Voice In Ear: Spoofing-Resistant and Passphrase-Independent Body Sound Authentication

YANG GAO, University at Buffalo, State University of New York, USA
YINCHENG JIN, University at Buffalo, State University of New York, USA
JAGMOHAN CHAUHAN, University of Southampton, UK
SEOKMIN CHOI, University at Buffalo, State University of New York, USA
JIYANG LI, University at Buffalo, State University of New York, USA
ZHANPENG JIN, University at Buffalo, State University of New York, USA

With the rapid growth of wearable computing and increasing demand for mobile authentication scenarios, voiceprint-based authentication has become one of the prevalent technologies and has already presented tremendous potentials to the public. However, it is vulnerable to voice spoofing attacks (e.g., replay attacks and synthetic voice attacks). To address this threat, we propose a new biometric authentication approach, named *EarPrint*, which aims to extend voiceprint and build a hidden and secure user authentication scheme on earphones. *EarPrint* builds on the speaking-induced body sound transmission from the throat to the ear canal, i.e., different users will have different body sound conduction patterns on both sides of ears. As the first exploratory study, extensive experiments on 23 subjects show the *EarPrint* is robust against ambient noises and body motions. *EarPrint* achieves an Equal Error Rate (EER) of 3.64% with 75 seconds enrollment data. We also evaluate the resilience of *EarPrint* against replay attacks. A major contribution of *EarPrint* is that it leverages two-level uniqueness, including the body sound conduction from the throat to the ear canal and the body asymmetry between the left and the right ears, taking advantage of earphones' paring form-factor. Compared with other mobile and wearable biometric modalities, *EarPrint* is a low-cost, accurate, and secure authentication solution for earphone users.

CCS Concepts: • **Security and privacy** → **Authentication**; **Biometrics**; • **Human-centered computing** → **Ubiquitous and mobile devices**.

Additional Key Words and Phrases: Voiceprint, authentication, earphones

ACM Reference Format:
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. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 12 (March 2021), 25 pages. https://doi.org/10.1145/3448113

Authors' addresses: Yang Gao, University at Buffalo, State University of New York, Department of Computer Science and Engineering, Buffalo, NY, 14260, USA, ygao36@buffalo.edu; Yincheng Jin, University at Buffalo, State University of New York, Department of Computer Science and Engineering, Buffalo, NY, 14260, USA; Jagmohan Chauhan, University of Southampton, Department of Electronics and Computer Science, Southampton, UK; Seokmin Choi, University at Buffalo, State University of New York, Department of Computer Science and Engineering, Buffalo, NY, 14260, USA; Jiyang Li, University at Buffalo, State University of New York, Department of Computer Science and Engineering, Buffalo, NY, 14260, USA; Zhanpeng Jin, University at Buffalo, State University of New York, Department of Computer Science and Engineering, Buffalo, NY, 14260, USA.

2474-9567/2021/3-ART12 $15.00
https://doi.org/10.1145/3448113

1 INTRODUCTION

Biometrics utilizing traditional human physiological and behavioral characteristics, like fingerprints, faces, voices, and finger touches, have been widely adopted in mobile devices for user authentication purposes. With the advances of wearable technologies, a wide variety of smart wearable devices have been enormously popularized by the general public, such as smartwatches, smart wristbands, and especially wireless earphones that have been significantly developed and widely used in recent years. It is reported that the global earphones and headphones market size is expected to reach 126.7 billion dollars by 2027, expanding at a CAGR (Compound Annual Growth Rate) of 20.3% from 2020 to 2027 [43]. With the new trend of smart devices getting grown and more data being stored, there is an increasing demand for secure authentication. Meanwhile, such smart wearable devices as earphones themselves would also provide potentials for more accessible, convenient, and secure authentication approaches.

Among all existing mobile biometric modalities, voiceprint has become one of the most prevalent methods due to the ubiquitous applicability with wide-scale adoption in smartphones and smart devices. Compared with high-cost and privacy-concerned face biometrics (i.e., FaceID [3]) and spoofing-vulnerable fingerprints authentication, voiceprints possess higher social acceptance [6] and more enormous commercial potentials (e.g., Google Trusted Voice [9], Wechat Voiceprint Lock [57]). However, existing voiceprint-based authentication also has its own limitations. First of all, voice is easily exposed to the public, which raises the risk of voice spoofing and replay attacks [49]. Moreover, unexpected body conditions (e.g., coughing, sore throat, and hoarse voice) would also degrade the robustness of voice authentication.

It is a known fact that when a sound wave passes through human tissues and bones, the wave energy is partially absorbed, scattered, and degraded along with the propagation. The energy loss can also be represented in the time and frequency domains. When the user speaks, the resulting sound wave propagates from the throat and can be collected by the inward-facing microphone in the ear canal. The energy loss of the sound wave is dependent on the characteristics of the propagation media, such as tissues and bones. Our key contribution is to unveil that the in-body propagation of speaking-induced sound waves embed the unique characteristics of the human body and can be considered a biometric trait. Since every individual possesses a unique body characteristic, we hypothesize that different people's body sound conduction should also hold distinct patterns (EarPrint). This work aims to explore the feasibility and validation of a new voice-evoked body conduction biometric modality and open discussions for mobile and wearable human authentication research.

In this paper, without using any additional wearable device, we propose a novel earphone-based user authentication system, named "*EarPrint*", utilizing the sounds collected by dual microphones to verify the user's identity as shown in Fig 1. Compared with existing mobile and wearable authentication approaches (e.g., FaceID, fingerprint, voiceprint), *EarPrint* possesses two advantages in real-world application scenarios:

1) **Ubiquitousness:** We make use of the low-cost microphones in off-the-shelf earphones, and this biometric trait is adoptable across earphones and other hearables with diverse flexibility.
2) **Anti-spoofing:** Different from the traditional voiceprint authentication, our proposed *EarPrint* leverages both the in-air voice and the in-ear body sound to extract the multi-level unique body conduction features, with the advantages of preventing voice spoofing attacks.

With the goal of developing an earphone-oriented daily authentication solution, we first identify and discuss several technical challenges that need to be addressed: (1) *EarPrint* is built on the propagation and transmission of body sounds. How to extract and represent the unique features between the voice and the in-ear body sound? (2) To ensure a high level of usability, it is imperative to provide users with passphrase freedom. How to combine the passphrase-independent model with the body sound feature embedding? (3) Assuming the user's voice has been eavesdropped, how to evaluate our system's spoofing resistance?

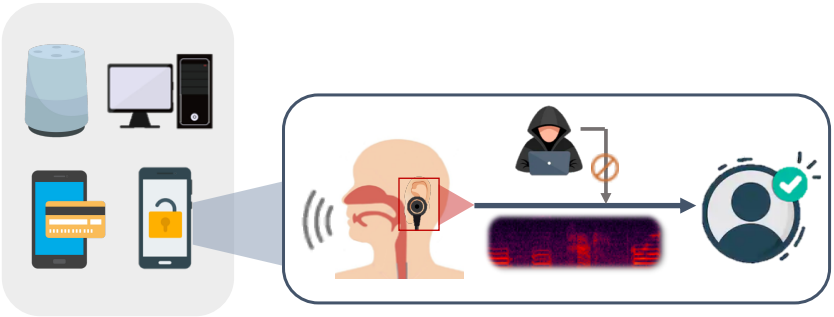


Fig. 1. EarPrint: A multi-level, spoofing-resistant authentication scheme in ear using voice-evoked, unique body conduction.

In the rest of this paper, we begin with presenting the scientific rationale for the uniqueness of sound wave propagation in the human body in Section 3.2.2, and perform the feasibility study in Section 3.3. Then we introduce the overview of *EarPrint* system in Section 4. In Section 5, we discuss the front-end signal processing scheme, including microphone layout design, audio segmentation, body sound enhancement, and augmentation strategy. Afterward, we introduce the novel multi-level descriptors in Section 6. In Section 7.2, we propose the Encoder-to-Decoder based feature transfer learning and passphrase-independent user authentication model with the Gaussian Mixture Model and Universal background Model (GMM-UBM). Finally, the system implementation and a comprehensive evaluation are performed with 23 subjects focusing on the usability and robustness of our proposed system in Section 8 and 9. In the end, we discuss the limitations and conclusion in Sections 10 and 11 respectively.

Specifically, we make the following contributions in this work:

- We explore a novel in-ear body sound-based biometric approach for user authentication. Our study reveals that when the user speaks, the dual, in-ear body sound channels contain intrinsic, unique body signatures.
- We characterize the uniqueness of body sound conduction during speaking by utilizing sound collected from both inward-facing and outward-facing microphones of two sides, and thus propose *EarPrint*, an encoder-to-decoder based multi-level biometric system to provide a convenient, ubiquitous, and secure authentication solution for earphone users.
- We perform extensive experimental evaluations about the system's performance under diverse application scenarios. Our results indicate that *EarPrint* can achieve very low EER (3.64%), high resistance against malicious attacks, and potential robustness against ambient noises (< 1% EER drop) and body motions (< 2% EER drop).

2 RELATED WORK

In this section, we review two major biometric modalities related to *EarPrint* in mobile and wearable authentication. Furthermore, to provide a better background of earphone-based sensing, we also introduce some recent researches about in-ear sensing applications.

2.1 Voiceprint-based User Authentication

Among various mobile and wearable user authentications, the voiceprint-based approach is one of the most convenient and promising approaches [21, 64, 65]. However, as discussed in Section 1, traditional voiceprint represents the physiological characteristics of the speaking behavior embedded in the speech voices without a clear indication of the liveness information of the user, which implies its vulnerability to various voice spoofing

techniques, such as the most accessible replay attacks by playing back the recorded speech. It was reported that [56], playback attacks could achieve a 40% attack rate to the voiceprint-based speaker verification, which indicates the voiceprint authentication has poor resiliency to replay attacks. Some recent researches [8, 40] have aimed to design end-to-end speaker verification systems with the replay attack detectors. Zhang *et al.* [64] proposed the VoiceGesture, an articulatory gesture-based liveness detection on smartphones to enhance authentication security. The system emitted high-frequency acoustic sound from the smartphone speaker and captured the reflected echoes when a user speaks a passphrase. The Doppler shift patterns of the reflected echoes caused by articulatory gestures were then analyzed for liveness detection. Similarly, Lu *et al.* [29] developed an FMCW acoustic-based user authentication by sensing the user's vocal tract during the speaking. However, active acoustic sensing is highly sensitive to position variations. As claimed by the authors, the relative position between the smartphone and the vocal tract is a major limitation, reducing system usability and robustness.

2.2 Body Sound-based User Authentication

In addition to those voiceprint-based approaches which utilize air-conducted sound, many studies have explored and validated that the body-conducted sound can also serve as a biometric trait. Yegnanarayana *et al.* [47, 62] conducted a feasibility study of a throat-microphone-based speaker recognition system in a noisy environment. The authors compared the performance of both the throat-mounted microphone and the close-speaking microphone, and they concluded that the throat-contact microphone is more robust to noise and reverberation. Similar research efforts on robust speaker recognition using the throat microphone have been widely reported in literature [46, 66]. A hybrid speaker verification method using both bone-conduction microphone and air-conduction microphone was proposed by Tsuge *et al.* [53]. Their results showed that the integration of bone-conduction and air-conduction microphones could achieve an average 9.5%-11.6% Identification Error Rate (IER) among 99 subjects, higher than the sole air-conducted microphone. A recent study of body sound authentication [27] implemented a wearable microphone prototype and tested it with two different machine-learning methods to verify whether the device is on the speaker's body. However, it is less practical and inconvenient for daily users to wear the extra device mounted on the throat or body.

2.3 In-ear Sensing

Recent advances in wearable devices such as earphones have brought great opportunities to bring human physiological sensing (e.g., PPG, EEG) [16, 54] and behavioral analysis (e.g., jaw movements, eating behaviors) [4, 33] out of the clinic and into people's daily life. Park *et al.* [38] designed a piezoelectric sensor measuring the pressure variances near the ear canal surface and estimated the user's heart rates. Bi *et al.* [5] proposed Auracle, an earpiece-based eating behavior recognition system, by capturing and analyzing the body sound of chewing conducted through the bone and tissue in the head. Goverdovsky *et al.* [17] proposed the "Hearables", an earpiece with multi-modal sensors platform, including EEG electrodes and a microphone. This system can monitor the user's brain, cardiac, and respiratory activities simultaneously. EarEcho [13] utilized the built-in speaker and microphone of the earbud to capture the acoustic profile of the user's ear canal for authentication. Similarly, Amesaka *et al.* [2] developed a facial expression recognition system utilizing the user's ear canal transfer functions corresponding to different facial muscle movements. Bui *et al.* [7] proposed a device called "eBP" to monitor blood pressure inside the user's ear canal, which brought a huge impact on users' daily health monitoring with a high comfort level. Besides the physiological signal monitoring, in-ear sensing can also provide a new way of human-computer interaction. The "Earable" [39] was an ear-worn bio-signal sensing device that can monitor the user's cognitive state and further serve as the human-computer interaction platform.

3 THREAT MODEL AND FEASIBILITY STUDY

In this section, we first introduce the attack scenarios of existing voiceprint-based speaker verification systems. We further provide the background on speaking-induced body sounds and the rationale behind the underlying uniqueness from the perspectives of physical and physiological body properties. We also perform a feasibility study as the proof-of-concept.

3.1 Attack Scenarios

Speaking is one of the most common activities in our daily lives, which carries the content information and also embeds the unique characteristics of each individual based on the physical configuration of the speaker's mouth and throat, the so-called "voiceprint." Specifically, the voiceprints refer to the acoustic frequency spectrum that carries the speech information in a human voice. Like fingerprints, voiceprints have unique biometric signatures, are individual-specific, and can function as an identification method. However, existing voiceprint-based authentication often suffer from various voice spoofing attacks [1, 48, 61]. Here we list two primary types of attacks for off-the-shelf voiceprint-based authentication.

Voice Replay Attacks: As the voiceprint-based authentication process is mostly conducted in an open space, the malicious attackers have a high chance to eavesdrop the user's voice information by recording or side-channel attacks. Then the adversary would be able to spoof the system by playing back the eavesdropped voice [40].

Synthetic Voice Attacks: In some scenarios (e.g., the attacker can not be physically close to the user), it might not be easy to directly record the user's voice of passphrase information. There still are possibilities of incorporating the user's non-passphrase voice and mouth movement to implement synthetic attacks [11] by utilizing deep learning-based Text-to-Speech (TTS) engines [36, 55].

3.2 Attack-resistant User Authentication through Body Sound (EarPrint)

3.2.1 Behavior of Speaking-induced Body Sound. The human vocalization system generally consists of three parts: the lungs, vocal folds, and articulators, including the tongue, palate, and lip. The lungs first generate enough airflow to vibrate the vocal folds [32], and then the muscles of the larynx fine-tune the vocal folds to produce a specific pitch and tone. Lastly, the articulators in the mouth and nose articulate and filter the sound from the larynx and finalize it as the actual voice. Meanwhile, the vibration of vocal folds in the throat also generates the body sound wave that is conducted through the tissues and bones, which is shown in Fig. 2.

Generally, the uniqueness of the vocal tract is not only represented by the air-conducted voice but also embedded in the body-conducted sound. Except for the vocal tract, the human body tissues and bones also hold the individual uniqueness in terms of physical and physiological properties, which affects the propagation of the body sound (will be discussed in Section 3.2.2). Since the ear canal is close to the throat, the conducted body sound has not been entirely absorbed or degraded by the tissues. Hence, both the vocal tract and tissue embed the individual uniqueness to the in-ear body sound, EarPrint, which demonstrates the potentials of a new biometric trait that could be integrated with earphones.

3.2.2 Characterizing EarPrint for User authentication. As the voice is essentially a type of energy being generated by vibrations of the throat and propagated through various media (e.g., air, tissues and bones), based on [58], we model the body sound propagation process in the following equations (1-6). Firstly, we define the general wave propagation equation as below:

$$\nabla^2 p = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} \tag{1}$$

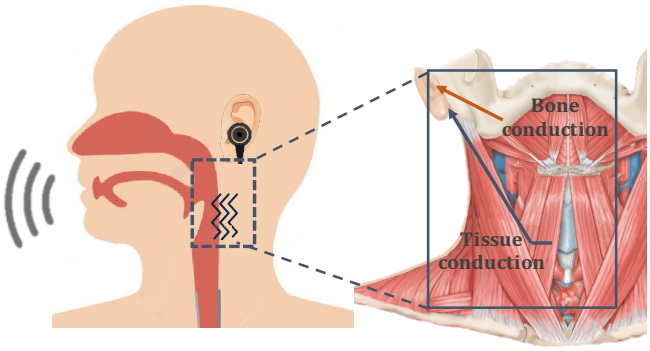


Fig. 2. The speaking-induced body sound wave propagates through the tissue conduction and bone conduction and can be captured by the inward-facing microphone.

where c is the propagation speed in the medium, p is the sound wave. Considering that no reflection boundary is in the propagation, the sound wave can be expressed as:

$$p(t,x) = p_0 cos(\omega t - kx),\ k = \omega/c \tag{2}$$

where ω is the sound angular frequency. Given μ is the wave viscosity of the tissue. The wave equation for sound propagation in tissue can be defined as:

$$\frac{\partial^2 p}{\partial x^2} + \frac{\mu}{B}\frac{\partial^2 p}{\partial x \partial t} - \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} = 0 \tag{3}$$

$$p(t,x) = p_0 cos(\omega t - kx)\exp(-\alpha_\eta x) \tag{4}$$

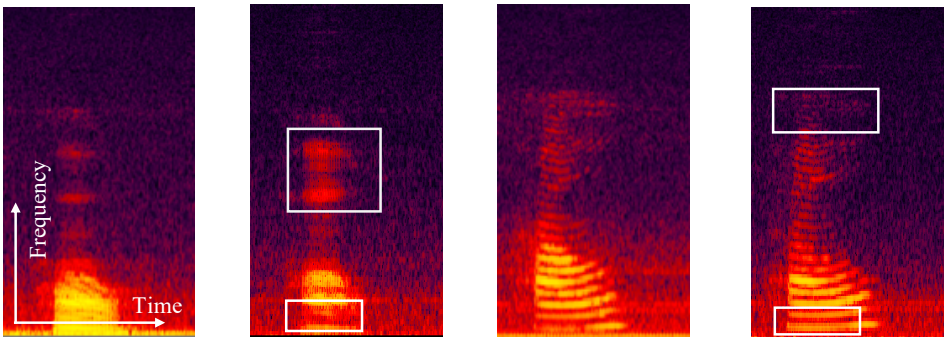
where α is the decay parameter of a tissue's viscosity, which is defined by μ, B, and ω. Considering the sound decay caused by tissue absorption and scattering, given the sound source pressure P_0, the propagation distance x, the received wave P is represented as:

$$P = P_0 \exp(-\alpha x) \tag{5}$$

The decay parameter α indicates the decay phenomenon of sound wave propagation through human soft tissues, including viscosity decay α_η, relaxation decay α_r, and scattering decay α_s (i.e., heat conduction decay is ignored due to the low sound wave frequency), which is defined as:

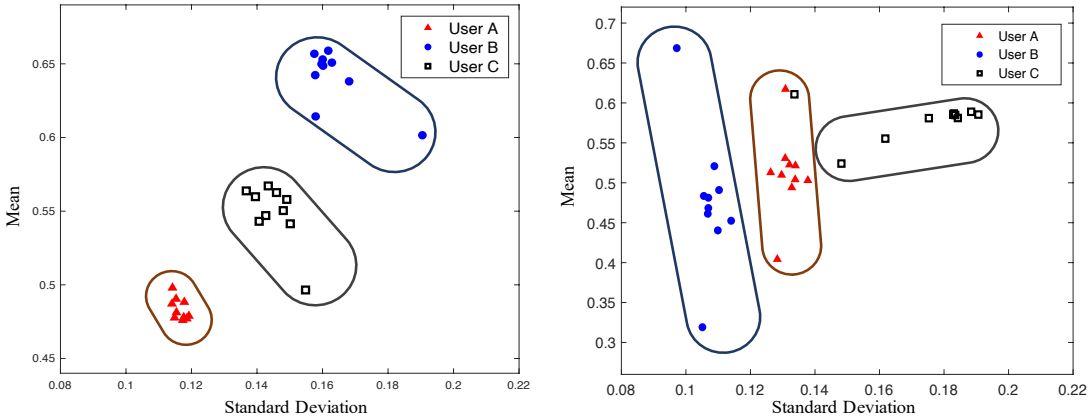
$$\begin{aligned} \alpha_\eta &= \frac{2\pi^2 f^2}{2}\left(\frac{3}{4}\eta_1 + \eta_2\right) \\ \alpha_r &= \frac{\omega^2}{2\rho c^3}\frac{\eta_0}{1+\omega^2\tau^2} \\ \alpha_s &= \frac{8}{9}(2\pi f/c)^2 r^2 \\ \alpha &= \alpha_\eta + \alpha_r + \alpha_s = \mathcal{F}(f,\tau) \end{aligned} \tag{6}$$

The received sound wave P in ear is determined by the decay parameter α, which is hard to be quantified and varies with wave frequency and relaxation time. To better capture the uniqueness of the decay parameter, We utilize the spectrogram to profile the sound wave information. As shown in Fig. 3, the white boxes indicate the differences in the spectrum between air-conducted and body-conducted sounds. The locations of differences vary along with both the frequency and time. In addition, those changes are also dependent on different users.



(a) User 1: air voice. (b) User 1: in-ear voice. (c) User 2: air voice. (d) User 2: in-ear voice.

Fig. 3. Examples of differences in the time-frequency domain caused by body conduction by speaking "how" by two different users.



(a) Level I Descriptor (b) Level II Descriptor

Fig. 4. A pilot study of the distinguishability of the EarPrint biometric modality. The stand deviation and mean of transfer functions between in-ear body sound and air voice. (a) Level I descriptor of the body conduction between the in-ear sound and air voice; (b) Level II descriptor of the asymmetry between the left and right in-ear sounds.

3.3 Pilot Feasibility Study

3.3.1 Proof-of-concept Setup. To investigate the feasibility of characterizing and utilizing body conduction properties of the human voice for user authentication, we conducted a preliminary study (based on 3 subjects) to perform speaking with randomly picked conversation topics, ten times per subject. Taking into the consideration of the variations of earphone wearing positions, we asked three subjects to put on and take off the earphones for each recording. We recorded both the left and right in-ear sounds using a pair of inward-facing microphones. The subjects' air voices were captured by a pair of outward-facing microphones. All sounds were collected with a 48 kHz sampling rate. As the preliminary study, we performed the experiments under the controlled environment with low ambient background noise (40 dB).

3.3.2 Feasibility Analysis. We divide EarPrint into two levels of descriptors, including the body conduction profile between the air voice and the in-ear body sound (*Level-I Descriptor*), and the body asymmetry information between the left and the right in-ear sounds (*Level-II Descriptor*). As a preliminary study, we first choose the transfer function to describe the mapping relationship between two audio signals which is defined as:

$$H_1(f) = \frac{P_{yx}(f)}{P_{xx}(f)} \tag{7}$$

where $H_1(f)$ is the estimated transfer function, P_{yx} and P_{xx} are the cross power spectral density of sound x and sound y respectively, which can be estimated using the Welch's averaged periodogram method [59] by dividing the time-series signal into non-overlapping successive blocks. Figure 4 illustrates the variations against the mean and standard deviation of descriptors after normalization. Each EarPrint (i.e., transfer function estimation) represents a data point on the scatter plot and the points from multiple trials from the same subject exhibit a cluster.

3.3.3 Insights and Summary. Our preliminary analysis reveals that: (1) different subjects hold the observable distinctions of body conduction patterns between the in-ear sound and air voice as shown in Figure 4(a), (2) every subject also holds a clear and distinguishable body asymmetry pattern between the left and right in-ear sounds as shown in Figure 4(b), and (3) variations in earphone wearing positions have very limited effects on the identification of both the body conduction and body asymmetry information.

To better extract the underlying uniqueness, we further adopt the well-designed machine learning models to extract the features in both Level I and Level II descriptors from the collected in-ear body sound and air voice. In the following section, we will discuss the implementation of EarPrint for earphone users and provide extensive experiments on robustness evaluation before real-world deployment.

4 SYSTEM OVERVIEW

In this study, we propose the *EarPrint*, an earphone-based user authentication system, which utilizes the dual microphones of the earphone to characterize the user's body sound conduction. The end-to-end methodological framework and processing flow are shown in Fig. 5, which consists of three major components: signal processing, feature representation, and authentication model.

Signal Processing. We utilize the dual microphones (i.e., outwards-facing and inwards-facing) in the earphones to record the voice sounds propagated in air and conducted by the body, respectively. We segment the sound frames through voice activity detection (VAD). In the enhancement state, the segmented sound is filtered by a band-pass filter to remove the high-frequency leaked-in noise. To reduce the noises caused by the variations of earphone wearing positions, we also augment the sound with the time-warping and frequency-time masking.

Feature Representation. To learn the body conduction uniqueness and provide noise-resistant features, the air sound and in-ear sound features are fed into an encoder-to-decoder based feature transfer model to generate enhanced body sound features as the Level I descriptor of *EarPrint*. We further extract asymmetric information as the Level II descriptor between the body sounds collected from the left and right in-ear microphones by utilizing transfer function estimations.

Authentication Modeling. Based on the transferred features of the Level I descriptor, we construct a GMM-UBM based passphrase-independent authentication model. The user's Level II descriptor of asymmetric information is fed into an ensemble decision tree classifier. In the end, we perform the soft voting on the outputs from the two classifiers to verify the user's identity.

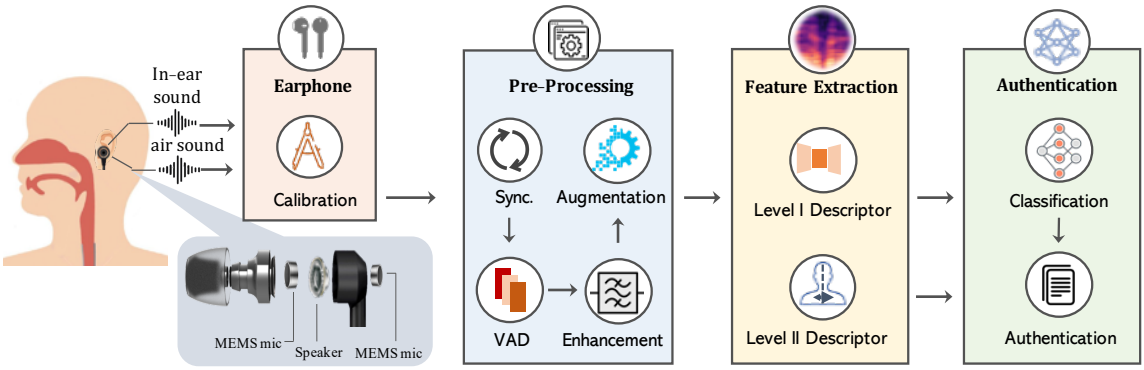


Fig. 5. Diagram of *EarPrint* methodological framework, consisting of an signal pre-processing, feature extraction module, and an authentication module to verify the user's identity.

5 EARPRINT PROCESSING

In this section, we will first introduce the dual microphone setting in modern earphones and then further present the event detection, data augmentation, body sound enhancement, and multi-level EarPrint.

5.1 Earpiece Microphones Selection and Setting

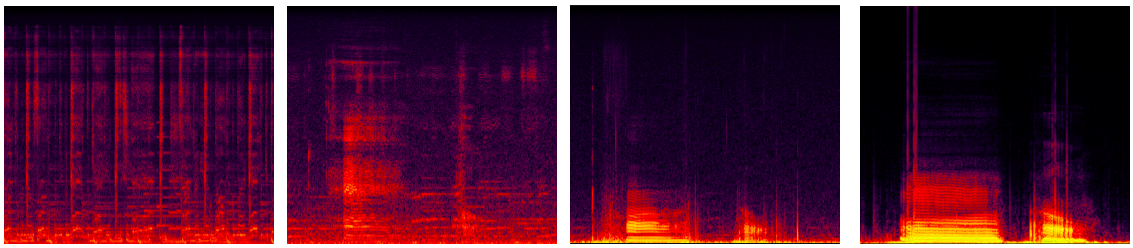
As a ubiquitous biometric solution for mobile and wearable scenarios, it is critical and imperative to provide an affordable, accessible, and highly usable hardware. Different from some existing body sound-based authentication solutions [19, 27] that require additional contact microphones (e.g., throat-mounted, chest-mounted), we choose to utilize the existing earphone design with embedded dual microphones on both sides of the ears (e.g., Apple AirPods Pro), as shown in Fig. 5. The dual microphones were originally designed for the active noise cancellation [41], where the outward-facing microphone collects the background noise and the inward-facing microphone also captures the unwanted sound inside the ear for further anti-noise processing. In our design, when the user speaks for voice authentication, the inward-facing microphone (fully attached to the ear canal with isolation of a sponge) will collect the in-ear body sound, and the outward-facing microphone listens to the voice vocalized from the user's mouth and propagating in the air. To validate that our inward-facing microphone design can filter out the air-conducted sound and capture the body sound, we asked the participant to speak a few words in a noisy environment (i.e., loud background music) and then recorded the sounds via outward-facing, inward-facing, and contact microphones respectively. As shown in Fig. 6, the sound captured by the inward-facing microphone (c) shows similar patterns as the contact microphone (d) and is not affected by the air-conducted noise compared with the outward-facing microphone (b).

5.2 Signal Synchronization

As the multiple microphones (i.e.,g left and right, inward-facing and outward-facing microphones) all have individual channels, we connect those microphones into the Bela board, an ultra-low-latency embedded platform, which could provide multiple audio I/Os. To further ensure a better synchronization between multi-channel audios, we compute the cross-correlations between three pairs of audio signals. Then we adjust the audios by the relative time lags that are indicated by the maximum values of the cross-correlations.

5.3 Segmentation and Event Detection

During the earphone's daily usage, to eliminate the effects of the low signal-to-noise ratio (SNR) segments and to increase the energy efficiency especially for the continuous authentication scenarios, we adopt the VAD algorithm



(a) Background (b) Outward-facing Mic. (c) Inward-facing Mic. (d) Contact Mic.

Fig. 6. Spectrum of recorded sounds when the user speaks the word "apple" in a noisy environment. (a) represents the spectrum of background music; (b) shows the mixture sound captured by the outward-facing conventional microphone; (c) shows body sound captured by the inward-facing microphone; (d) is the body sound recorded by the contact microphone.

in WebRTC [15] to filter out undesired low-power density and silent audio segments. We first divide the audio spectrum into 6 sub-bands (i.e., 80 Hz-250 Hz, 250 Hz-500 Hz, 500 Hz-1000 Hz, 1 kHz-2 kHz, 2 kHz-3 kHz, and 3 kHz-4 kHz), and then calculate the corresponding sub-band energy as feature vectors of the Gaussian Mixture Model (GMM). Given the assumption that the audio input only contains speeches and noises, we create two GMMs of speech and noise respectively for every sub-band. Given that:

$$P(x_k|z, r_k) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_k-\mu_z)^2}{2\sigma^2}}, \; x \sim N(\mu, \sigma^2) \tag{8}$$

where x_k is the sub-band energy, r_k is the combination of μ_z and σ, $z = 0$ represents the probability of noise frame, and $z = 1$ represents the probability of speech frame. For every sub-band, the 2D likelihood ratio (LR) is defined as:

$$\begin{aligned} L_i(x_k) &= \log(\frac{P_s(x_k, i)}{P_n(x_k, i)}) \\ L_t(x_k) &= \sum_{i=1}^{M} K_i L_i(x_k), \; M = 6 \end{aligned} \tag{9}$$

where $L_i(x_k)$ is the log-likelihood ratio of each sub-band, $L_t(x_k)$ is the weighted sum of log-likelihood ratios for all sub-bands, K_i represents the weighted LR parameter. The probability of speech detection is:

$$F_{\text{vad}} = \begin{cases} 1, & L_i > T_\tau \;||\; L_t > T_\alpha \\ 0, & else \end{cases} \tag{10}$$

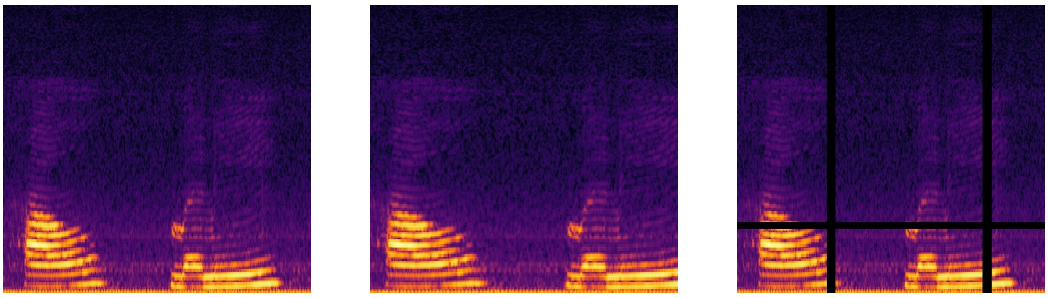
where T_τ is the threshold of sub-band log-likelihood ratio, and T_α is the threshold of weighted log-likelihood ratio.

5.4 Body Sound Enhancement.

As the skull and human tissues conduct low-frequency sound better than the air [26, 44], which is the reason why a person's voice sounds different when it is recorded and played back, the power spectrum of in-ear sound covers from 20 Hz to 4000 Hz. We thus design a Butterworth low-pass filter with stop frequency at 4,000 Hz and feed the raw in-ear sound into the low-pass filter to remove undesired high-frequency leaked-in noise.

5.5 Data Augmentation

To achieve better robustness against noises, we aim to generate additional noisy training data caused by variations of earphone wearing positions for our deep neural network model. Inspired by the recent success of data



(a) Raw spectrum (b) Time Warping. (c) Frequency-Time Masking.

Fig. 7. Augmentation policies applied to the Mel-spectrogram, including time warping and time-frequency masking.

augmentation in the speech recognition domain, we utilize the augmentation policy proposed by Park *et al.* [37], which consisted of warping the features along with time steps, masking blocks of frequency channels, and masking blocks of time steps. As shown in Fig. 7, given an Mel-spectrogram with τ time steps, time warping is applied by fixing anchor points on the boundary - four corners and two mid-points of the vertical edges, and warping the random points along with the horizontal line within the time step $(W, \tau - W)$ to either left or right by a certain distance $w \in [0, W]$, where W is the time-warp parameter. Time-Frequency is applied by masking consecutive frequency channels $[f_0, f_0 + \Delta f)$ and time steps $[t_0, t_0 + \Delta t)$, where Δf is chosen from a uniform distribution from 0 to the frequency mask parameter F, and $f_0 \in [0, v - \Delta f]$, and Δt is chosen from a uniform distribution from 0 to the time mask parameter T, and $t_0 \in [0, \tau - \Delta t)$. In this paper, we use both the time warping and time-frequency masking with customized augmentation parameters, as shown in Fig. 7.

6 EARPRINT FEATURE REPRESENTATION

The speaking-induced body sound waves conduct omnidirectionally through our body and reach to our ear canal of both sides. Unlike the traditional body sound detection approaches, which primarily consider one sound source location, earphones have the advantages of dual-channel placement. Thus, we propose a novel multi-level descriptor of intrinsic body sound that combines the uniqueness of body conduction and the uniqueness of body asymmetry into the EarPrint design, as shown in Fig. 8. The detailed pipeline is discussed below.

6.1 Level I Body Conduction Descriptor

In *EarPrint*, Level I features describe the physical uniqueness during the body sound conduction from the vocal to the ear canal. These features are embedded in both the time and frequency domains, as discussed in Section 3.2.2, which can be represented by the Mel-spectrogram of the body sound and vocal voice. However, vocal voice is usually mixed with environment noises, and body sound is robust against ambient noise but is sensitive to the earphone friction caused by body motions (also known as stethoscope effect [30]) and other body noises such as swallows and yawns, which barely affect the vocal voice. Therefore, directly using raw in-ear body sounds as the body conduction features would be easily affected by body motions. To characterize the body conduction and also generate the enhanced body sound for authentication (i.e., filtering out undesired environment noises and body noises), we design the encoder-to-decoder based feature transfer from vocal voice features to in-ear sound features. Specifically, to capture the minor distinction between in-ear sound and vocal voice, in the Level I body conduction descriptor, we select Mel-spectrograms with 36 Mel filter bands and 64-frame window with a 5 ms frame period.

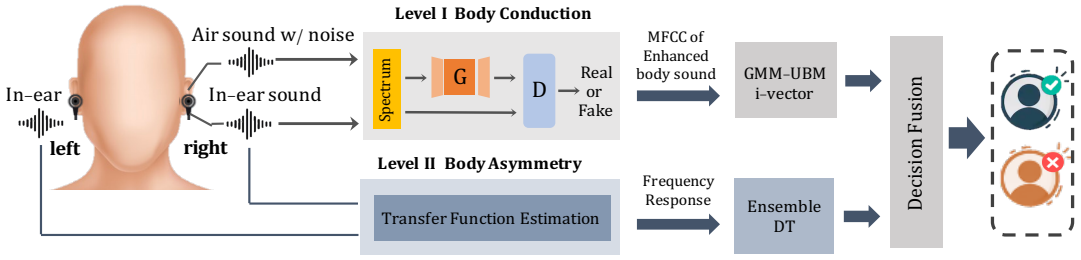


Fig. 8. Multi-level uniqueness descriptor, including body conduction and body asymmetry.

6.2 Level II Body Asymmetry Descriptor

Although the Level I body conduction descriptor represents a certain degree of uniqueness of the human body's physical structures, to provide a more unique and hidden biometric amongst other existing solutions, similar with biometrics of asymmetrical faces and body shapes [14, 25], the inherent tissues and skeletal structures (e.g., jaws) also shows asymmetry [35] which affects the body sound conduction. Thus, we propose the Level II descriptor based on the asymmetry between the left and right in-ear body sounds, taking advantage of the pairwise configuration of earphones. We extract the transfer function estimations (discussed in Section 3.3.2) between the body sounds of each side to describe the body asymmetric information. Finally, we choose 20 Hz to 4,000 Hz frequency response coefficients with the 200 ms Hanning window and 100 ms overlap length.

7 EARPRINT AUTHENTICATION MODELING

7.1 Encoder-Decoder Network Design for Feature-based Transfer Learning

We hypothesize that there exists an intrinsic mapping between the air voice and body sound, including a uniform representation and user-level uniqueness. As discussed in 6.1, to model the connection between those two sound sources and generate enhanced in-ear body sound robust against both ambient noises and body noises, we propose a cycle-consistent adversarial network (CycleGAN) [23] based feature representation transfer learning. To enhance the robustness of generated features against environmental noises, we first partially select the collected clean air voice, and mix with white noises (45 dB) as background. Afterward, we extract Mel-spectrogram features from both the noise-mixed air voice and clear in-ear body sound as the input and output for the generator adversarial model.

Figure. 9 shows the general architecture of the CycleGAN-based model, including the generator and the discriminator. The input of the CycleGAN generator (direction is from the air voice to the body sound) is the spectrogram features of air voice, and the output is the estimated spectrogram features of the body sound. The generator adopts the encoder-to-decoder scheme to find the mapping from the source domain (air voice) to the target domain (body sound). The task of the discriminator is to distinguish the generated data (output of the generator) and the real data (real spectrogram of body sound). By minimizing the cycle consistency loss, the cycleGAN can well capture the mapping from the air voice to the body sound.

Speech/body sound has its uniqueness from sequential (i.e., time-domain) and hierarchical (i.e., frequency-domain) structures. An effective way to represent sequential structures would be to use an RNN, but it is computationally demanding due to the difficulty of parallel implementations. Instead, we configure a CycleGAN using gated CNNs [10] that not only allow parallelization over sequential data but also achieve the state-of-the-art in speech modeling. In a gated CNN, gated linear units (GLUs) [10] are used as an activation function. A GLU is a data-driven activation function, and the gated mechanism allows the information to be selectively propagated depending on the previous layer states.

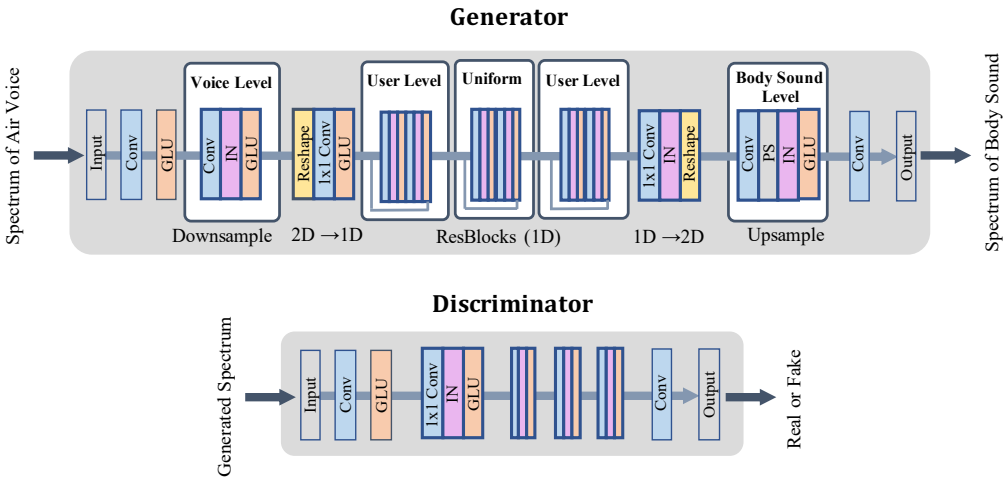


Fig. 9. Architecture of the cycleGAN for feature-based transfer learning, where *IN* denotes instance normalization, *GLU* denotes gated linear unit, *PS* denotes pixel shuffler.

7.2 User Authentication Model

7.2.1 Level I Descriptor. To implement the passphrase-independent user authentication, we adopt a Gaussian Mixture Model and Universal Background Model (GMM-UBM) with i-vectors [31]. Given spectrograms of the transferred body sounds from the output of CycleGAN, we first calculate spectrograms into Mel-frequency Cepstral Coefficients (MFCCs) which are beneficial with linear models such as the GMM. Then GMM-UBM models the statistical characteristics of individual uniqueness without containing text related information.

7.2.2 Level II Descriptor. Given the Level II body asymmetry features, we explore several classifiers which are very efficient in user identification tasks, including the Support Vector Machine (SVM), Random Forest (RF), Ensemble of Decision Trees, and Linear Discriminant Analysis (LDA). According to the experiment results, the Gentle-AdaBoost ensemble decision tree [60] with surrogate splits achieved the best performance.

Gentle AdaBoost learning is known for numerically robust to noisy data and avoiding the overfitting [63]. Given a set of training data as $\{(x_1, y_1), \ldots, (x_N, y_N)\}$, where $y_i \in \{+1, -1\}$ is the label of the training data x_i. Boosting learning generally offers an additive model in a sequential manner by the form of $F(x) = \sum_{m=1}^{M} f_m(x)$, where $f_m(x)$ is a weak learner, M is the number of ensemble learners and $F(x)$ is the strong learner. Gentle AdaBoost adopts adaptive Newton steps to minimize the cost, which is defined as $J = E[e^{-yF(x)}]$. In each step, the weak classifier $f_m(x)$ is designed to minimize the weighted squared error as below:

$$J_w = \sum_{i=1}^{N} \omega_i (y_i - f_m(x_i))^2 \tag{11}$$

where N is the number of training samples, ω_i is the training-based weight for each sample and is updated by each iteration.

7.2.3 Decision Fusion. After obtaining the results from GMM-UBM and Ensemble DT as P_I and P_{II}, we implement the soft voting mechanism to generate the final user verification result. The assembled prediction result P_{total} can be defined as:

$$P_{\text{total}} = \alpha_1 P_{\text{I}} + \alpha_2 P_{\text{II}} \tag{12}$$

where, α_1 and α_2 are the classifier decision weights. In this work, based on the empirical analysis, body-sound-based GMM-UBM is the dominant classifier compared with the Ensemble DT using the asymmetry information from both sides of the ears. Thus, we assign the weight to the GMM-UBM and Ensemble DT as 0.6 and 0.4, respectively, to decide the legitimate user.

8 SYSTEM IMPLEMENTATION

8.1 Dataset

8.1.1 Preparation. We conducted extensive experiments to validate the effectiveness and robustness of our proposed *EarPrint*. Human participants were instructed to sit on the chair in a casual and comfortable position. We asked the participants to put on and wear the earphones in their daily way of life with a comfortable position and great stability. The sounds were recorded by and sent to a desktop equipped with an Nvidia GeForce GTX 1080ti 11GB GPU.

8.1.2 Participants. 23 subjects (5 females and 18 males) aged from 24 to 30 years were recruited in our experiments. These participants included both native and non-native English speakers. All subjects had experiences of using voiceprint-based applications. The experiment was approved by the Internal Review Board (IRB) of the [University name is hidden for double-blinded review] for human subjects.

8.1.3 Data Collection. To ensure the generality and practicability of our experiment, we randomly chose some daily conversations under five different topics, including family, restaurant, book, travel, and website, in the English speaking website [51]. Overall, our reading materials cover 100 sentences and approximately 500 unique words. During the experiment, all participants were asked to wear the earphones prototype and read the materials for over 10 minutes in a controlled lab environment (40 dB). Both air-conducted voices and body sounds are recorded at the sampling rate of 48,000 Hz. To collect clean in-ear body sound data for training the CycleGAN, we asked participants to sit down without any big body motion during the data collection. In addition, to mimic the earphone's usage in daily lives, participants were asked to take off and put back on the earphones for 5 times in total during intervals of different speaking sessions. In total, we have over 180 minutes data for training and testing.

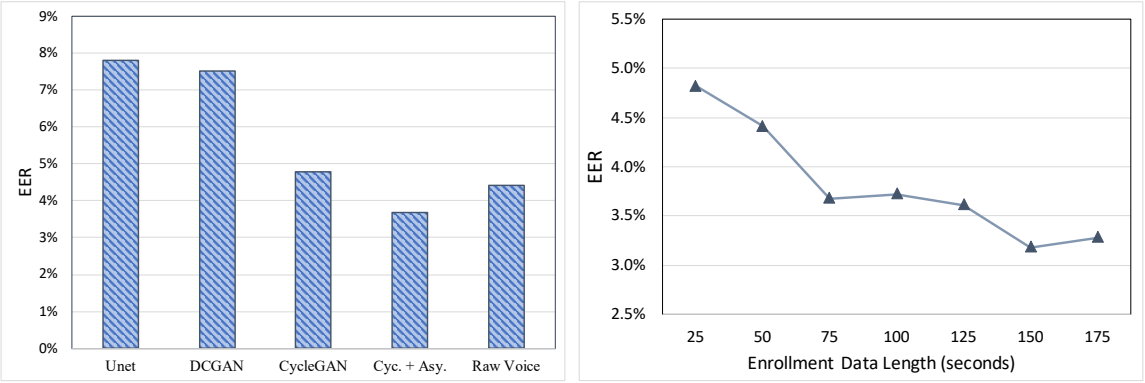
8.2 Models Implementation

Feature Transfer Learning. We implemented the proposed CycleGAN in Tensorflow. We pre-trained the base model on the VCC2018 dataset [28] in a non-parallel setting, which is a public dataset for voice conversion tasks from a source speech to a target speech and contains 12 subjects with 81 sentences per subject. We used the Adam optimizer with an initial generator learning rate of 0.001, a discriminator learning rate of 0.0005, $\lambda_{cyc} = 10$, and $\lambda_{id} = 5$ to train the CycleGAN.

Authentication Model. We first selected a sub-set which contains 100 subjects of the VoxCeleb dataset [34] (i.e., originally contains over 100,000 utterances for 1,251 celebrities from videos on YouTube) to extract 20-coefficients MFCCs with 25 ms Hann windows and 10 ms hop-length, and then pre-train the base Gaussian UBM with 2,058 components. Afterward, the speaker-dependent total variability space model with 512 dimensions and i-vectors were trained. Then, an LDA and WCCN based projection matrix was created to indicate which i-vector corresponds to which speaker. During the enrollment, we randomly selected 75 seconds of data for each subject as the enrollment data, and the remaining dataset was the testing data.

8.3 Evaluation Metrics

In this authentication problem, we introduce the equal error rate (EER) as the evaluation metrics, which describes the overall accuracy of a biometric system. Specifically, EER describes the point where the FRR and FAR are equal



(a) (b)

Fig. 10. (a) EERs with 75 seconds enrollment data over different network structures. (Cyc. means CycleGAN, and Asy. means asymmetric information) (b) EERs under different subject enrollment data lengths.

along with different sensitivity (i.e., decision threshold). Given the true positive (TP), false negative (FN), false positive (FP), and true negative (TN), FAR and FRR can be defined as:

$$\text{FAR} = \frac{FP}{FP + TN}, \quad \text{FRR} = \frac{FN}{FN + TP} \tag{13}$$

9 PERFORMANCE EVALUATION

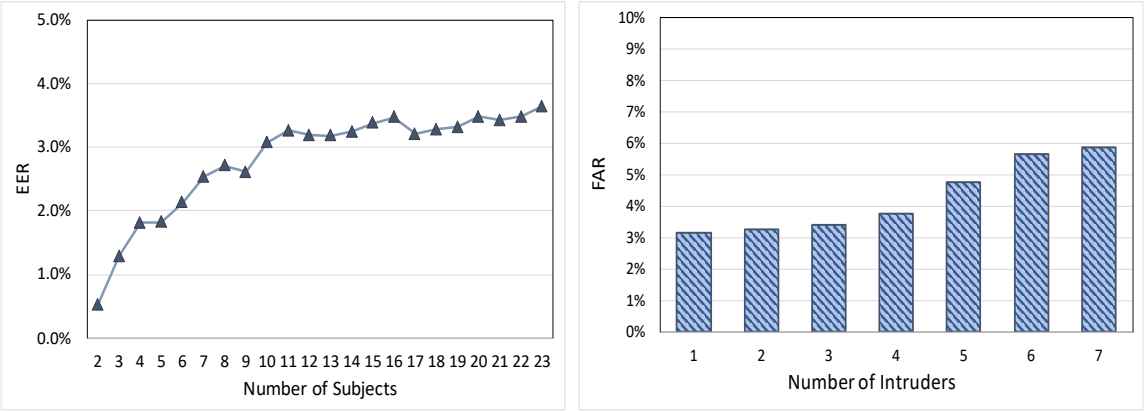
In this section, we aim to evaluate the *EarPrint* from the following three perspectives: 1) overall accuracy for different numbers of subjects and enrollment data lengths, 2) resistance to replay attacks, and 3) robustness and reliability against environmental factors.

9.1 Recognition Performance

9.1.1 Performance over Different Network Structures. To evaluate the performance of the encoder-decoder based transferred features with CycleGAN, we compared different encoder-decoder network structures, including Pixel2Pixel GAN [20], DCGAN [42], and the raw voice. Fig. 10(a) shows the detailed EERs for different feature transfer strategies. It is shown that CycleGAN outperforms the Pixel2Pixel GAN and DCGAN, especially when we incorporated the body asymmetry features into the authentication decision.

9.1.2 Performance over Different Enrollment Data Lengths. New user enrollment is a major factor in measuring the usability of any mobile and wearable authentication solutions. It is necessary to evaluate how the data collection amount would affect the system's performance. Fig. 10(b) shows the *EarPrint*'s performance along with the increasing enrollment data demands for each new subject. Theoretically, the more enrollment data are collected for training, the lower EER our system could achieve. However, a too-long enrollment duration will reduce the user experience significantly. To get a balanced trade-off between performance and usability, we select 75 seconds as our required enrollment data length.

9.1.3 Performance over Enrolled Users. A well-designed user authentication should have a stable performance over an increasing number of enrolled users. To evaluate the performance of different numbers of subjects, we varied the number of enrolled subjects and tested the EERs with cross-validations. According to the results in Fig. 11(a), as the number of subjects increases, the EERs also increase slightly. With over 11 subjects, our model shows promising potential for generalization with a stable EER (< 4%). However, due to the wide-spreading pandemic,



(a) (b)

Fig. 11. (a) Equal Error Rate under different numbers of enrolled new users with 75 seconds enrollment data. (b) False Acceptance Rates under different numbers of intruders with 75 seconds enrollment data.

we were only able to performed evaluations based on a limited subject pool. A comprehensive and sufficient evaluation of increasing subjects for EarPrint will be included in future work.

9.1.4 Performance over Intruders. To examine the vulnerability of *EarPrint* against an intruder (i.e., the user's data is entirely unknown by the authentication system and is solely used for testing purposes), we randomly chose 16 subjects as the enrolled users, and the remaining 7 subjects were selected as intruders for testing. The averaged False Acceptance Rate (FAR) is shown in Fig. 11(b). It is observed that, as the number of intruders increases, the chance of similar body sound features between enrolled users and intruders also increases, which causes the higher system FARs. Considering the relatively small training dataset, compared with the intruder dataset we evaluated, the FARs are still reasonable. With more subjects enrolled in the training model, we would continue to investigate the generalization of EarPrint among unknown users. Moreover, it is argued that, similar to other widely-adopted biometric modalities, the success of a biometric approach relies on not only the fundamental characteristic uniqueness, but also a large amount of training subject samples accumulated over time.

9.2 Attack Resistance Performance

We assume that the malicious attacker knows the underlying mechanism of *EarPrint* and has eavesdropped the legitimate user's voice during an access attempt through a high-resolution microphone at a distance of 30 cm. In our experiment, 5-minutes legitimate user's voice was recorded.

Attack via Air-conducted Voice: The recording was replayed to the target earphone at the same distance of 30 cm. In total, 60 replay-attack attempts are made, among which, 56.67% of attempts pass the standalone GMM-UBM based voiceprint verification system, but all 60 attempts were rejected by *EarPrint*, which indicates the direct replay attacks via speakers are ineffective.

Attack via Simulated Body Conduction and Air-conducted Voice: Different from the direct air-conducted replay attacks, we consider a scenario that the malicious user attempts to forge both the body sound and air voice for spoofing. As shown in Fig. 12, a 2-Watt bone-conduction transducer is attached on the medical silicon sponge (i.e., similar texture to the human skin). Then the pre-recorded audio is fed into the transducer to generate the body sound vibrations via the Bela audio platform with the built-in amplifier. Meanwhile, the other speaker is simultaneously playing the audio to mimic the air-conducted voice. We also tried multiple sets of attack attempts

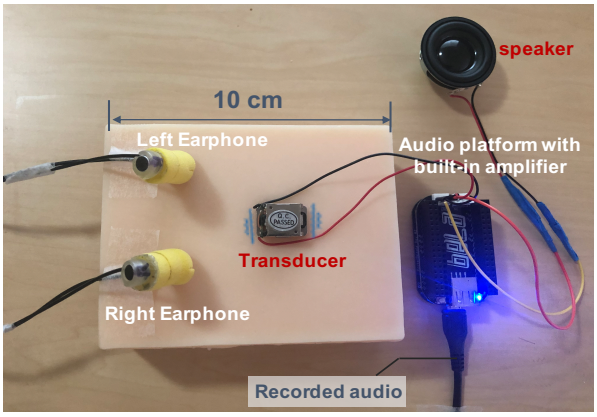


Fig. 12. Spoofing attempts of replayed voice and simulated body conduction via a transducer and the recorded audio.

under different distances between the transducer and the target earphones from 1 cm to 5 cm. All 300 attack attempts were thwarted by the EarPrint system.

Attack via Hybrid Replay and Mimic Voice: We also consider the scenario when the adversary wears the earphones and speaks during the authentication process while the target's voice is also played by a smart speaker nearby. With the same setting as in "Air-conducted Voice", all 60 attempts were rejected by *EarPrint*.

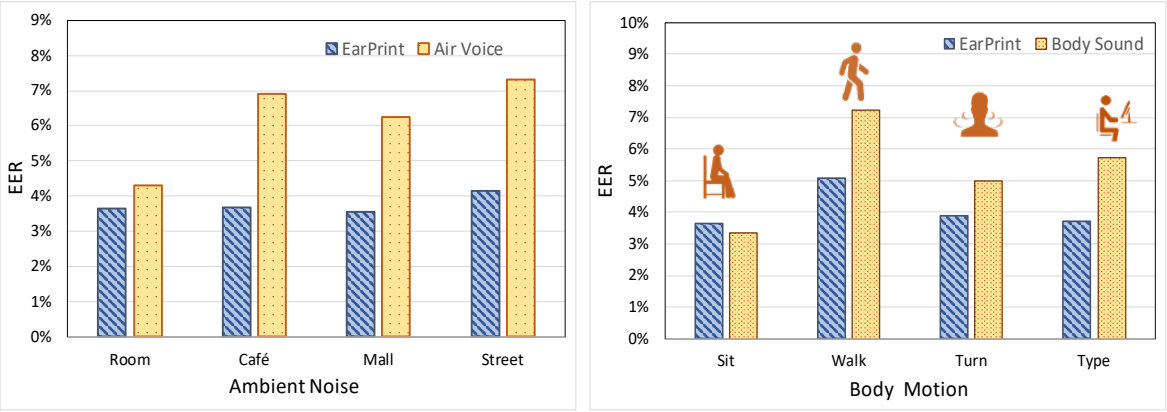
9.3 Robustness Quantification

In this section, we evaluate the robustness of *EarPrint* in terms of background noises and body motions.

9.3.1 Resistance to Ambient Noises. With the popularity of earphones and convenience of voice-based human-computer-interaction, people are more and more willing to use voice-based authentication in many scenarios. However, diverse application scenarios also bring the challenge of being resistance to various background noises. Thus, we examine *EarPrint* system under four types of environment with different ambient noises, including a controlled room (40 dB), a cafe with background music (55 dB), a crowded shopping mall (65 dB), and a noisy street (70 dB). To ensure the replicability and controllability of the experiments, during the training phase, we used the data collected in the controlled room with later mixed white noises and verify the data collected in the real-world noisy environment by playing background noises at the corresponding sound pressure levels. We used dual smart speakers to play the background noises with 40 cm away from the subjects. The results are shown in Fig. 13(a). Compared with the results (4.21%-7.28%) using the GMM-UBM model with air-conducted voice collected by outward-facing microphones, *EarPrint* has lower EERs (2.95%-4.06%) among various noisy environments.

Insights: With the help of body sound enhancement and Encoder-to-Decoder based feature representation, *EarPrint* shows the evidence of being potentially robust against ambient noises. Moreover, the earphone's physical isolation might also help block the noises from the ear-canal cavity.

9.3.2 Reliability on Body Motions. Besides the ambient noises, users' body motions (e.g., walking, turning around, and typing/writing) during the authentication process would also introduce extra body sound noises to be collected by the inward-facing microphones. To evaluate the robustness of *EarPrint* over standalone in-ear sound against body motions, we asked a sub-group (N=2) among the enrolled subjects to implement additional experiments for body motions. We tested our system for different body motions: walking, head movements, and typing.



(a) (b)

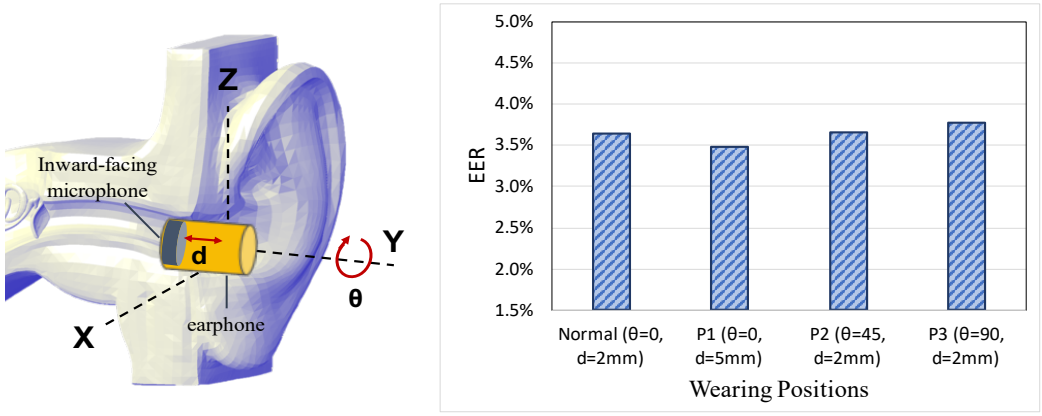
Fig. 13. (a) Evaluation among different ambient noises. (b) Evaluation among different body motions.

Insights: In Fig. 13(b), compared with the EER while baseline sitting, we can observe that the walking behavior brings the largest EER variance, and other body motions such as waving head and typing have relatively lower EER increases. In addition, compared with the performance of the pure in-ear body sound, *EarPrint* has lower EERs, which validates the effectiveness of the enhanced body sound generated by CycleGAN of the Level I descriptor as discussed in Section 6.1.

9.3.3 Reliability on Earphone Wearing Positions. Putting on and taking off the earphones are common actions in people's daily use of earphones. Each time this type of actions may cause a slight different wearing position (e.g., angles and depths) of the earphones relative to the ear canal. To evaluate the system performance in terms of the effects of various earphone wearing positions, we chose a subgroup (N=2) among the total enrolled subjects to implement the study. The earphone wearing positions are quantified into two factors, rotation angle θ (i.e., $\theta = \{0, 45, 90\}$) and in-ear depth d (i.e., $d = \{2\text{mm}, 5\text{mm}\}$), as shown in Fig. 14(a).
Insights: It can be observed from Fig. 14(b) that, wearing position variations of the earphones have all resulted in very similar, low EERs, which demonstrate the stability of the system over varying wearing positions within the normal ranges of daily usage. In particular, a deeper microphone insertion (e.g., d=5mm) results in a smaller EER, partially because the deep insertion provides more solid position fixation and better surface contact. It is worthy to note that, although regular actions to put on or take off the earphones may cause slight variations of earphone wearing positions, those differences are usually quite limited, as people tend to stay with the most comfortable positions in their normal daily usage of earphones, which is restricted by an individual's ear shape and contour.

9.4 On-Device Evaluation

As a wearable and mobile biometric solution, it is necessary to investigate the real-time performance of our system. Thus, we implement the EarPrint in mobile-end. We collect the data on earphones and then evaluate the latency of executing *EarPrint* on different smartphones. As shown in Table. 1, given a 5-second audio sample as an authentication input, we divide the system's latency into two parts, including EarPrint generation (MFCC feature transferring) and inference (authentication model). The generation step consumes more computational resources with a higher latency compared with the inference step. Overall, the entire latency is quite low (400 - 500 ms) on all smartphones and could well satisfy the daily-usage requirement. The latency could be further reduced by data pre-processing and model optimization .



(a) (b)

Fig. 14. (a) Simulation of earphone wearing rotation θ and depth d. The default setting is $\theta = 0$, $d = 2$mm. (b) Evaluation results for different wearing positions.

Table 1. The averaged latency on different smartphones

Latency[1] (ms)	iPhone 6S	iPhone X	iPhone 11
EarPrint Generation	316	263	256
EarPrint Inference	168	132	133
Total	484	395	389

[1] Results are tested given the same 5-second sample and averaged for 10 runs.

9.5 User Experience Study

To evaluate the comfort and acceptance of EarPrint, we conducted a survey with 10 participants to collect their feedback on a 7-point Likert Scale in terms of usability, effort, performance, and safety. To make the survey more comparable and easy to understand, we requested the participants to rate our system in comparison with the commercial smartphone-based voiceprint approach. The result is shown in Fig. 15. The higher score means the better user-experience. Based on the feedback, we can observe that compared with smartphone-based voiceprint authentication, earphone-based EarPrint possesses the advantages of being more usable (e.g., potential robustness in noisy environments), requiring less operational efforts (e.g., hands-free operation during speech), and providing higher security due to the lower chances of eavesdropping and strong resistance against replay and mimic attacks.

10 DISCUSSIONS

10.1 Advantages over Voiceprint Biometric

One of the major challenges for voiceprints is the vulnerability to voice spoofing attacks, especially the replay attacks. In Section 9.2, compared with the standalone voiceprint authentication system, *EarPrint* demonstrates better spoofing resistance. Even though, in recent years, there are several mobile-based anti-spoofing liveness detection approaches for voiceprint, such as VoiceGesture [64], REVOLT [40] and VocalLock [29]. These methods either require the involvement of WiFi infrastructures, or the user needs to hold the smartphone next to the mouth at a fixed position while speaking, which can not provide a robust and convenient user experience during access attempts. With the earphones and earphone-based voice assistance becoming more and more popular,

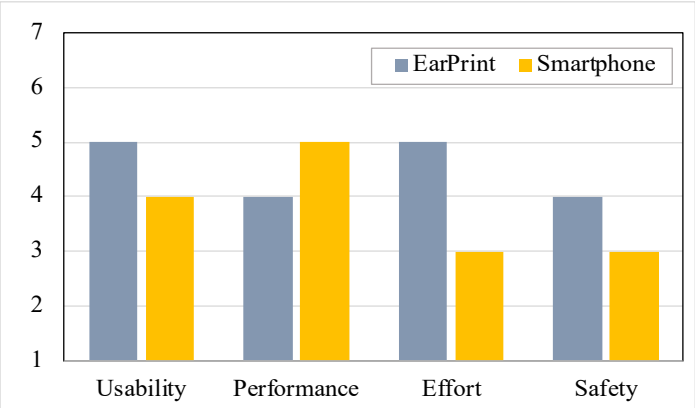


Fig. 15. Results of the user experience study (rate from 1 to 7) in comparison with the existing smartphone-based voiceprint solution.

thus, we propose *EarPrint* as the first ear-worn voiceprint-based authentication that could provide users with a reliable and secure application.

10.2 Comparison with Existing Body-Sound Authentication Solutions

Taking advantage of the body conduction of sound energy [18, 22], prior research has explored using body-conducted sound as a biometric identifier for authentication purpose. For instance, Vocal Resonance [27], a body sound based speaker verification system, captures the sound of the person's voice travelling through the body by an attached contact microphone (in locations like throat, neck, or back of the ear) and then verifies the user's identity. However, just as the fact that the airborne sound-based voice authentication is vulnerable to ambient noises, body-conducted sound-based authentication is sensitive to noises from body movements (e.g., eating, breathing) [18, 50]. The results reported in Vocal Resonance were collected in a quiet lab environment without considering possible noises caused by ambient sounds and body motions. In our work, by utilizing an encoder-to-decoder transferring mechanism to represent the intrinsic mapping correlation between the air-conducted and body-conducted voices, *EarPrint* shows potentials of maintaining robustness in noisy environment including ambient noises and body motions.

In addition to the unique body signatures embedded in the air-to-body voice mapping, we also took advantage of the unique body features reflected from the asymmetry information of both sides of the ears. This multi-modal fusion based authentication can bring an extra level of protection in terms of unique body characteristics, higher authentication accuracy, and stronger resistance against attacks. In addition, we simulate the air-conducted and body-conducted voice attacks, and the result indicates our system holds the strong resistance to those replay attacks. Lastly, the earphone platform in the proposed EarPrint would become an accessible, convenient, and easy-to-use gadget to meet people's daily authentication needs, without any extra device worn or attached to human body restricting their usability and scalability.

10.3 Aging Effects

Permanence is one of the common concerns of any biometric design. It is important to provide a stable performance with low false reject rate over time. Thus, to evaluate the aging effects of our system, we asked one subject to participate in the longitudinal study and collected the subject's data every week for a period of one month. We enrolled the user into our system by collecting 10-minute data as the reference day, and verified the FRR of the user by testing the remaining data. We observed an increase of 2.58 % FRR after three weeks.

Table 2. Comparison with Earphones-based or Speaking-evoked Authentication Approaches

Interfaces	**EarPrint**	VocalLock [29]	EarEcho [13]	Vocal Resonance [27]	SilentKey [52]
Features	Body sound	Vocal Tract	Ear Canal	Body sound	Vocal Tract
Capability	23 Subjects	25 subjects	20 Subjects	29 Subjects	50 Subjects
Devices	Earphone	Smartphone	Earphone	Wearable Microphones	Smartphone
Portable	Yes	Yes	Yes	No	Yes
Accuracy	96.36%	91.1%	94.2%	94.2%-96.1%%	78%-87%

10.4 Comparison with Earphones-based or Speaking-evoked Authentication

We compare the performance of *EarPrint* with other earphones-based or speaking-evoked user authentication solutions deployed on mobile and customized wearable devices, including VocalLock [29], EarEcho [13], Vocal Resonance [27], SilentKey [52]. As shown in Table 2, *EarPrint* shows the highest accuracy with the potential generalizability over increasing subjects. In addition, different from the listed research, *EarPrint* is the first of its kind to characterize the body asymmetry uniqueness embedded underlying the body-conducted sound for authentication by leveraging the pairwise configuration of earphones.

10.5 Limitations and Future Work

In this study, the proposed *EarPrint* presents a new promising way of mobile authentication in daily life. However, the proposed technique still exhibits several limitations in its current stage. In order to further enhance the *EarPrint*'s accuracy and usability, we discuss the future work from the following aspects:

10.5.1 Small Subject Pool. In this work, due to the COVID-19 pandemic, we were only able to involve 23 subjects to participate in our performance evaluation, which is less sufficient for providing a generalized accuracy of user verification. As the future work, we would continue to recruit more subjects and validate the system performance in terms of accuracy and robustness on the expanded dataset.

10.5.2 Robustness on Body Conditions. We also aim to assess the performance of *EarPrint* on users when they get sick as sore or inflamed throat, coughing, and tonsillitis would affect the user's vocalization behaviors.

10.5.3 Music-Listening Scenarios. Our proposed EarPrint approach possesses a great promise in people's daily authentication because it can be easily deployed on any commodity earphones. However, on the other side, there is a concern about the potential conflict with and influence of earphones' regular sound-playing functionality. In the current study, all of our evaluations were based on the assumption that there is no other sounds or music being played by the earphones. However, people often need to request and gain authentication while they are listening to music. In future, to design a highly usable authentication system that can be seamlessly integrated into the user's daily-used earphones, we seek to deploy the interference cancellation module that removes the interference of the sound or music emitted from the earpiece speaker on the inward-facing microphone by analyzing the transfer function and correlation between the played music and received audio.

11 CONCLUSION

In this paper, we proposed an earphone-based user authentication system named *EarPrint*, leveraging the dual microphones to capture and recognize the user in-body sound conduction features. Extensive experiments under real-life scenarios with various simulated configuration and environmental variations have shown the effectiveness and robustness of EarPrint to precisely authorize the user with EER 3.64%.

REFERENCES

[1] Muhammad Ejaz Ahmed, Il-Youp Kwak, Jun Ho Hua, Iljoo Kim, Taekkyung Oh, and Hyoungshick Kim. 2020. Void: A fast and light voice liveness detection system. In *29th USENIX Security Symposium (USENIX Security'20)*. 2685–2702.

[2] Takashi Amesaka, Hiroki Watanabe, and Masanori Sugimoto. 2019. Facial Expression Recognition Using Ear Canal Transfer Function. In *Proceedings of the 23rd International Symposium on Wearable Computers (ISWC '19)*. 1–9. https://doi.org/10.1145/3341163.3347747

[3] Apple. 2020. About Face ID advanced technology. https://support.apple.com/en-us/HT208108. [Online; accessed 30-July-2020].

[4] Abdelkareem Bedri, David Byrd, Peter Presti, Himanshu Sahni, Zehua Gue, and Thad Starner. 2015. Stick it in your ear: Building an in-ear jaw movement sensor. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*. 1333–1338.

[5] Shengjie Bi, Tao Wang, Nicole Tobias, Josephine Nordrum, Shang Wang, George Halvorsen, Sougata Sen, Ronald Peterson, Kofi Odame, Kelly Caine, Ryan J. Halter, Jacob Sorber, and David Kotz. 2018. Auracle: Detecting Eating Episodes with an Ear-mounted Sensor. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 92.

[6] David Braue. 2020. Voiceprint authentication starts to go mainstream in Australia. https://www.csoonline.com/article/3546188/voiceprint-authentication-starts-to-go-mainstream-in-australia.html. [Online; accessed 10-Feb-2020].

[7] 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.

[8] S. Chen, K. Ren, S. Piao, C. Wang, Q. Wang, J. Weng, L. Su, and A. Mohaisen. 2017. You Can Hear But You Cannot Steal: Defending Against Voice Impersonation Attacks on Smartphones. In *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*. 183–195.

[9] Computerworld. 2020. Google Smart Lock: The complete guide. https://www.computerworld.com/article/3322626/google-smart-lock-complete-guide.html. [Online; accessed 30-June-2020].

[10] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *International conference on machine learning*. 933–941.

[11] P. L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga. 2012. Evaluation of Speaker Verification Security and Detection of HMM-Based Synthetic Speech. *IEEE Transactions on Audio, Speech, and Language Processing* 20, 8 (2012), 2280–2290.

[12] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. 2010. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing* 19, 4 (2010), 788–798.

[13] 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.

[14] Afzal Godil, Patrick Grother, and Sandy Ressler. 2003. Human identification from body shape. In *Fourth International Conference on 3-D Digital Imaging and Modeling, 2003. 3DIM 2003. Proceedings*. IEEE, 386–392.

[15] Google. 2016. WebRTC. https://webrtc.org/. [Online; accessed 30-June-2020].

[16] Valentin Goverdovsky, David Looney, Preben Kidmose, and Danilo P Mandic. 2016. In-ear EEG from viscoelastic generic earpieces: Robust and unobtrusive 24/7 monitoring. *IEEE Sensors Journal* 16, 1 (2016), 271–277.

[17] Valentin Goverdovsky, Wilhelm von Rosenberg, Takashi Nakamura, David Looney, David J Sharp, Christos Papavassiliou, Mary J Morrell, and Danilo P Mandic. 2017. Hearables: Multimodal physiological in-ear sensing. *Scientific Reports* 7, 1 (2017), 6948.

[18] Tatsuya Hirahara, Makoto Otani, Shota Shimizu, Tomoki Toda, Keigo Nakamura, Yoshitaka Nakajima, and Kiyohiro Shikano. 2010. Silent-speech enhancement using body-conducted vocal-tract resonance signals. *Speech Communication* 52, 4 (2010), 301 – 313. Silent Speech Interfaces.

[19] Chenyu Huang, Huangxun Chen, Lin Yang, and Qian Zhang. 2018. BreathLive: Liveness detection for heart sound authentication with deep breathing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–25.

[20] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1125–1134.

[21] RG Maduranga M Jayamaha, Maduri RR Senadheera, T Nuwan C Gamage, KD Pavithra B Weerasekara, Gayan A Dissanayaka, and G Nuwan Kodagoda. 2008. VoizLock - human voice authentication system using hidden markov model. In *2008 4th International Conference on Information and Automation for Sustainability*. IEEE, 330–335.

[22] Takeshi Joyashiki and Chikamune Wada. 2020. Validation of a Body-Conducted Sound Sensor for Respiratory Sound Monitoring and a Comparison with Several Sensors. *Sensors* 20, 3 (2020), 1–16.

[23] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. 2019. CycleGAN-VC2: Improved CycleGAN-based Non-parallel Voice Conversion. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.

[24] Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre Dumouchel. 2007. Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing* 15, 4 (2007), 1435–1447.

[25] Leonid Kompanets. 2004. Biometrics of asymmetrical face. In *International Conference on Biometric Authentication*. Springer, 67–73.

[26] Kazuhiro Kondo, Tomoe Fujita, and Kiyoshi Nakagawa. 2006. On equalization of bone conducted speech for improved speech quality. In *2006 IEEE International Symposium on Signal Processing and Information Technology*. IEEE, 426–431.
[27] Rui Liu, Cory Cornelius, Reza Rawassizadeh, Ronald Peterson, and David Kotz. 2018. Vocal Resonance: Using internal body voice for wearable authentication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–23.
[28] Jaime Lorenzo-Trueba, Junichi Yamagishi, Tomoki Toda, Daisuke Saito, Fernando Villavicencio, Tomi Kinnunen, and Zhenhua Ling. 2018. The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods. *arXiv preprint 1804.04262* (2018).
[29] Li Lu, Jiadi Yu, Yingying Chen, and Yan Wang. 2020. VocalLock: Sensing Vocal Tract for Passphrase-Independent User Authentication Leveraging Acoustic Signals on Smartphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2 (2020), 1–24.
[30] MAONO. 2018. What Is the Stethoscope Effect? http://m.maonotech.com/info/what-is-the-stethoscope-effect-29859990.html. [Online; accessed 10-July-2020].
[31] Pavel Matějka, Ondřej Glembek, Fabio Castaldo, Md Jahangir Alam, Oldřich Plchot, Patrick Kenny, Lukáš Burget, and Jan Černocky. 2011. Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4828–4831.
[32] Deirdre D. Michael. 2018. About the voice. http://www.lionsvoiceclinic.umn.edu/page2.htm#physiology101. [Online; accessed 19-Jan-2020].
[33] Mark Mirtchouk, Christopher Merck, and Samantha Kleinberg. 2016. Automated estimation of food type and amount consumed from body-worn audio and motion sensors. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 451–462.
[34] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. VoxCeleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612* (2017).
[35] Kunihiko Nojima, Taishi Yokose, Takenobu Ishii, Makoto Kobayashi, and Yasushi Nishii. 2007. Tooth axis and skeletal structures in mandibular molar vertical sections in jaw deformity with facial asymmetry using MPR images. *The Bulletin of Tokyo Dental College* 48, 4 (2007), 171–176.
[36] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).
[37] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin Dogus Cubuk, and Quoc V Le. 2019. SpecAugment: A Simple Augmentation Method for Automatic Speech Recognition. (2019).
[38] Jang-Ho Park, Dae-Geun Jang, Jung Park, and Se-Kyoung Youm. 2015. Wearable sensing of in-ear pressure for heart rate monitoring with a piezoelectric sensor. *Sensors* 15, 9 (2015), 23402–23417.
[39] Nhat Pham, Taeho Kim, Frederick M Thayer, Anh Nguyen, and Tam Vu. 2019. Earable–An Ear-Worn Biosignal Sensing Platform for Cognitive State Monitoring and Human-Computer Interaction. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. 685–686.
[40] Swadhin Pradhan, Wei Sun, Ghufran Baig, and Lili Qiu. 2019. Combating replay attacks against voice assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–26.
[41] AirPods Pro. 2020. Apple. https://www.apple.com/airpods-pro/. [Online; accessed 30-July-2020].
[42] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv:cs.LG/1511.06434
[43] Grand View Research. 2020. Earphones Headphones Market Size Worth 126.7 Billion By 2027. https://www.grandviewresearch.com/press-release/global-earphones-headphones-market. [Online; accessed 30-July-2020].
[44] Sheldon M Retchin and Martin Lenhardt. 2007. Recreational bone conduction audio device, system. US Patent 7,310,427.
[45] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn. 2000. Speaker verification using adapted Gaussian mixture models. *Digital signal processing* 10, 1-3 (2000), 19–41.
[46] Md. Sahidullah, Rosa Gonzalez Hautamäki, Dennis Alexander Lehmann Thomsen, Tomi Kinnunen, Zheng-Hua Tan, Ville Hautamäki, Robert Parts, and Martti Pitkänen. 2016. Robust Speaker Recognition with Combined Use of Acoustic and Throat Microphone Speech. In *INTERSPEECH 2016*. 1720–1724. https://doi.org/10.21437/Interspeech.2016-1153
[47] A. Shahina and B. Yegnanarayana. 2007. Mapping Speech Spectra from Throat Microphone to Close-Speaking Microphone: A Neural Network Approach. *EURASIP Journal on Advances in Signal Processing* 087219 (2007).
[48] J. Shang, S. Chen, and J. Wu. 2018. Defending Against Voice Spoofing: A Robust Software-Based Liveness Detection System. In *2018 IEEE 15th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*. 28–36.
[49] Maliheh Shirvanian, Summer Vo, and Nitesh Saxena. 2019. Quantifying the Breakability of Voice Assistants. In *2019 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 1–11.
[50] Masaki Shuzo, Shintaro Komori, Tomoko Takashima, Guillaume Lopez, Seiji Tatsuta, Shintaro Yanagimoto, Shin'ichi Warisawa, Jean-Jacques Delaunay, and Ichiro Yamada. 2010. Wearable eating habit sensing system using internal body sound. *Journal of Advanced Mechanical Design, Systems, and Manufacturing* 4, 1 (2010), 158–166.

[51] Basic English Speaking. 2020. ESL Conversation. https://basicenglishspeaking.com/. [Online; accessed 30-June-2020].
[52] Jiayao Tan, Xiaoliang Wang, Cam-Tu Nguyen, and Yu Shi. 2018. SilentKey: A new authentication framework through ultrasonic-based lip reading. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 36.
[53] Satoru Tsuge, Takashi Osanai, Hisanori Makinae, Toshiaki Kamada, Minoru Fukumi, and Shingo Kuroiwa. 2008. Combination method of bone-conduction speech and air-conduction speech for speaker recognition. In *Ninth Annual Conference of the International Speech Communication Association*.
[54] Boudewijn Venema, Johannes Schiefer, Vladimir Blazek, Nikolai Blanik, and Steffen Leonhardt. 2013. Evaluating innovative in-ear pulse oximetry for unobtrusive cardiovascular and pulmonary monitoring during sleep. *IEEE Journal of Translational Engineering in Health and Medicine* 1 (2013), 2700208–2700208.
[55] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135* (2017).
[56] Zhi-Feng Wang, Gang Wei, and Qian-Hua He. 2011. Channel pattern noise based playback attack detection algorithm for speaker recognition. In *2011 International conference on machine learning and cybernetics*, Vol. 4. IEEE, 1708–1713.
[57] WeChat. 2015. Voiceprint: The New WeChat Password. https://blog.wechat.com/tag/voiceprint/. [Online; accessed 30-June-2020].
[58] Kang Weixin, Gong Xue, Wang Hongru, and Pan Dawei. 2017. Frequency characteristic of ultrasonic based on soft tissue attenuation model. In *2017 13th IEEE International Conference on Electronic Measurement & Instruments (ICEMI)*. IEEE, 441–446.
[59] Peter Welch. 1967. The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics* 15, 2 (1967), 70–73.
[60] Shuqiong Wu and Hiroshi Nagahashi. 2015. Penalized AdaBoost: Improving the Generalization Error of Gentle AdaBoost through a Margin Distribution. *IEICE Transactions on Information and Systems* E98-D, 11 (2015), 1906–1915.
[61] Chen Yan, Yan Long, Xiaoyu Ji, and Wenyuan Xu. 2019. The Catcher in the Field: A Fieldprint Based Spoofing Detection for Text-Independent Speaker Verification. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. 1215–1229.
[62] Bayya Yegnanarayana, A. Shahina, and M.R. Kesheorey. 2004. Throat microphone signal for speaker recognition. In *INTERSPEECH-2004, 8th International Conference on Spoken Language Processing (ICSLP)*.
[63] Lun Zhang, Rufeng Chu, Shiming Xiang, Shengcai Liao, and Stan Z Li. 2007. Face detection based on multi-block lbp representation. In *International Conference on Biometrics*. Springer, 11–18.
[64] Linghan Zhang, Sheng Tan, and Jie Yang. 2017. Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 57–71.
[65] Linghan Zhang, Sheng Tan, Jie Yang, and Yingying Chen. 2016. Voicelive: A phoneme localization based liveness detection for voice authentication on smartphones. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. 1080–1091.
[66] Zhengyou Zhang, Zicheng Liu, M. Sinclair, A. Acero, Li Deng, J. Droppo, Xuedong Huang, and Yanli Zheng. 2004. Multi-sensory microphones for robust speech detection, enhancement and recognition. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 3. iii–781.

A APPENDIX

A.1 CycleGAN

The total loss function for traditional CycleGAN can be defined as:

$$\mathcal{L}_{full} = \mathcal{L}_{adv}(G_{X\to Y}, D_Y) + \mathcal{L}_{adv}(G_{Y\to X}, D_X) + \lambda_{cyc}\mathcal{L}_{cyc}(G_{X\to Y}, G_{Y\to X}) + \lambda_{id}\mathcal{L}_{id}(G_{X\to Y}, G_{Y\to X}) \tag{14}$$

where λ_{cyc} and λ_{id} are tradeoff parameters for cycle-consistency loss and identity-mapping loss, respectively. In this model, two-step adversarial loss is adopted to mitigate the over-smoothing issue of the cycle-consistency loss, the adversarial loss on the circularly converted features is updated as:

$$\mathcal{L}_{\text{adv}}(G_{X\to Y}, G_{Y\to X}, D'_X) = \mathbb{E}_{x\sim P_X(x)}[logD'_X(x)] + \mathbb{E}_{x\sim P_X(x)}[log(1 - D'_X(G_{Y\to X}(G_{X\to Y}(x))))] \tag{15}$$

where D'_x is the additional discriminator.

A.2 GMM-UBM

GMM-UBM is one of the most classic approaches for speaker verification. It models the statistical characteristics of individual uniqueness without containing text related information. It is known that GMM usually requires lots of training data, but in the real world, the enrolled user's voice samples are always limited and insufficient to train the standard GMM. Therefore, GMM-UBM utilizes a UBM [45] (i.e., a special GMM type) which is pre-trained by many non-target users' data (i.e., also known as background knowledge) to build a prior speaker-independent distribution of features. Given the enrolled user's data, we only need to adapt the parameters in multiple Gaussian distributions in the GMM by a form of Bayesian adaptation (i.e., MAP: maximum a posteriori estimation). Given the pre-trained UBM and the training data from the enrolled user, $X = \{x_1, x_2, ...x_T\}$, we first obtain the probabilistic alignment between the enrolled vectors and the UBM mixture distributions as:

$$\Pr(i|x_t) = \frac{\omega_i p_i(x_t)}{\sum_{j=1}^{M} \omega_j p_j(x_t)} \tag{16}$$

where ω_i is the mixture weights that satisfy the constraint $\sum_{i=1}^{M} \omega_i = 1$. After the expectation step in the Expectation-Maximization (EM) algorithm, then the new parameters such as weight ω, mean μ, and variance σ^2 can to updated step by step based on $\Pr(i|x_t)$ and x_t.

However, one of the main difficulties of GMM-UBM systems involves intersession variability. Joint factor analysis (JFA) [24] was then proposed to compensate for this variability by separately modeling speaker variability and channel variability. Some works [12] later proved that JFA channel factors still contained speaker-related information. To provide a better channel robustness and lower computation cost, i-vector was proposed by combining the channel and speaker spaces into a total variability space $M = m + T\omega$, where m is the UBM Ω supervector, T is the rectangular matrix of low rank, and ω is the a standard normal distributed vector of size M. The i-vector of a given utterance can be estimated as:

$$\tilde{F}(u) = \sum_{t=1}^{L} P(c|y_t, \Omega)(y_t - m_c) \tag{17}$$

$$\omega = (I + T^t \sum^{-1} N(u)T)^{-1} T^t \sum^{-1} \tilde{F}(u) \tag{18}$$

where $\tilde{F}(u)$ is the first-order Baum-Welch statistics based on the UBM mean mixture components, m_c is the mean of UBM mixture component c, y_t is one frame of the given utterance. T is the total varability matrix. Afterwrd, intersession variability was then compensated for by using linear discriminant analysis (LDA) and within-class covariance normalization (WCCN). To this end, the cosine similarity score between the target user's utterance and test user's utterance can be calculated to verify the user's identity.