jin, "Voice in ear: Spoofing-resistant and passphrase-independent body sound authentication," in *Proc. ACM Conf. Interactive, Mobile, Wearable Ubiquitous Technol.*, 2021, pp. 1–25.
[55] T. G. Leighton, "Are some people suffering as a result of increasing mass exposure of the public to ultrasound in air?," in *Proc. Roy. Soc. A: Math., Phys. Eng. Sci.*, vol. 472, 2016, Art. no. 20150624.
[56] S. Schneegass, Y. Oualil, and A. Bulling, "SkullConduct: Biometric user identification on eyewear computers using bone conduction through the skull," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2016, pp. 1379–1384.
[57] X. Xu et al., "Touchpass: Towards behavior-irrelevant on-touch user authentication on smartphones leveraging vibrations," in *Proc. ACM Conf. Mobile Comput. Netw.*, 2020, pp. 1–13.
[58] Z. He et al., "Eyeauth: Smartphone user authentication via reflexive eye movements," *Front. Comput. Sci.*, vol. 19, no. 9, p. 199810, 2025.