PiezoBud: A Piezo-Aided Secure Earbud with Practical Speaker Authentication

Gen Li[†], Huaili Zeng[†]
ligen4@msu.edu
zenghuai@msu.edu
Michigan State University

Hanging Guo
guohanqi@hawaii.edu
University of Hawaii at Manoa

Yidong Ren
renyidon@msu.edu
Michigan State University

Aiden Dixon
dixonaid@msu.edu
Michigan State University

Zhichao Cao
caozc@msu.edu
Michigan State University

Tianxing Li
litianx2@msu.edu
Michigan State University

ABSTRACT

With the advancement of AI-powered personal voice assistants, speaker authentication via earbuds has become increasingly vital, serving as a critical interface between users and mobile devices. However, existing audio-based speaker authentication methods fail to defend against voice spoofing threats such as replay and deepfake attacks. To counteract these risks, we introduce PiezoBud, a pioneering multi-modal user authentication system that is truly practical and lightweight for earbuds. PiezoBud uses miniature piezoelectric sensors to detect micro-vibrations on the skin, extracting user-specific biometric data to authenticate legitimate access on the local smartphone and protect against malicious attacks. Our exploratory study, involving 85 participants, demonstrates the effectiveness of PiezoBud in various everyday scenarios, including ambient noise, body movement, and in-ear media playing. Using only **15 seconds** of enrollment data, PiezoBud achieves an Equal Error Rate (EER) of **1.05%** and attain a mean authentication latency of **0.06 seconds** on mobile devices. We also evaluate PiezoBud's effectiveness in countering challenging adaptive attack scenarios and its overall performance in various real-world situations. Our evaluation highlights that PiezoBud stands out as a **practical**, **resilient**, **responsive**, and **secure** option for earbuds users.

CCS CONCEPTS

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

KEYWORDS

Multi-modality, Piezoelectric, Earbuds, User Authentication

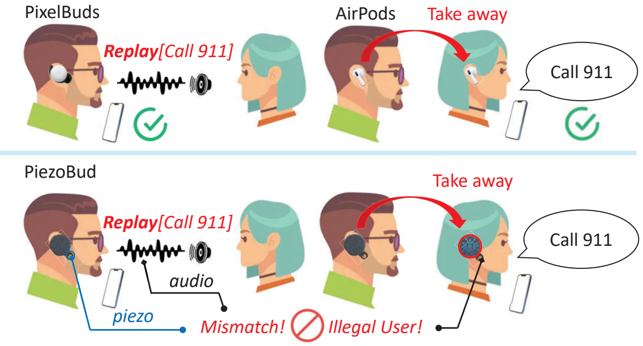


Figure 1: Attackers can compromise existing VA systems on smartphones via COTS earbuds like Google PixelBuds and Apple AirPods through replay/mimic attack. PiezoBud eliminates such threats by adding a tiny piezo sensor on earbuds, capturing the skin vibration signals that are hard to be replayed or mimicked and feasible to establish bio-matching with audio signals to authenticate legal users.

ACM Reference Format:
Gen Li[†], Huaili Zeng[†], Hanging Guo, Yidong Ren, Aiden Dixon, Zhichao Cao, and Tianxing Li. 2024. PiezoBud: A Piezo-Aided Secure Earbud with Practical Speaker Authentication . In *The 22nd ACM Conference on Embedded Networked Sensor Systems (SENSYS '24), November 4–7, 2024, Hangzhou, China.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3666025.3699358

1 INTRODUCTION

Speaker authentication via earbuds has become increasingly critical. With a booming market share [10], earbuds now become a crucial platform for voice assistant (VA) interactions on mobile devices, offering enhanced privacy protection and portability for users in their daily lives. Furthermore, the proliferation of AI-based VAs is encouraging users to manage private matters through these devices [11, 12]. However, existing audio-based user authentication solutions [13, 14] fail to defend against prevalent threats such as replay attacks [15], side-channel attacks [14], and mimic attacks [16]. By compromising victims' earbuds, attackers can easily circumvent robust authentication methods on mobile devices, thereby accessing sensitive data stored on them. Although personalized keyword recognition by training VA with authorized speakers' sentences [17] is widely used to extract unique voiceprint-based bio-features to counter potential threats, sole voiceprint-based authentication or liveness detection on earbuds fails to defend against advanced attacks: As shown at the top of Figure 1, existing earbuds like PixelBuds [18] fail to defend against replay attacks [19] and deepfake threats [20]. Others, such as AirPods [21], only detect

[†]Both authors contributed equally to this research.

SENSYS '24, November 4–7, 2024, Hangzhou, China

ACM ISBN 979-8-4007-0697-4/24/11.
https://doi.org/10.1145/3666025.3699358

	Modalit(ies)	Defense Reliability				User-Friendly	
		# subjects	ACC (%)	FAR (%)	EER (%)	Enrollment length (s)	Latency (s)
[1]	In/out ear sound	23	-	0	< 4	75	0.389 − 0.484
[2, 3]	Ear canal	20 − 24	95.16 − 97.38	0.18 − 5.3	-	- / 120	-
[4]	PPG + Audio	25	94.84	-	-	-	-
[5–7]	ACC/IMU	18 − 41	95 − 97	< 7	-/1.28	-	0.3 − 2
[8, 9]	Piezo (41 mm ϕ) + Audio	8 / 29	96 / 97	- / 3.6	-	N.A. / 107	- / 4.53
PiezoBud	Piezo (**10** mm ϕ) + Audio	**85**	**99.21**	**0**	**1.05**	**15**	**0.041 − 0.094**

Table 1: Comparison with other state-of-the-art authentication methods.

whether the device is worn, ignoring user identity [14]. Thus, attackers can easily compromise the security of mobile devices by attacking the victim's earbuds. For example, recent studies like EchoAttack [14] have demonstrated that attackers can easily deceive existing commercial off-the-shelf (COTS) voiceprint-based earbuds using a single ultrasound speaker to activate private voice assistants on smartphones. These insecure systems can result in significant privacy breaches and economic losses [22–26].

Although earbuds are susceptible to attacks, they offer manufacturers the opportunity to incorporate various modalities beyond voice to enhance security. Designing a practical earbuds system with reliable multi-modal speaker authentication requires software-hardware co-design that meets the following criteria:
i) **Modality Resilience**: Sensors and their auxiliary hardware components must be compact enough for integration into COTS earbuds, and the cost of the selected sensor should also be minimized. Therefore, the newly integrated modality must still deliver consistent optimal signal quality with a high signal-to-noise ratio (SNR).
ii) **Defense Reliability**: Introducing an additional modality should enhance the security robustness of the earbuds. The new authentication system must perform well in various daily user scenarios across a large user group and remain immune to advanced attacks, even if stolen. Additionally, the system should be text-independent, meaning the earbuds must authenticate every speech they perceive, not just specific activation keywords (e.g., Hey Siri).
iii) **User-Friendly**: The authentication system should be user-friendly, with negligible latency to remain unobtrusive. The enrollment should be quick and require minimal processing time. The authentication process should not negatively impact user experience and health. Both enrollment and authentication are expected to be performed locally on mobile devices. Compared to cloud-based solutions, local processing enhances privacy by keeping sensitive biometric data on the device. It also reduces latency from device-to-cloud communication and ensures the authentication system remains accessible without an Internet connection. Furthermore, efficient battery life and a comfortable fit should also be optimized.

Nevertheless, existing solutions fail to meet all the aforementioned requirements simultaneously. As shown in Table 1, firstly, [1] verifies user identity by comparing in-ear and out-ear sounds. However, the long data required for enrollment (75 seconds) negatively impacts user experience. Secondly, some methods [2, 3] enhance security using the unique structure of the human ear, but playing ultrasound into the ear may risk hearing damage [27] and requires the speaker to remain inactive, reducing earbud functionality [28]. Moreover, its false accept rate (FAR) can reach to as high as 5.3%, leaving the space to launch a mimic attack. Thirdly, although combining a Photoplethysmography (PPG) sensor with audio [4] has also been explored, it fails to provide accurate authentication (<95% accuracy). Fourthly, some studies have used accelerometer (ACC)/inertial measurement units (IMU) [5–7], or piezoelectric sensors [8, 9] to capture body vibration or non-audible murmur (NAM) signals to enhance the security. Nevertheless, the low sampling rate of IMUs results in information loss (Section 2.2). Thus, users must speak fixed long sentences to collect sufficient information, hindering efficient text-independent authentication [6] and increasing latency (up to 2 seconds). Additionally, these piezoelectric-aided methods offer poor authentication (up to 3.6% FAR), require long enrollment (up to 107 seconds), and have high latency (up to 4.53 seconds). Furthermore, large piezoelectric sensors with cumbersome hardware are demanded to boost SNR, which is difficult to integrate into COTS earbuds [8, 9]. Finally, since the lack of sufficient subjects (<35) for evaluation, whether these methods could work in a large group of users is still questionable.

In this paper, we propose PiezoBud, a practical piezo-audio earbuds solution with a carefully co-designed hardware-software approach to meet end-to-end requirements, addressing both user experience and security concerns. As shown at the bottom of Figure 1, PiezoBud surpasses its predecessors by overcoming all the aforementioned drawbacks. PiezoBud enables user authentication when paired with a mobile device, such as a smartphone. By leveraging the newly introduced piezoelectric modality, PiezoBud allows users to enroll and authenticate themselves on the local smartphones with an enhanced level of security. PiezoBud excels in *real-time*, *text-independent*, *local*, and *highly secure* user authentication, functioning as *COTS earbuds without burdening the user experience*.

However, several challenges must be addressed for efficient PiezoBud deployment in complex environments:
i) **Embedding Piezoelectric Sensors in Earbuds**: Minimal sensors are desirable for integrating into small earbuds. However, their sizes significantly influence the signal strength (Section 2.3). Thus, the NAM piezoelectric signal will be much weaker than the voice audio signal and might be overshadowed by irrelevant noise. It is challenging to enhance piezoelectric signal strength while keeping sensor size small.
ii) **Integrating Two Different Modalities for Higher Security**: Unlike existing piezo-audio methods [8, 9], PiezoBud enhances security by leveraging the inherent bio-feature mapping between

voice audio signals and NAM piezoelectric signals, offering more user-specific information without the need for additional security measures. However, effective piezo-audio integration is challenging as audio data often contains more information than piezoelectric data. Relying on audio skews the balance, and fusion network [29–31] or simple concatenation does not enhance security.

iii) **Achieving Real-Time Authentication on Mobile Devices with Limited Resources**: To optimize user experience, we need to shorten the latency of model computation, especially executing it on mobile devices with a limited energy/computational budget. This exposes the challenge of requiring a lightweight model while reserving a high-security level.

To address these challenges, firstly, we design a **low-cost hardware platform** that seamlessly incorporates a miniature piezoelectric sensor into earbuds, ensuring the user experience and signal quality remain unaffected with a unique cavity design and amplification circuit. Secondly, a **novel authentication network** consists of two carefully designed pipelines to project the distributions of two modalities into a latent space. By doing so, we effectively fuse features between the modalities in a balanced manner to guarantee security. Moreover, a **signal processing module** pre-validates input audio and piezoelectric data, and a **compact universal feature extractor** derives high-dimensional features from both modalities. Together, they reduce training overhead, enabling efficient real-time performance on resource-constrained devices.

We implement PiezoBud using COTS hardware components. Table 1 compares PiezoBud with state-of-the-art (SOTA) methods. PiezoBud demonstrates superior performance in several key metrics: they provide long battery life (10 continuous hours with CR2032 coin cell), achieve the highest identification accuracy of 99.21%, maintain the lowest EER (1.05%) with shortest enrollment time (15 s), ensure 100% defense against spoofing/mimic attacks, and support real-time (< 94 ms) text-independent authentication on smartphone locally with low cost (~$ 30). The main contributions are summarized below:

- We developed PiezoBud, the first earbud with a miniature piezoelectric sensor, as an open source [1] and low cost platform for practical speaker authentication. Superior performance (1.05% EER with 15 seconds of enrollment data) and negligible latency (41-94 ms on six tested phones) enable PiezoBud to offer mobile-friendly, real-time, text independent, and local authentication with robust protection against adaptive attacks.
- We proposed FusionSecNet, a cascaded neural network to meticulously extract high-dimensional features across dual modalities. This framework handles two modalities in parallel, skillfully fusing them to produce a robust, personalized embedding vector. This vector, resistant to various attacks, enhances the system's user authentication capabilities.
- We conducted a comprehensive study on human voice and surface vibration, assembling a dataset for speaker verification — the first of its kind to our knowledge. This study involved 85 participants, from whom we recorded 516 minutes of audio using both a microphone and a piezoelectric sensor.

2 FEASIBILITY STUDY

Our feasibility study validates the proposed concept and potential of utilizing piezoelectric sensors to secure earbuds.

2.1 Can we use the NAM signal to identify?

Human Phonation Principles The human vocalization system consists of three key components, each fulfilling a unique but interrelated function in speech production [32]. Figure 2(a) illustrates the process starting with lungs that generate necessary airflow. The airflow vibrates the vocal folds and creates the initial raw sound. The mouth, tongue, lips, and other articulators refine this sound, acting as filters to shape the voice and produce a diverse speech spectrum. Additionally, tissue vibrations generate NAM signals on the skin [1, 33, 34], which piezoelectric sensors can capture.

Biometric Analysis of Piezoelectric and Audio To verify individual biometric traits in NAM signals, we attach a piezoelectric sensor and a microphone to the same place near the right ear on the face. We then evaluate the Cross-Power Spectral Densities (CPSD) between the audio and piezoelectric data using the transfer function $H(f) = P_p(f)/P_a(f)$, where $P_p(f)$ and $P_a(f)$ represent the power spectral densities of the piezoelectric and audio data, respectively. These densities are calculated using Welch's method [35]. Results for three subjects are displayed using Principal Component Analysis [36] (PCA), with the first two components in Figure 2(b). The clustering of points indicates a consistent relationship between piezoelectric and audio data within each subject and variations among different ones.

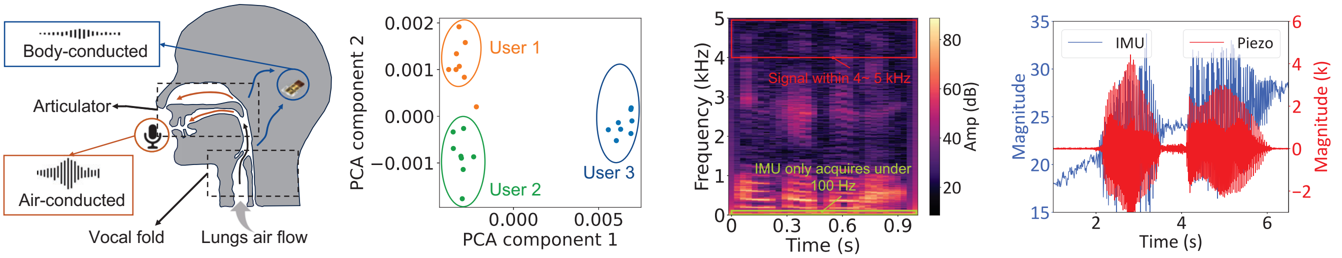
2.2 Why choose piezoelectric over IMU?

Existing studies primarily utilized IMU sensors as the additional or sole modality for voiceprint-based user authentication, as these sensors are already integrated into some earbuds [5, 6, 37]. This raises a critical question: why choose piezoelectric sensors over IMU sensors? To address this, we conducted an experiment comparing the performance of both sensor types. Three participants read an article for one minute, with a piezoelectric sensor [38] placed in front of each subject's ear and an IMU sensor [39] attached next to it. Both sensors recorded the NAM signals while participants spoke. The piezoelectric sensor was sampled at 10 kHz, 100 times higher than typical IMU sensors [40]. The NAM signals captured by the piezoelectric sensor were converted into a Short-Time Fourier Transform (STFT) spectrum after applying a 100 Hz high-pass filter (HPF). As shown in Figure 2(c), the piezoelectric sensor captures information up to 5 kHz, while the IMU sensor loses high-frequency components. Besides, as shown in Figure 2(d), IMU data drift also introduces errors. Although methods [41] exist to mitigate this issue, they can't recover the information in high frequencies.

2.3 What is the impact of sensor size?

To assess if the piezoelectric sensor size affects PiezoBud's performance, we compared different sensor sizes in an experiment. We used three piezoelectric sensors differing only in dimensions: 20 × 10 mm [38], 30 × 15 mm [42], and 30 × 30 mm [43]. Each sensor was placed on a speaker, emitting a 1000 Hz tone at a consistent volume. We recorded ten seconds of piezoelectric signals for each

[1] https://github.com/HuailiZ/PiezoBud



(a) The principles of human phonation.

(b) Users' transfer function depicted using first two components of PCA.

(c) The highest frequency of piezoelectric data can reach 5 kHz.

(d) IMU sensor captured NAM signal suffers from drift errors.

Figure 2: Feasibility Study.

trial, repeating ten times. The average signal strengths were 60.24 dB, 71.5 dB, and 79.67 dB, respectively. These findings show that piezoelectric sensor signal strength increases with size. Hence, the raw signal must be enhanced to integrate miniature sensors into earbuds without compromising signal quality.

3 THREAT MODEL

Beyond the ability to authenticate different users, PiezoBud should be resilient to various attacks. We outline the attacker capabilities and threat model for PiezoBud. The adversary aims to trick the authentication system into accessing sensitive services on mobile devices. We assume that attackers might gain possession of PiezoBud, but not the user's authentic piezoelectric data. Because this would require the attacker to physically attach sensors to the user's facial area, which is highly improbable. This paper examines four attack scenarios: targeting audio data alone (Scenario 1) and compromising both audio and piezoelectric modalities (Scenarios 2, 3, and 4).

Scenario 1. Attackers only launch audio replay/mimic attacks. In this scenario, we assume the attackers lack access to the user device and are unaware of the piezoelectric modality, targeting audio-only, replay attacks [14, 19]. Under this assumption, we consider two cases: i) **Replay:** the attackers use previously recorded audio of the victim, and ii) **Mimic:** the attackers synthesize the victim's voice using advanced voice generation technologies [44–46]. The victim is either silent or talking during the attacks. If the victim is silent, we assume the piezoelectric sensor can only pick up random noise. If the victim is talking, the attackers may use a loudspeaker to overwhelm the victim's audio following the criteria in [14] with an ultrasound speaker, ensuring that victims will not notice them.

Scenario 2. Attackers falsify piezoelectric data to mimic victims'. In this scenario, we assume attackers obtain the victim's PiezoBud. Aware of the piezoelectric protection, they attempt to mimic the victim's signals. Attackers might attempt several methods to deceive PiezoBud: i) generating voltage glitches to mimic the piezoelectric data, and ii) wearing PiezoBud themselves to use their own piezoelectric data in an effort to imitate the victim. In such cases, the audio modality is the victim's, while the attackers could use their own piezoelectric signals or introduce unrelated voltage changes to deceive PiezoBud.

Scenario 3. Attackers use recorded audio data as input via different media. Understanding that both modalities originate from the same source, attackers may attempt to deceive the system with a synthetic human skull model. Attackers could attempt to place a speaker near the throat area of a skull model while positioning PiezoBud on the ear side of the model, trying to replicate the piezoelectric data through the speaker and skull model. This fools the system into identifying this setup as the legitimate user.

Scenario 4. Attackers train a voice conversion network to obtain synthetic NAM signals from compromised audio. In a less likely scenario, attackers could develop a voice conversion network [47] to convert audio modality into piezoelectric modality, attempting to uncover the unique bio-information tied to a user in both audio and piezoelectric data. However, given the limited availability of piezoelectric data and the substantial computational demands of such a network, it is presumed that attackers would only have access to a constrained dataset, and the network would face significant limitations in complexity and size.

4 SYSTEM OVERVIEW

As shown in Figure 3, PiezoBud consists of two primary components: a custom-designed hardware platform and an authentication framework FusionSecNet.

PiezoBud Hardware: We design a custom hardware prototype consisting of a pair of two printed circuit boards (PCBs) and a shell that combines a small piezoelectric sensor with the sensing boards. The cavity, designed to house the piezoelectric sensor, improves the quality of collected piezoelectric data. Furthermore, a specially designed amplification circuit significantly enhances the raw piezoelectric signal.

FusionSecNet Modeling: FusionSecNet is optimized for efficient use on mobile devices such as smartphones. It includes a feature extractor SynthEx and an authentication model flowAuth. After pre-validation, PiezoBud employs the compact universal SynthEx to extract high-dimensional features using cascaded scaling Res2Blocks [48]. These features capture user-specific information from each modality across multiple scales. Those features from both modalities are then processed by flowAuth, a blockwise-paired network that maps each user's distribution into a latent space combining both modalities. The final output of flowAuth is individual-distinctive and attack-resistant, giving subsequent authentication.

Overall Progress: To authenticate a new user, PiezoBud collects enrollment data from both audio and piezoelectric sensors. This data is processed locally by FusionSecNet to generate user-specific

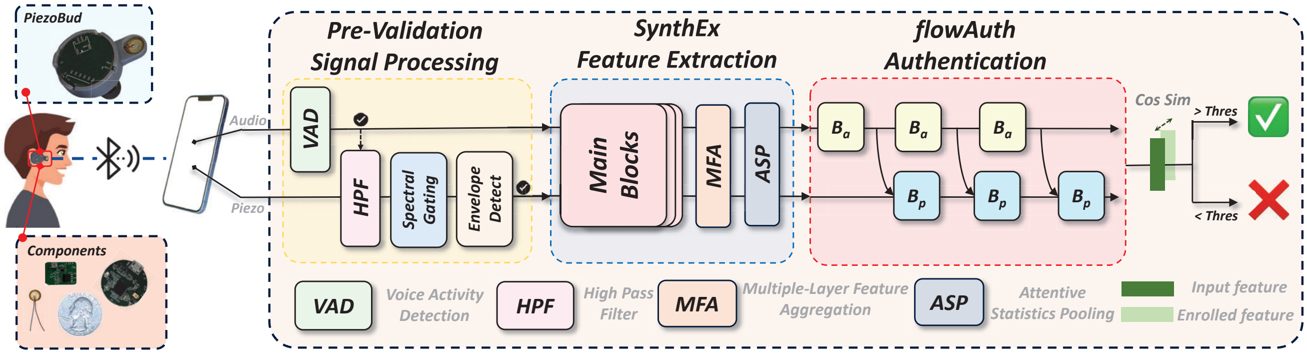


Figure 3: PiezoBud Overview: The system involves three main procedures: signal pre-validation, feature extraction, and authentication. SynthEx derives high-level features, and flowAuth combine two modalities for authentication. PiezoBud further computes the cosine similarity between the embedding vectors of the input data and the enrolled data. If the similarity exceeds the threshold, PiezoBud accepts the input; otherwise, it is rejected.

embedding vectors on the device. During regular use, PiezoBud continuously captures audio and piezoelectric data while the user is speaking, feeding it into FusionSecNet to produce verification embedding vectors. Authentication proceeds with PiezoBud performing binary classification based on the cosine similarity between these embedding vectors. The process concludes with a binary output, where a label of ***True*** signifies successful user authentication.

5 PIEZOBUD HARDWARE DESIGN

As shown in Figure 4, our hardware setup features a pulse-density modulated (PDM) microphone, a miniature piezoelectric sensor, and a Bluetooth Low-Energy (BLE) chip for the microcontroller (MCU). It is powered by a coin cell battery or MicroUSB connector, with programming via Serial Wire Debug (SWD). As the piezoelectric sensor's SNR declines with its size decreasing, we aim to develop a dedicated amplification circuit to improve signal quality before processing. The right side of Figure 4 highlights key components: a differential amplification module to amplify raw piezoelectric data, an impedance matching module to reduce signal reflection, a high-pass filter to eliminate baseline voltage, and a second amplification module for further enhancement. The MCU captures the enhanced data via the differential successive approximation register (SAR) analog-to-digital converter (SAADC) channel. After capturing both audio and piezoelectric data, they are transmitted to the end device via Bluetooth.

Additionally, we create a shell to enhance the quality of data collected by the piezoelectric sensor. The housing structure is optimized for better SNR of piezoelectric data. As shown in Figure 5(b), the sensor is centrally positioned with distinct materials layered above and below it. The material above reflects most ambient noises, while the material below conducts most NAM signals. We assessed the sensor's performance with various materials placed above and below it. For each material combination, we conducted two experiments: i) placing the material between the sensor and the skin (inwards) to assess NAM signal conduction, and ii) positioning it above the sensor (outwards) to evaluate ambient noise reflection. In the inwards test, higher amplitudes indicate better signal conduction, while in the outwards test, lower amplitudes signify better noise reflection.

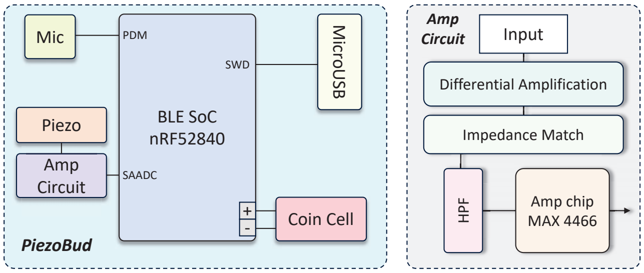


Figure 4: Hardware block diagram: PiezoBud includes a PDM microphone, a piezoelectric sensor with a custom amplifier, a coin cell battery, and a MicroUSB port.

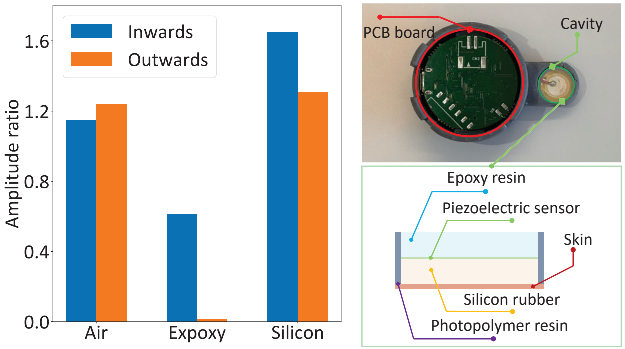
For the inward tests, a volunteer consistently pronounces the vowel *o*. For the outward tests, another volunteer stands in front of the sensor and repeats the same sound at the same volume. Figure 5(a) shows the average amplitude ratios from each experiment, compared to the baseline where the piezoelectric sensor is attached directly to the skin without any material. We found that silicone significantly amplifies the inward sounds (target signal) due to its effective transmission of NAM signals, while epoxy suppresses most ambient noise. Based on these findings, we engineered the cavity to fit the piezoelectric sensor, as shown in Figure 5(b). The sensor is positioned between silicone rubber (below) and epoxy resin (above). The significant density differences between the epoxy resin and air reduce the ambient noise picked by the piezoelectric sensor. This cavity design enhances PiezoBud's defense against spoofing/mimic attacks and improves the quality of the collected data.

Figure 6 illustrates the daily use of PiezoBud and its size compared to a quarter. The compact design ensures a comfortable fit and user experience while maintaining piezoelectric data quality.

6 MULTI-MODAL FUSIONSECNET DESIGN

6.1 Pre-Validation Data Processing

Feeding raw collected data directly into the next procedure is impractical due to high computational and power demands. To conserve battery and improve signal quality, we perform pre-validation



(a) Materials affect the piezoelectric sensor's performance.

(b) We place epoxy on the top of sensor and silicon below it.

Figure 5: The cavity layered with various materials helps reduce outside noise and enhance signal quality.

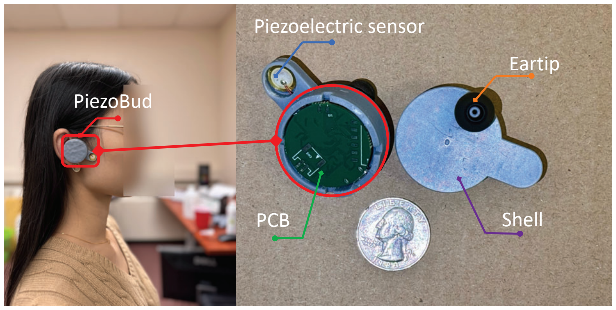


Figure 6: PiezoBud integrates a piezoelectric sensor into earbuds without compromising user comfort.

to prevent simple attacks, like audio replay. Specifically, we segment active speech audio clips and verify their alignment with corresponding piezoelectric data clips.

We use the Voice Activity Detection (VAD) algorithm from WebRTC [49] to isolate valid human speech clips, allowing PiezoBud to activate only in response to detected speech. First, we apply a HPF to remove low-frequency components related to facial muscle movements and human activities [5]. We then reduce steady-state noise (e.g., hiss or hum) using spectral gating [50]. Next, we derive the envelope of the piezoelectric data by applying the Hilbert transform [51]. Finally, we determine the range of the normalized envelope. If the envelope range for the piezoelectric data falls below an empirically configured threshold, it indicates an absence of voice activity, meaning piezoelectric input is considered to be either silent or inactive in terms of vocalization. The pre-validation scheme assesses the alignment between the audio and piezoelectric modalities, ensuring that PiezoBud proceeds only when both inputs are validated. Additionally, it enhances the quality of the validated piezoelectric data. Detailed parameter settings are provided in Section 7.2.

6.2 Feature Extractor SynthEx Modeling

After processing the raw signals, we aim to extract high-dimensional features from both modalities instead of simply using the raw data. The reason is that using the data directly, which contains surplus information, would increase model parameters, complicate training, and risk system overload. We developed SynthEx as a universal compact feature extractor for both modalities, producing high-dimensional, concise features that retain modality-specific information. The high sampling rate of the piezoelectric sensor enables using the same model structure for both piezoelectric data and audio.

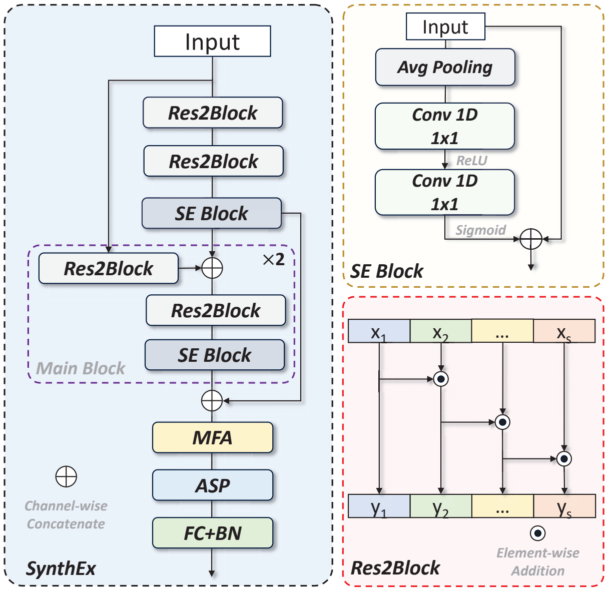


Figure 7: SynthEx Overview: It contains three main blocks, each with 2 Res2Blocks and 1 SE block. Outputs are concatenated and processed through MFA and ASP, followed by an FC and BN layer. In the Res2Block with scale s, the input data x is divided into s equal-length segments. During the forwarding, the segments are residually added. Then, all the segments are concatenated to generate the final output y.

As shown in Figure 7, SynthEx contains three sequential main blocks. Each consists of two Res2Blocks [52] and one SE block [53]. These Res2Blocks and SE blocks enable the model to process features at different scales, enhancing the representativeness of the extracted features and improving model efficiency. We apply a different scale s for each main block, where s refers to dividing the input feature map into s subbands along the channel dimension. Larger scales divide the feature map into smaller channel bands, allowing the block to focus on more detailed features. Conversely, smaller scales direct the main block to concentrate on more general features across the channels and reduce the number of training parameters [54]. However, reducing training parameters can lead to potential performance degradation. To address this, we added an extra Res2Block to each main block, forming a double Res2Blocks chain. This structure preserves the original spectrum and helps counteract performance decline. The outputs from different main blocks are combined residually, and features are extracted using multiple-layer feature aggregation (MFA) and attentive statistic pooling (ASP) [52], followed by a fully connected (FC) layer and a batch normalization (BN) layer. By incorporating cascaded Res2Blocks, SynthEx effectively extracts user-specific information across multiple scales, from small to large. This unique structure comprehensively distills vital features, countering performance degradation without significantly increasing the number of parameters, thus maintaining simplicity.

6.3 Multi-Modality Authentication flowAuth Modeling

We aim to integrate distinct features from audio and piezoelectric modalities for user authentication, surpassing simple concatenation, as mere concatenation cannot defend against complex attacks. Drawing on the insight that the characteristics of audio and piezoelectric data each follow unique distribution patterns, we observe that different modalities like audio and piezoelectric inherently exhibit distinct features within their respective distributions. As stated in Section 2.1, our findings indicate that the relationship between audio and piezoelectric modalities remains consistent for the same individual. Thus, we realize the object function as follows:

$$z = g_\theta(p|a) \tag{1}$$

In this context, $g_\theta(\cdot)$ is the function defined by the model, and z symbolizes the distribution of piezoelectric data p conditioned on audio a within a specific latent space. Ideally, z should be able to integrate the biometric information of those two different modalities. Based on this, we developed flowAuth to meet our unique requirements. The structure of flowAuth is depicted in Figure 8. It accepts embedding vectors of audio and piezoelectric data as E_a and E_p, then produces a user-specific embedding vector, denoted as E_z. As shown in the left bottom part of Figure 8, the block contains two flows [55], and each flow contains three layers: actnorm, invertible 1×1 conv, and affine coupling layers. Each non-final block splits its output: one half progresses as the *hidden* output to the subsequent block, and the other half as the block's *open* output. flowAuth's unique block pairwise alignment design balances the information-rich audio data with the less information-rich piezoelectric data, effectively preventing a potential skew towards the audio data. Additionally, by assigning different modalities to separate input pipelines, flowAuth projects the distribution of hidden user-specific biometric features into a latent space. The process is outlined as follows: i) Each modality is processed using a separate pipeline. These two pipelines have the same number of blocks and flows but operate with distinct parameters. We refer to them as P_A and P_P for the pipelines of audio and piezoelectric modalities, respectively. ii) During the training progress, we apply block-to-block alignment between P_A and P_P, and the output of blocks at the same levels in these pipelines can be articulated as follows:

$$\begin{aligned} \left[o_a^{i+1}, h_a^{i+1}\right] &= B_a^{i+1}\left(h_a^i\right) \\ \left[o_p^{i+1}, h_p^{i+1}\right] &= B_p^{i+1}\left(h_p^i, o_a^i\right) \end{aligned} \tag{2}$$

where a and p represent the audio and piezoelectric modalities respectively, and i denotes the level of block. In this notation, o and h correspond to the *open* and *hidden* outputs of the block. The *hidden* and *open* outputs work together to enable efficient and accurate modeling of complex multi-modal data through distribution projection. *Hidden* outputs capture essential features, while *open* outputs ensure the results are expressive and observable. The *hidden* output of i-th block in P_A is fed into the block B_a^{i+1}, its *open* output o_a^i is concatenated together with *hidden* output h_p^i of i-th block in P_P. Then, the concatenated output is fed into the block B_p^{i+1}. Finally, the open outputs of each block in P_P are reshaped and concatenated as the final output of E_z. In our design, the flowAuth model consists of three blocks, each containing two flows. These settings were chosen to achieve optimal results with the smallest possible configuration.

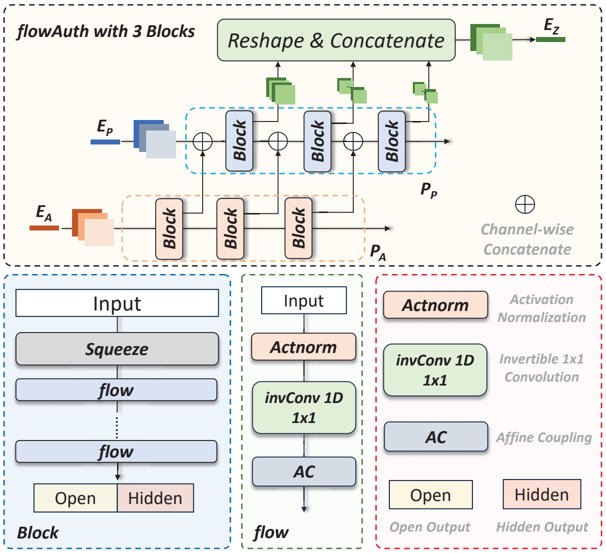


Figure 8: flowAuth Overview: it take embedding vectors E_p (piezoelectric) as input and E_a (audio) as condition. Each block (mid bottom) contains k flows. Each flow (right bottom) contains three essential layers.

6.4 Loss Function

FusionSecNet is expected to extract user-specific features and fuse multi-modal features with enhanced security. It is important to note that while PiezoBud is designed for authentication, multiple users are selected in each training batch. The model's objective is to maximize the distance between different users' embedding vectors while minimizing the distance within the same user's embedding vectors [52, 54, 56]. The detailed designs for the loss functions guiding the SynthEx and flowAuth are outlined below.

Feature Extractor SynthEx is designed to extract user-specific high-level features. It ensures that features extracted from a single modality of one user exhibit high similarity. To achieve this, we use Generalized End-to-End (GE2E) loss, which streamlines and enhances the training process for speaker verification systems [56]. GE2E loss is defined as:

$$\mathcal{L}_{\text{GE2E}} = \frac{1}{S \times U} \sum_{j=1}^{S} \sum_{i=1}^{U} \left[-\mathrm{S}_{ji,j} + \log \sum_{k=1}^{S} \exp(\mathrm{S}_{ji,k}) \right] \tag{3}$$

where S is the number of speakers, U is the number of utterances for each speaker. $\mathcal{S}_{ji,k}$ is the similarity between i-th utterance of speaker j and centroid of speaker k:

$$\mathrm{S}_{ji,k} = \begin{cases} w \cdot \cos(\mathbf{e}_{ji}, \mathbf{c}_j^{-i}) + b & \text{if } k = j \\ w \cdot \cos(\mathbf{e}_{ji}, \mathbf{c}_k) + b & \text{otherwise} \end{cases} \tag{4}$$

$\mathbf{c}_j^{-i}$ is speaker j's centroid excluding utterance i and $\mathbf{c}_k$ is speaker k's centroid. w and b are learnable parameters that scale and shift the similarity scores, respectively. By minimizing this loss, SynthEx learns to generate embedding vectors close to the same speaker and distant for different speakers. For the training of SynthEx, we apply the GE2E loss to two different modalities separately:

$$\mathcal{L}_{\text{SynthEx}} = \mathcal{L}_{\text{GE2E}}(E_a) + \mathcal{L}_{\text{GE2E}}(E_p) \tag{5}$$

Authentication Model The goal for flowAuth is to construct the integrated embedding vector E_z^v using the embedding vector of piezoelectric data E_a^v as the condition and the embedding vector of audio E_p^v as the input. To defend against various attack scenarios, we incorporated falsified inputs during training, including i) treating invalid piezoelectric modality input as white Gaussian noise (WGN), ii) using audio data for both audio and piezoelectric modalities, and iii) introducing a temporal mismatch between piezoelectric and audio data. These inputs are then concatenated with genuine E_z at the user dimension to compute the GE2E loss:

$$\mathcal{L}_{\text{flowAuth}} = \mathcal{L}_{\text{GE2E}}\left(\left[E_z^{(a,p)}, E_z^{(a,wgn)}, E_z^{(a,a)}, E_z^{\text{mis}(a,p)}\right]\right) \tag{6}$$

where wgn denotes WGN, a is the audio data, and p indicates the piezoelectric data. $E_z^{(a,p)}$ represents the authentic embedding when audio and piezoelectric data are aligned, and $E_z^{\text{mis}(a,p)}$ corresponds to the scenario where a and p are mismatched. This concatenation enhances flowAuth's robustness against multiple malicious attacks, enabling it to effectively distinguish legitimate input features from malicious ones. The overall objective function is as follows:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{SynthEx}} + \beta \cdot \mathcal{L}_{\text{flowAuth}} \tag{7}$$

where α and β are pre-defined hyper parameters to change to focus of FusionSecNet on different part.

7 IMPLEMENTATION

7.1 Hardware Prototype

The PiezoBud prototype consists of three components: an audio sampling module, a piezoelectric sampling module, and an earbud shell. We use a miniature piezoelectric sensor (PUI AB107B-LW100-R [57]), which interfaces with an amplification PCB board including LM358P [58] and MAX4466 [59] amplifiers[2]. This amplification PCB connects to the main PCB, which integrates a PDM microphone (MP23DB01HPTR) [60] and a BLE microcontroller (nRF52840) [61].

The amplification circuit was moved to a separate PCB to fit the earbuds. Figure 9 shows a size comparison between our custom-designed PCBs, the piezoelectric sensor, and a quarter coin. The left circular PCB (28 × 28 mm) is the main component, while the right rectangular PCB (15 × 20 mm) is the amplification circuit. The MAX4466-based module's amplification circuit has a trimmer pot to adjust the gain from 25× to 125× for research. The differential amplification module's gain is set to 200×, and the MAX4466-based module maintains a fixed 25× gain throughout the experiment. We developed PiezoBud's hardware schematic and layout using EasyEDA [62–64] and fabricated two 2-layer PCB boards through JLCPCB. The 3D-printed enclosures were designed with AutoDesk Fusion 360 and produced on a Creality Resin 3D Printer Halot-Mage. Table 2 provides the prototype's cost breakdown. PiezoBud serves as an initial prototype for performance assessment, with future versions potentially using cheaper components for large-scale production.

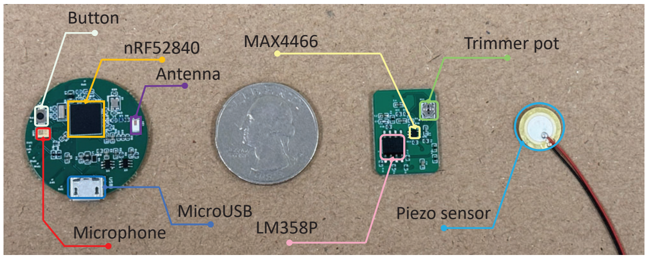


Figure 9: PiezoBud PCBs with miniature piezoelectric sensor when placed beside a quarter coin.

7.2 Software Implementation

We implement FusionSecNet using PyTorch with a learning rate of 1e-3 and a scheduler that decays at 0.97 every 100 epochs. We optimize using Adam with a weight decay of 2e-5. To enhance FusionSecNet's defensive capabilities, we set α to 0.3 and β to 0.7. We ensure thorough learning and convergence by training for 2,000 epochs.

Pre-Validation Signal Processing The cut-off frequency of the HPF is set to 100 Hz, and a threshold of 0.45 is determined through empirical analysis. For spectral gating noise reduction, we set the noise reduction ratio to 0.97. We configure the detection level to 2 and use a 30 ms frame length for VAD.

Feature Extraction Model To enhance SynthEx's verification performance, we pre-train it using Voxceleb1 [65], which features 1,211 speakers for training and 40 for testing. The pre-training process includes 153,000 utterances processed with a 25 ms Hanning window and a hop length of 10 ms. Two-second segments are extracted from each utterance. During this process, we utilize AAM-SoftMax loss [66] with a loss margin of 0.2 and a loss scale of 30.

Authentication Model To train flowAuth, we collect a dataset from 85 individuals (Section 8.1). For each subject, we allocate 80% of their utterances for training and reserve the remaining 20% for testing. This widely adopted strategy ensures ample training data while keeping the test data unseen during training [67]. Each training epoch consists of batches containing 20 subjects. For each subject, ten utterances (each lasting 500 ms) are selected to produce an 80-band Mel Spectrum using the same Hann window and hop length settings as in SynthEx.

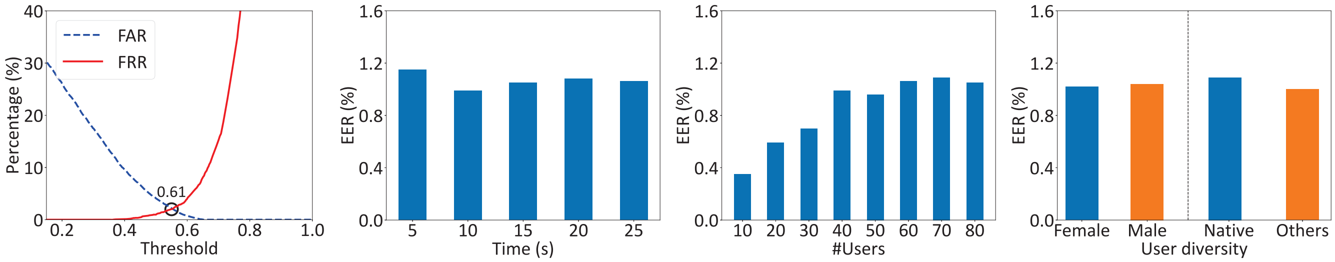
8 SYSTEM EVALUATION

8.1 Experimental Setup

Participants To evaluate the FusionSecNet authentication model in Section 6, we recruited 85 subjects [3] (21 females, 64 males) aged

[2] An impedance mismatch exists between the amplifier and the piezoelectric sensor. Our prototype serves as preliminary validation, using a common COTS amplifier with acceptable signal strength in tests. For better performance, an impedance-matched amplifier is preferred.

[3] The study received Institutional Review Board (IRB) approval from our institute.



(a) EER of PiezoBud on training dataset.

(b) EER v.s. Enroll time.

(c) EER v.s. #Users.

(d) EER v.s. User Diversity.

Figure 10: PiezoBud Performance over different enroll time and #users.

Components and Manufacturing		Price ($)	Total ($)
Components	Piezoelectric Sensor	0.34	9.17
	PDM Microphone	0.78	
	BLE Chip	0.67	
	Others	7.38	
PCB and SMT Service	Main component	17.47	20.65
	Amplification Circuit	3.18	

Table 2: Detailed price of PiezoBud prototype.

18 to over 40, including 45 native English speakers, all experienced with voiceprint-based applications. After training the authentication model with data from 85 subjects, we recruited 10 additional volunteers (8 males and 2 females) to evaluate the performance of PiezoBud under various impact factors and to validate system robustness against malicious attackers, as discussed in Section 3.

Data Collection During the experiment, participants were instructed to wear PiezoBud comfortably and read an article at a natural speed and volume. The PDM microphone was sampled at 16 kHz, and piezoelectric data was sampled at 8 kHz to reduce power consumption. All signals were recorded synchronously and transmitted to a MacBook Pro 2021. In the study, we gathered a unique dataset of 516 minutes, plus 10 more minutes data of 10 additional users (Section 8.3 and Section 8.4). All data were stored and trained locally.

Reading Materials To ensure PiezoBud is text-independent, we used three types of reading materials: a speech script, a fairy tale, and scientific content. Each participant was randomly given one to introduce content diversity and mitigate bias. Each material is about 1000 words long to ensure consistent reading duration.

Enrollment and Verification During the enrollment phase, embedding vectors of the piezoelectric (E_p^e) and audio (E_a^e) signals are fed into the flowAuth model, producing the output E_z^e. Centroids for both modalities and the output are determined by averaging their respective embedding vectors as C_a^e, C_p^e, and C_z^e. For verification, the embedding vectors E_p^v and E_a^v are used as inputs and conditions to generate E_z^v. Subsequently, we compute the average cosine similarity between the embedding vectors and centroids of each modality:

$$S_c = \frac{1}{3}\sum_i S_c^i(E_i^v, C_i^e) = \frac{1}{3}\sum_i \frac{E_i^v \cdot C_i^e}{\|E_i^v\|\|C_i^e\|} \quad i \in \{a, p, z\} \tag{8}$$

Next, we assess whether the cosine similarity value exceeds a threshold T generated from training data to determine if the current input is authentic. Authentication is granted only when the cosine similarity values surpass the threshold.

Evaluation Metrics We utilize the Equal Error Rate (EER) as our principal evaluation metric [68], which identifies the point at which the False Reject Rate (FRR) and the False Accept Rate (FAR) are equal across different thresholds. Generally, FAR and FRR are defined as follows:

$$\text{FAR} = \frac{\text{FP}}{\text{FP} + \text{TN}}, \quad \text{FRR} = \frac{\text{FN}}{\text{FN} + \text{TP}} \tag{9}$$

where TP, FN, FP, and TN denote true positive, false negative, false positive, and true negative, respectively. Figure 10(a) illustrates the variation of FAR and FRR on the training set. When the threshold T rises, the FRR rises while the FAR decreases. Our goal is to minimize EER to prevent both unexpected FN and FP simultaneously. The EER reaches its minimum value of 0.61% when the threshold is set at 0.56.

8.2 FusionSecNet Authentication Performance

Overall Performance We assess the overall performance using 15 seconds of enrollment data for each of the 85 users. PiezoBud achieves optimal performance, reaching the lowest EER of 1.05% with the threshold-obtained over training set. Besides, PiezoBud realizes 99.21% accuracy on speaker classification [65]. Compared to SOTA baselines in Table 1, PiezoBud boosts accuracy by 1.83% to 11.71%. Furthermore, PiezoBud demonstrates an 18% to 73.8% lower EER and require 5× to 8× shorter enrollment time.

Impact of Enrollment Data Lengths We investigate the influence of enrollment data lengths (from 5 to 25 s) on PiezoBud's performance. Figure 10(b) shows consistent EER values across the range from 1.15% to 1.06%. This analysis is vital, as the short enrollment time ensures user experience and system practicality.

Impact of User Number We analyze PiezoBud's performance relative to the number of users. This evaluation will assess the model's scalability, ensuring it maintains stable accuracy as the user base grows. Since a well-designed authentication system should deliver consistent performance regardless of user base size, we adhere to the same evaluation principles described in [1]. For each user, we obtain the enrollment embedding by randomly selecting 15 seconds of data from the testing set and inputting it into FusionSecNet. We tested PiezoBud with groups ranging from 10 to 80 subjects. The results, shown in Figure 10(c), demonstrate that the EER rises as the number of users increases from 10 to 40 (0.35% to 0.99%) but remains consistent as the number of users increases from 40 to 80 (0.99% to 1.09%). With an overall EER of 1.05% at 85 users, these findings indicate PiezoBud's adaptability to varying user profiles and its potential scalability to accommodate a larger user base.

Impact of User Diversity: We evaluate PiezoBud across diverse users, considering gender and English proficiency. Figure 10(d) shows an EER of 1.02% for females, 1.04% for males, 1.09% for native English speakers, and 1.00% for non-native speakers. These results highlight the consistent performance of PiezoBud regardless of gender or native language.

8.3 PiezoBud in Practical Scenarios

After training the authentication model with the 85-user dataset, we evaluate the robustness of PiezoBud against factors such as background noise, body movement, audio playback, and device re-wearing. The additional ten volunteers (eight males, two females), who are not part of the original 85 subjects, participate in this evaluation. We collect 30 seconds of data from each of the 10 volunteers. This non-overlapping setup ensures that the evaluation scenarios closely resemble real-world conditions.

Resistance to Ambient Noises In this experiment, we evaluate the performance of PiezoBud with five different ambient noise types: a) an office with white noise (40 dB), b) a meeting room with conversational noise (45 dB), c) a café with background chatter (55 dB), d) a bustling restaurant (80 dB), and e) an active construction site (85 dB). To ensure consistent noise levels, we play each type of noise at fixed volumes matching the previously mentioned dB levels using a MacBook Pro 2021. The noise simulation setup is shown on the left side of Figure 11(a). The results, depicted in Figure 11(b), show PiezoBud's superior performance, with lower EERs ranging from 1.05% to 3.84% across different noise environments, compared to 3.06% to 5.05% with only audio signals. This improvement is due to the piezoelectric modality capturing stable surface vibrations from the body, unaffected by environmental noise. The epoxy also further shields the piezoelectric sensor from external interference.

Reliability on Body Gestures We also conduct an experiment focusing on daily gestures that could disrupt the system, including a) walking, b) turning around, c) typing, and d) clapping. Participants perform these gestures, as shown on the right side of Figure 11(a), while speaking naturally. We collect a 30-second data sample for each gesture from each participant. The results are displayed in Figure 11(c). PiezoBud exhibits superior performance, with EERs ranging from 1.05% to 2.41% across diverse body movements. This is an improvement over EERs of 3.80% to 6.95% when solely utilizing the piezoelectric modality. This enhancement can be attributed to the stability of the audio signal, which remains unaffected by human body movement.

Solidness on Body Gestures with Noises We expand our evaluation of PiezoBud to include complex conditions, testing its performance with user motion and ambient noise. Experiments simulate common activities in various scenarios: a) typing in an office, b) typing in a cafe, c) walking on a construction site, and d) clapping in a meeting room. Noise levels match those previously described. As illustrated in Figure 11(d), PiezoBud shows superior performance, with EERs from 2.58% to 3.85%. In comparison, using only audio yields EERs between 4.60% and 5.30%, and using only the piezoelectric modality results in EERs from 4.90% to 7.35%. These results highlight PiezoBud's stability and effectiveness in challenging real-world conditions.

Resilience on Media Playing In addressing a common daily scenario, we investigate the performance of PiezoBud during media playback on earbuds. Existing works demand that these devices can't produce additional sounds or music during voice-based authentication processes [1–3], significantly restricting user experience. To assess PiezoBud's performance in realistic settings, participants spoke at their normal volume while media played through the earbuds, including TV shows, music, and movies. The results in Figure 12(a) show that PiezoBud's EER increases from 1.05% to 1.15%. The unaffected performance suggests PiezoBud's efficiency.

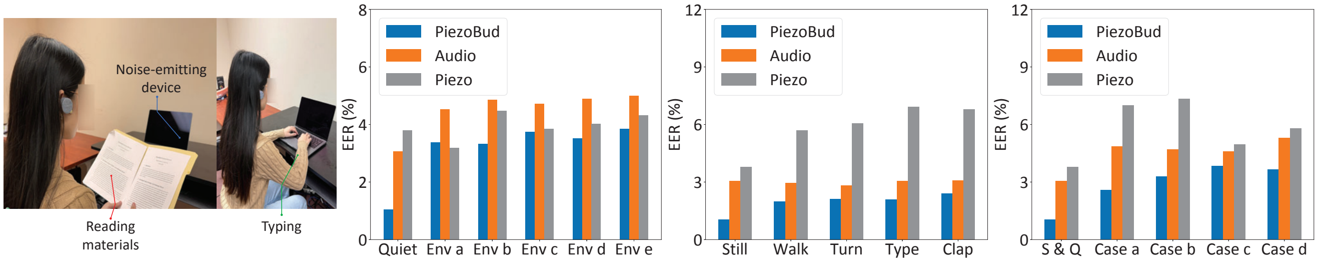
Robustness on Position Changes We assess PiezoBud's resilience to the practice of putting on and taking off earbuds, which may induce slight variations in sensor placement. Participants read the material, re-wore the device, and repeated this process three times. The results shown in Figure 12(b) reveal that the performance of PiezoBud remains consistent after re-wearing, with a minor change of 0.02%. PiezoBud's EER remain stable across both scenarios, indicating minimal impact from user positional shifts.

8.4 PiezoBud in Attack Scenarios

Following the protocol outlined in Section 8.3, we continue using the 10 volunteers who are not part of the original 85 participants. We ensure that PiezoBud has no prior interaction with these volunteers before the experiment. We collect 30 seconds of data from each of the 10 volunteers, including voice assistant commands and everyday conversations. Each participant's characteristics are represented by a unique centroid. We then assess PiezoBud's ability to counter the attack scenarios described in Section 3, evaluating the security effectiveness of the PiezoBud software-hardware co-design. The defense success rates are presented in Table 3.

Scenario 1. Attackers only launch audio replay/mimic attacks. In this scenario, attackers use only audio replay attacks without engaging the piezoelectric sensor. We record volunteers' voices using an iPhone 14 Pro Max and process them with voice cloning technologies (e.g., [44–46]) for mimic attacks. The deceptive audio is then played through a MacBook Pro 2021 while volunteers wear PiezoBud. When the victim is silent, the played audio is input into PiezoBud as the audio component, while the piezoelectric input is invalid due to no NAM signal. When the victim speaks, we employed an ultrasound speaker [69] to replay the audio, superimposing it over the victim's voice imperceptibly, and the piezoelectric modality remains the victim's own. The pre-validation step rejects all attacks due to the audio and piezoelectric data mismatch. We conducted 100 attacks for each combination, and none breached PiezoBud. These results demonstrate PiezoBud's effectiveness in defeating replay/mimic attacks by integrating piezoelectric data.

Scenario 2. Attackers falsify piezoelectric data to mimic victims'. We also consider a scenario where attackers gain direct



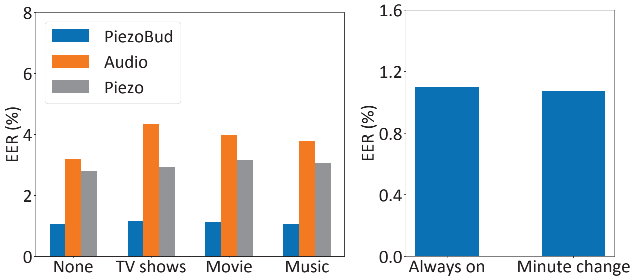
(a) Participants wear PiezoBud in different noise environments and motions.

(b) EER against ambient noises.

(c) EER against body gestures.

(d) EER against motion combined with noise. 'S&Q' indicates still and quiet.

Figure 11: PiezoBud performance over different noise types and body gestures compared with utilizing only audio or piezoelectric modality, respectively. The results show that PiezoBud has a more stable performance over different interference than the single modality.



(a) EER against different medias played.

(b) EER against position changes.

Figure 12: PiezoBud shows effective performance during media playback and maintain stable performance after re-wearing, even with minor position changes.

Attack Scenarios			Defense Success Rate
Scenario 1	Replay Only		100%
	Mimic	ResmbleAI [44]	100%
		PlayHT [46]	
		Vall-E [45]	
Scenario 2: Replay + Attacker's Piezo Scenario 3: Replay + Skull Generated Piezo Scenario 4: Replay + Generative AI Piezo			100%

Table 3: PiezoBud defends all the attack scenarios.

access to the victim's device. We examine the possibility of attackers using unrelated data with the victim's recorded audio to deceive PiezoBud. In this scenario, attackers steal the victim's PiezoBud. They may attempt to activate the piezoelectric input by directly tapping the sensor, or wear the device and mimic the victim's speech patterns to replicate their unique piezoelectric data. The recorded audio is played at a high volume to drown out the attacker's voice, making the audio input closely resemble the victim's. 100 attacks were conducted for each possible attack combination. Despite these efforts, PiezoBud successfully rejected all 200 attack attempts. PiezoBud's flowAuth component, which integrates user-specific audio and piezoelectric modalities, detects input inconsistencies, thwarting all malicious attacks.

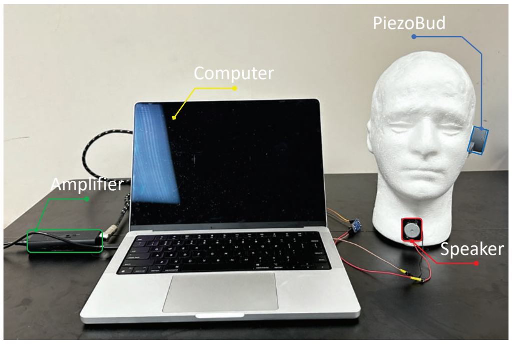


Figure 13: PiezoBud is attached to a skull's ear. A computer plays audio through an amplifier and speaker.

Scenario 3. Attackers use recorded audio data as input via different media. Furthermore, we envision attackers using a speaker within a synthetic human skull model to produce fake audio and piezoelectric data simultaneously, mimicking a human user (as shown in Figure 13). PiezoBud successfully countered all 100 attack attempts. The emulated piezoelectric modality lacks the user's distinct bio-metric properties, allowing PiezoBud to effectively reject these attacks.

Scenario 4. Attackers train piezo-to-audio converter networks with limited datasets. Attackers might use voice conversion networks to create synthetic piezoelectric data from recorded audio. Assuming a limited dataset (10 people), we use RVC [70] to generate piezoelectric data from audio. We select 40 minutes of data from 10 random users out of the previous 85 subjects. After training, we use newly collected volunteer audio to generate synthetic piezoelectric data. We disassemble PiezoBud and feed the recorded audio and synthetic piezoelectric data into the PCB pins. None of the 100 attacks bypassed PiezoBud, proving voice conversion techniques can't mimic the unique audio-piezoelectric relationship.

8.5 Latency and Energy Consumption

Since PiezoBud aims to perform the entire authentication process locally, we assess the model's complexity and feasibility by measuring processing times on various mobile devices, as shown in Table 4. Even on a five-year-old smartphone (Pixel 4), PiezoBud completes the authentication process in up to 100 ms, significantly shorter than the 500 ms sampling window and works in Table 1. This result validates the successful design of FusionSecNet on smartphones.

Apart from that, we also assess PiezoBud's battery performance using a Monsoon power monitor [71]. Power was supplied from a 3.3V DC source, simulating a coin cell battery. The system's average power consumption was 23.624 mW. With a CR2032 coin cell battery (210 mAh), PiezoBud can run continuously for about 10 hours.

8.6 User Experience Study

We survey 85 participants using a 5-point Likert scale. Our prototype receives ratings of 4.52 for comfort, 4.79 for size, and 4.35 for weight, indicating high satisfaction. Integrating piezoelectric sensors does not compromise performance, as 88% of participants report minimal awareness of the sensor. Additionally, 82% of respondents expressed interest in purchasing earbuds with authentication and protection against malicious attacks, underscoring the demand for enhanced security features.

9 RELATED WORKS

9.1 Sole Voiceprint User Authentication

Voiceprint-based authentication is widely adopted in multiple areas ranging from mobile [14] to IoT devices [13]. However, it overlooks the aspect of liveness [72], as it solely focuses on the physiological traits of speech. This oversight makes it vulnerable to spoofing and replay attacks [15]. LiVoAuth [73] and VoiceGesture [74] tackle voice authentication issues using vector sequences and articulatory gestures, while CaField [75] combats loudspeaker-based spoofing attacks by utilizing the 'fieldprint', a physical field of acoustic energy created as the sound propagates over the air. Despite their innovations, [73] and [74] struggle with environmental noise and user-device positioning, and [75] faces challenges in maintaining authentication across different sessions. In contrast, PiezoBud overcomes environmental and movement limitations by incorporating a miniature piezoelectric sensor, thereby enhancing effectiveness.

9.2 Multi-Modal Voiceprint Authentication

Recent studies have delved into voiceprint-based security using various sensor modalities, yet they face significant challenges. [76] employs mmWave to bolster VA security, but its integration into COTS earbuds is hindered by the requirement for large and complex devices. [3] and [2] enhance authentication by exploiting the unique shape of the ear canal, while [74] utilizes articulatory gestures. However, these methods are vulnerable to environmental noise and bodily changes. [1] introduces NAM signal collection via an in-ear microphone as an additional modality, but the prolonged data required for enrollment (75 seconds) hinders its practicality. [5–7] adopt IMU/ACC for user verification, facing constraints like limited attack scenarios and lack of real-time text independence. [8, 9] use piezoelectric sensors for NAM signal detection, yet the sensors' size and the computational intensity of the algorithms present practical limitations. Meanwhile, [8] functions solely as liveness detection, overlooking the risk of device theft. Additionally, it faces challenges in scenarios where user identification is needed [77]. [9] necessitates extensive enrollment data (107 seconds) and experiences latency ranging from 2.17 to 4.53 seconds, which constrains its practical application. PiezoBud overcomes previous limitations by providing liveness detection, user verification, and user identification with superior performance. Besides, its open-source platform allows integration with COTS earbuds, enhancing security without compromising user experience.

Phone Model	Running time (ms)		
	SynthEx	flowAuth	Overall
iPhone 13	23.08	3.25	49.41
iPhone 14 Pro Max	20.79	2.18	43.76
iPhone 15 Pro	19.92	2.01	41.85
Pixel 4	44.27	6.02	94.56
Pixel 4 XL	43.22	5.82	92.26
Pixel 6a	33.43	3.35	70.21

Table 4: Processing time of each phone model.

10 DISCUSSION

Human Voice Volume: Since the NAM signal is directly related to body vibrations, PiezoBud may experience reduced performance when users speak at lower volumes, as the resulting vibrations are weaker and more challenging to capture. In future work, we aim to address this limitation by refining the hardware design to amplify the raw signal strength, potentially through more sensitive piezoelectric sensors, improved amplification circuits, or optimized sensor placement to better capture subtle body vibrations during soft speech. This enhancement will allow PiezoBud to maintain high performance even in low-volume speech scenarios.

Impedance Match: In our current prototype, we employed a commonly used COTS amplifier for convenience and accessibility. However, this amplifier does not achieve impedance matching with the piezoelectric sensor. The impedance mismatch can lead to suboptimal signal transfer, resulting in reduced sensitivity and potential loss of important signal details captured by the sensor. For better performance in future designs, we could further improve the performance by integrating an amplifier specifically impedance-matched to the piezoelectric sensor.

11 CONCLUSION

This paper introduces PiezoBud, a cost-efficient, multi-modal wireless earbuds authentication system with open-source hardware. Integrating a miniature piezoelectric sensor, PiezoBud preserves user experience. Using 516 minutes of data from 85 subjects, PiezoBud achieves a low EER of 1.05% with 15 seconds of enrollment, outperforming SOTA baselines. PiezoBud counters advanced attacks, adapts to environmental changes, resists body motions, handles daily scenarios, enhances privacy, offers low latency and energy consumption, and provides local text-independent authentication.

12 ACKNOWLEDGEMENT

We sincerely thank our anonymous reviewers and shepherd for their invaluable feedback, which significantly improved our work. This work was partially supported by NSF CAREER Award 2338976.

REFERENCES

[1] Yang Gao, Yincheng Jin, Jagmohan Chauhan, Seokmin Choi, Jiyang Li, and Zhanpeng Jin. Voice in ear: Spoofing-resistant and passphrase-independent body sound authentication. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 5(1):1–25, 2021.

[2] Yang Gao, Wei Wang, Vir V. Phoha, Wei Sun, and Zhanpeng Jin. Earecho: Using ear canal echo for wearable authentication. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 3(3), 2019.

[3] Zi Wang, Sheng Tan, Linghan Zhang, Yili Ren, Zhi Wang, and Jie Yang. Eardynamic: An ear canal deformation based continuous user authentication using in-ear wearables. *Proceedings of Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 5(1), 2021.

[4] Seokmin Choi, Junghwan Yim, Yincheng Jin, Yang Gao, Jiyang Li, and Zhanpeng Jin. Earppg: Securing your identity with your ears. In *Proceedings of ACM IUI*, 2023.

[5] Huan Feng, Kassem Fawaz, and Kang G Shin. Continuous authentication for voice assistants. In *Proceedings of ACM MobiCom*, 2017.

[6] Jianwei Liu, Wenfan Song, Leming Shen, Jinsong Han, and Kui Ren. Secure user verification and continuous authentication via earphone imu. *IEEE Transactions on Mobile Computing*, 22(11):6755–6769, 2023.

[7] Tanmay Srivastava, Shijia Pan, Phuc Nguyen, and Shubham Jain. Jawthenticate: Microphone-free speech-based authentication using jaw motion and facial vibrations. In *In Proceedings of ACM SenSys*, 2023.

[8] Jiacheng Shang and Jie Wu. Enabling secure voice input on augmented reality headsets using internal body voice. In *Proceedings of IEEE SECON*, 2019.

[9] Rui Liu, Cory Cornelius, Reza Rawassizadeh, Ronald Peterson, and David Kotz. Vocal resonance: Using internal body voice for wearable authentication. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(1):1–23, 2018.

[10] WireCutter. Your wireless earbuds are trash (eventually). https://www.nytimes.com/wirecutter/blog/your-wireless-earbuds-are-trash-eventually/, Retrieved by June 23 2024.

[11] Apple. Siri. https://www.apple.com/siri/, Retrieved by June 23 2024.

[12] Google. Google assistant. https://assistant.google.com/.

[13] Yun-Tai Chang and Marc J. Dupuis. My voiceprint is my authenticator: A two-layer authentication approach using voiceprint for voice assistants. In *2019 IEEE SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI*, 2019.

[14] Gen Li, Zhichao Cao, and Tianxing Li. Echoattack: Practical inaudible attacks to smart earbuds. In *Proceedings of ACM MobiSys*, 2023.

[15] Zhi-Feng Wang, Gang Wei, and Qian-Hua He. Channel pattern noise based playback attack detection algorithm for speaker recognition. In *Proceedings of IEEE ICMI*.

[16] Yuan Gong and Christian Poellabauer. Crafting adversarial examples for speech paralinguistics applications. *arXiv preprint arXiv:1711.03280*, 2017.

[17] Rafizah Mohd Hanifa, Khalid Isa, and Shamsul Mohamad. A review on speaker recognition: Technology and challenges. *Computers & Electrical Engineering*, 90:107005, 2021.

[18] Google. Which pixel buds are right for you? https://store.google.com/us/magazine/pixel_buds_compare?hl=en-US, Retrieved by June 23 2024.

[19] Galina Lavrentyeva, Sergey Novoselov, Egor Malykh, Alexander Kozlov, Oleg Kudashev, and Vadim Shchemelinin. Audio replay attack detection with deep learning frameworks. In *Proceedings of ISCA Interspeech*, 2017.

[20] Anton Firc, Kamil Malinka, and Petr Hanáček. Deepfakes as a threat to a speaker and facial recognition: An overview of tools and attack vectors. *Heliyon*, 9(4):e15090, 2023.

[21] Apple. Airpods (3rd generation)). https://www.apple.com/airpods-3rd-generation/, Retrieved by June 23 2024.

[22] Jingjin Li, Chao Chen, Mostafa Rahimi Azghadi, Hossein Ghodosi, Lei Pan, and Jun Zhang. Security and privacy problems in voice assistant applications: A survey. *Computers & Security*.

[23] Nikolay Ivanov, Chenning Li, Qiben Yan, Zhiyuan Sun, Zhichao Cao, and Xiapu Luo. Security threat mitigation for smart contracts: A comprehensive survey. *ACM Computing Surveys*, 2023.

[24] Guangjing Wang, Ce Zhou, Yuanda Wang, Bocheng Chen, Hanqing Guo, and Qiben Yan. Beyond boundaries: A comprehensive survey of transferable attacks on ai systems. *arXiv preprint arXiv:2311.11796*, 2023.

[25] Hanqing Guo, Guangjing Wang, Yuanda Wang, Bocheng Chen, Qiben Yan, and Li Xiao. Phantomsound: Black-box, query-efficient audio adversarial attack via split-second phoneme injection. In *In Proceedings of ACM RAiD*, 2023.

[26] Yuanda Wang, Hanqing Guo, Guangjing Wang, Bocheng Chen, and Qiben Yan. Vsmask: Defending against voice synthesis attack via real-time predictive perturbation. In *In Proceedings of ACM WISEC*, pages 239–250, 2023.

[27] Mark D Fletcher, Sian Lloyd Jones, Paul R White, Craig N Dolder, Timothy G Leighton, and Benjamin Lineton. Effects of very high-frequency sound and ultrasound on humans. part ii: A double-blind randomized provocation study of inaudible 20-khz ultrasound. *The Journal of the Acoustical Society of America.*

[28] Jasper Lastoria. The 7 best true wireless earbuds - summer 2024 reviews. https://www.rtings.com/headphones/reviews/best/by-type/truly-wireless-earbuds.

[29] Maolin Gan, Yimeng Liu, Li Liu, Chenshu Wu, Younsuk Dong, Huacheng Zeng, and Zhichao Cao. Poster: mmleaf: Versatile leaf wetness detection via mmwave sensing. In *Proceedings of ACM MobiSys*, 2023.

[30] Ruihao Wang, Yimeng Liu, and Rolf Müller. Detection of passageways in natural foliage using biomimetic sonar. *Bioinspiration & Biomimetics*, 2022.

[31] Yimeng Liu, Maolin Gan, Huaili Zeng, Liu Li, Younsuk Dong, and Zhichao Cao. Hydra: Accurate multi-modal leaf wetness sensing with mm-wave and camera fusion. In *Proceedings of ACM MobiCom*, 2024.

[32] Hans Von Leden. The mechanism of phonation: a search for a rational theory of voice production. *Archives of Otolaryngology*, 74(6):660–676, 1961.

[33] Huaili Zeng, Wei Xu, Bo Dong, Changyuan Yu, Wei Zhao, Yishan Wang, and Wenye Sun. Beat-to-beat heart rate estimation from mzi-bcg signal based on hierarchical clustering. In *2021 Opto-Electronics and Communications Conference (OECC)*, 2021.

[34] Huaili Zeng, Wei Xu, Bo Dong, Changyuan Yu, Wei Zhao, Yishan Wang, and Wenye Sun. Non-invasive highly sensitive under mattress vital signs monitoring based on fiber sagnac loop. In *2021 Opto-Electronics and Communications Conference (OECC)*, 2021.

[35] Kurt Barbe, Rik Pintelon, and Johan Schoukens. Welch method revisited: non-parametric power spectrum estimation via circular overlap. *IEEE Transactions on signal processing*.

[36] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*.

[37] Mach1 Research. Imu enabled devices. https://research.mach1.tech/posts/imu-enabled-devices/, Retrieved by Apr 17 2024.

[38] TDK Corporation. Phua2010-049b-00-000. https://product.tdk.com/en/search/sw_piezo/speaker/piezolisten/info?part_no=PHUA2010-049B-00-000, Retrieved by Apr 17 2024.

[39] TDK Invensense. Mpu-9250. https://invensense.tdk.com/products/motion-tracking/9-axis/mpu-9250/, Retrieved by Apr 17 2024.

[40] Lin Zhou, Eric Fischer, Can Tunca, Clemens Markus Brahms, Cem Ersoy, Urs Granacher, and Bert Arnrich. How we found our imu: Guidelines to imu selection and a comparison of seven imus for pervasive healthcare applications. *Sensors*, 20(15):4090, 2020.

[41] Jingdong Zhao. A review of wearable imu (inertial-measurement-unit)-based pose estimation and drift reduction technologies. In *Journal of Physics: Conference Series*, volume 1087, page 042003, 2018.

[42] TDK Corporation. Phua3015-049b-00-000. https://product.tdk.com/en/search/sw_piezo/speaker/piezolisten/info?part_no=PHUA3015-049B-00-000, Retrieved by Apr 17 2024.

[43] TDK Corporation. Phua3030-049b-00-000. https://product.tdk.com/en/search/sw_piezo/speaker/piezolisten/info?part_no=PHUA3030-049B-00-000, Retrieved by Apr 17 2024.

[44] Resemble. Resemble.ai. https://www.resemble.ai/, Retrieved by Apr 17 2024.

[45] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models are zero-shot text to speech synthesizers, 2023.

[46] play.hy. Playht. https://play.ht/, Retrieved by Apr 17 2024.

[47] Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. A survey on neural speech synthesis, 2021.

[48] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of IEEE CVPR*, 2016.

[49] Google. Webrtc, Retrieved by Apr 17 2024. [Online; accessed 30-June-2020].

[50] Tim Sainburg, Marvin Thielk, and Timothy Q Gentner. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS computational biology*, 16(10):e1008228, 2020.

[51] Maans Klingspor. Hilbert transform: Mathematical theory and applications to signal processing, 2015.

[52] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Proceedings of ISCA Interspeech*, 2020.

[53] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of IEEE CVPR*, 2018.

[54] Zhenduo Zhao, Zhuo Li, Wenchao Wang, and Pengyuan Zhang. Pcf: Ecapa-tdnn with progressive channel fusion for speaker verification. In *Proceedings of IEEE ICASSP*, 2023.

[55] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation, 2015.

[56] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification, 2020.

[57] PUIaudio. Ab1070b-lw100-r. https://puiaudio.com/product/benders/ab1070b-lw100-r, Retrieved by Apr 17 2024.

[58] Texas Instruments. Lm358p. https://www.ti.com/lit/ds/symlink/lm358b.pdf?ts=1712387960629&ref_url=https%253A%252F%252Fwww.google.com%252F, Retrieved by Apr 17 2024.

[59] Analog. Max4466. https://www.analog.com/en/products/max4466.html, Retrieved by Apr 17 2024.

[60] STMicroelectronics. Mp23db01hp. https://www.st.com/en/mems-and-sensors/mp23db01hp.html, Retrieved by Apr 17 2024.

[61] Nordic. nrf52840. https://www.nordicsemi.com/Products/nRF52840, Retrieved by Apr 17 2024.

[62] Huaili Zeng, Gen Li, and Tianxing Li. Pyrosense: 3d posture reconstruction using pyroelectric infrared sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2024.

[63] Yidong Ren, Puyu Cai, Jinyan Jiang, Jialuo Du, and Zhichao Cao. Prism: High-throughput lora backscatter with non-linear chirps. In *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*, 2023.

[64] Yidong Ren, Wei Sun, Jialuo Du, Huaili Zeng, Younsuk Dong, Mi Zhang, Shigang Chen, Yunhao Liu, Tianxing Li, and Zhichao Cao. Demeter: Reliable cross-soil lpwan with low-cost signal polarization alignment. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, 2024.

[65] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *Proceedings of ISCA Interspeech*, 2017.

[66] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF CVPR*, 2019.

[67] Hanqing Guo, Qiben Yan, Nikolay Ivanov, Ying Zhu, Li Xiao, and Eric J. Hunter. Supervoice: Text-independent speaker verification using ultrasound energy in human speech. In *Proceedings of ACM CCS ASIA*, 2022.

[68] James L Wayman. Error rate equations for the general biometric system. *IEEE Robotics & Automation Magazine*, 6(1):35–48, 1999.

[69] Avisoft Biosacoustics. Ultrasonic dynamic speaker vifa. http://www.avisoft.com/playback/vifa/.

[70] RVC-Project. Retrieval-based-voice-conversion-webui. https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI?tab=readme-ov-file, Retrieved by Apr 17 2024.

[71] Monsoon. High voltage power monitor. https://www.msoon.com/online-store/High-Voltage-Power-Monitor-p90002590, Retrieved by Apr 17 2024.

[72] Lei Zhang, Yan Meng, Jiahao Yu, Chong Xiang, Brandon Falk, and Haojin Zhu. Voiceprint mimicry attack towards speaker verification system in smart home. In *Proceedings of IEEE INFOCOM*, 2020.

[73] Rui Zhang, Zheng Yan, Xuerui Wang, and Robert H. Deng. Livoauth: Liveness detection in voiceprint authentication with random challenges and detection modes. *IEEE Transactions on Industrial Informatics*, 19(6):7676–7688, 2023.

[74] Linghan Zhang, Sheng Tan, and Jie Yang. Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication. In *Proceedings of ACM CCS*, 2017.

[75] Chen Yan, Yan Long, Xiaoyu Ji, and Wenyuan Xu. The catcher in the field: A fieldprint based spoofing detection for text-independent speaker verification. In *Proceedings of ACM CCS*, 2019.

[76] Huining Li, Chenhan Xu, Aditya Singh Rathore, Zhengxiong Li, Hanbin Zhang, Chen Song, Kun Wang, Lu Su, Feng Lin, Kui Ren, et al. Vocalprint: exploring a resilient and secure voice authentication via mmwave biometric interrogation. In *Proceedings of ACM SenSys*, 2020.

[77] Meta. Meta quest for business. https://forwork.meta.com/quest/business-subscription/, Retrieved by Apr 17 2024.