AuthGlass: Benchmarking Voice Liveness Detection and Authentication on Smart Glasses via Comprehensive Acoustic Features

WEIYE XU, Tsinghua University, China
ZHANG JIANG, Tsinghua University, China
SIQI ZHENG, Tsinghua University, China
XIYUXING ZHANG, Tsinghua University, China
CHANGHAO ZHANG, Ant Group, China
JIAN LIU, Ant Group, China
WEIQIANG WANG, Ant Group, China
YUNTAO WANG*, Tsinghua University, China

With the rapid advancement of smart glasses, voice interaction has been widely adopted due to its naturalness and convenience. However, its practical deployment is often undermined by vulnerability to spoofing attacks, while no public dataset currently exists for voice liveness detection and authentication in smart-glasses scenarios. To address this challenge, we first collect a multi-acoustic-modal dataset comprising 16-channel audio data from 42 subjects, along with corresponding attack samples covering two attack categories. Based on insights derived from this collected data, we propose AuthG-Live, a sound-field-based voice liveness detection method, and AuthG-Net, a multi-acoustic-modal authentication model. We further benchmark seven voice liveness detection methods and four authentication methods across diverse acoustic modalities. The results demonstrate that our proposed approach achieves state-of-the-art performance on four benchmark tasks, and extensive ablation studies validate the generalizability of our methods under real-world constraints. Finally, we release this dataset, termed AuthGlass, to facilitate future research on voice liveness detection and authentication for smart glasses.

CCS Concepts: • **Security and privacy → Authentication**; • **Human-centered computing → Ubiquitous and mobile devices**; *Sound-based input / output*.

Additional Key Words and Phrases: smart glasses, voice authentication, liveness detection, acoustic sensing, sound field, multi-modal, dataset, benchmark

ACM Reference Format:
Weiye Xu, Zhang Jiang, Siqi Zheng, Xiyuxing Zhang, Changhao Zhang, Jian Liu, Weiqiang Wang, and Yuntao Wang. 2026. AuthGlass: Benchmarking Voice Liveness Detection and Authentication on Smart Glasses via Comprehensive Acoustic Features. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 10, 3, Article 180 (September 2026), 43 pages. https://doi.org/10.1145/3831980

*Corresponding author.

Authors' Contact Information: Weiye Xu, Tsinghua University, Key Laboratory of Pervasive Computing, Ministry of Education, Department of Computer Science and Technology, Global Innovation Exchange (GIX) Institute, Beijing National Research Center for Information Science and Technology (BNRist), Beijing, China, xuwy24@mails.tsinghua.edu.cn; Zhang Jiang, Tsinghua University, Beijing, China, jiangzhang20030102@gmail.com; Siqi Zheng, Tsinghua University, Beijing, China, ymzqwq@gmail.com; Xiyuxing Zhang, Tsinghua University, Beijing, China, zxyx22@mails.tsinghua.edu.cn; Changhao Zhang, Ant Group, Hangzhou, China, eleven.zch@antgroup.com; Jian Liu, Ant Group, Hangzhou, China, rex.lj@antgroup.com; Weiqiang Wang, Ant Group, Hangzhou, China, weiqiang.wwq@antgroup.com; Yuntao Wang (corresponding author), Tsinghua University, Beijing, China and Qinghai University, Xining, China, yuntaowang@tsinghua.edu.cn.



This work is licensed under a Creative Commons Attribution 4.0 International License.
© 2026 Copyright held by the owner/author(s).
ACM 2474-9567/2026/9-ART180
https://doi.org/10.1145/3831980

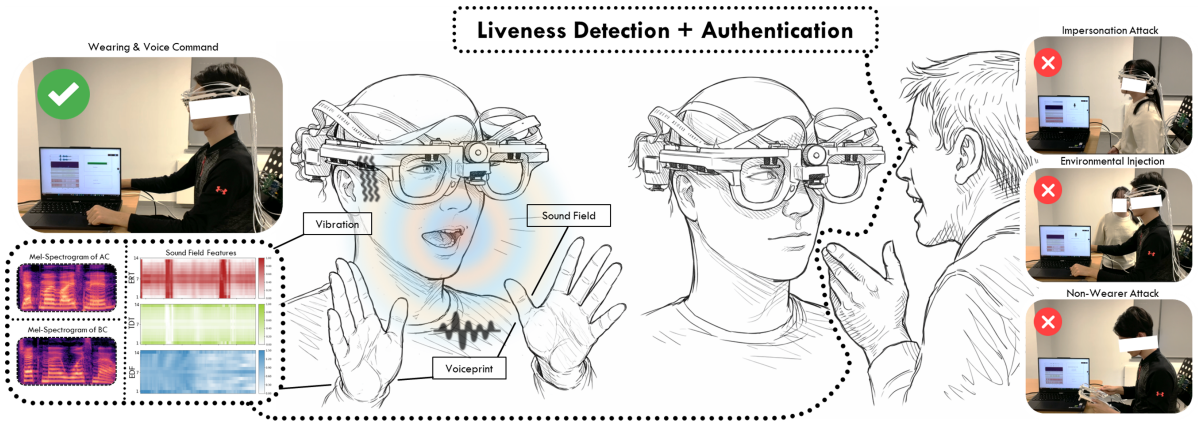


Fig. 1. Based on insights derived from our collected multi-acoustic sensing data, we develop AuthG-Live, which performs liveness detection using sound-field features, and AuthG-Net, which conducts authentication by jointly leveraging voiceprint, vibration, and sound-field features. The proposed framework is generalizable to different input-channel configurations. In this figure, liveness detection and authentication are demonstrated using five air-conductive microphones (consistent with Ray-Ban Meta [7]) mounted on the smart glasses and operating at a sampling rate of 16 kHz. As illustrated, the system successfully rejects three adversarial scenarios: impersonation attack, environmental injection and non-wearer input.

1 INTRODUCTION

Smart glasses have rapidly evolved from experimental prototypes to widely deployed wearable platforms, with commercial products such as Microsoft HoloLens [4] and Ray-Ban Meta Glasses [7] gaining increasing adoption. In emerging application scenarios such as driving [23, 35], cooking [19, 56], and industrial workflows [66], users rely on smart glasses to access information and execute commands while their hands are occupied. As these devices become tightly integrated with users' daily activities and increasingly handle sensitive personal or enterprise data, ensuring secure and reliable input mechanisms becomes critical [63].

Existing authentication solutions for smart glasses span multiple paradigms. Commercial systems often rely on paired smartphones to provide *device-level authentication* [7, 13], where access control is delegated to a trusted companion device. While effective for protecting device access, such approaches cannot distinguish between different users once the device is unlocked. Other user-level authentication methods, including password-based input [40], iris recognition [14, 45, 52], and gaze-based techniques [30], typically require explicit user interaction or specialized hardware, making them less suitable for seamless wearable use.

Voice interaction has emerged as the dominant input modality on smart glasses due to its hands-free nature and minimal hardware requirements [7, 8]. It also enables implicit authentication by leveraging existing voice interfaces without requiring explicit user actions or additional sensors. However, this openness to ambient audio exposes voice-based systems to practical attacks, including command injection via loudspeakers, replayed recordings, and synthetic speech impersonation [90, 96], necessitating robust liveness detection and authentication mechanisms.

Although voice authentication and liveness detection have been extensively studied on smartphones and other wearable devices [18, 29, 33, 44, 48, 58, 83, 90, 96], directly applying these approaches to smart glasses remains challenging. A key limitation lies not only in the lack of datasets collected on smart-glasses platforms, but also in their unique sensing characteristics. Unlike handheld or ear-worn devices, smart glasses integrate multiple microphones along a curved frame around the head, capturing rich yet underexplored acoustic cues such as speech-induced vibrations and spatial sound-field patterns coupled with human anatomy. Existing methods,

primarily designed for single-channel or regular-array configurations, do not account for these properties, making their effectiveness on smart glasses unclear.

To address these challenges, we develop a DIY smart-glasses platform that captures 16-channel synchronized multi-acoustic-modal audio at 96 kHz, and build *AuthGlass*, a dataset collected from 42 users with two representative attack sample sets. Based on this dataset, we analyze smart-glasses-specific acoustic characteristics and propose *AuthG-Live*, a sound-field-based liveness detection method, and *AuthG-Net*, a multi-modal authentication model that jointly leverages voiceprint, vibration, and spatial acoustic cues. We further benchmark seven liveness detection methods and four authentication methods across multiple modalities, establishing a systematic evaluation framework.

Our approach achieves over 96% liveness detection accuracy and 97% authentication accuracy, with strong generalization across attack types and robustness under cross-utterance settings. Extensive ablation studies demonstrate effectiveness under practical constraints and provide insights into microphone placement for smart glasses. To support future research, we open-source both the dataset and the smart-glasses platform implementation[1].

Our key contributions are summarized as follows:

- We propose and open-source the AuthGlass dataset, which contains high-sampling-rate, 16-channel synchronized multi-acoustic-modal recordings from 42 participants, along with two attack scenarios. We also release the corresponding data capture platform.
- We propose AuthG-Live, a voice liveness detection method using sound-field features that generalizes across different attacks, and AuthG-Net, a multi-acoustic-modal voice authentication model supporting utterance-independent authentication.
- We benchmark seven voice liveness detection methods and four voice authentication methods across four tasks, demonstrating that our approach achieves state-of-the-art performance.
- Through ablation studies, we show that our methods remain effective under limited acoustic channel settings, and provide design insights into microphone placement on smart glasses.

2 RELATED WORKS

This section summarize the related works, including voice interface on smart glasses (Section 2.1), liveness detection and authentication on wearable devices (Section 2.2), and benchmark for liveness detection and authentication on smart glasses (Section 2.3)

2.1 Voice Interface on Smart Glasses

Voice user interfaces (VUIs) are a natural and intuitive interaction modality, widely adopted across domains such as smart homes [38, 75, 79], in-vehicle systems [27, 41], and wearable devices [24]. For smart glasses, VUIs rely only on lightweight microphones [61], making them particularly suitable for resource-constrained form factors and thus widely deployed in commercial products such as Ray-Ban Meta [7], Xreal One Pro [11], and Rokid Glasses [8]. What's more, voice input supports hands-free command execution and conversational AI, especially with the emergence of large language models. For instance, Rokid enables voice-based payments, while Ray-Ban Meta integrates voice interaction with AI services. Prior work further highlights its benefits in accessibility (e.g., assisting visually impaired users [31, 60, 67, 77]) and in attention-critical scenarios such as medical and industrial settings [46, 99].

Despite these advantages, the security of voice interfaces on smart glasses remains underexplored. As devices increasingly handle sensitive personal data, they are exposed to realistic threats from ambient audio and adversarial inputs. Advances in speech synthesis and deepfake technologies [43, 64, 80] enable attacks such as replay and

[1]Project URL: https://github.com/xuy8w89/IMWUT2026-AuthGlass

synthetic speech injection [28, 42, 94], which are particularly concerning given the always-listening nature of smart glasses [21, 50]. Prior studies show that authentication methods relying solely on acoustic features are vulnerable to replay [82, 94] and synthetic attacks [85, 101], indicating limited robustness [49, 89].

While voice liveness detection and authentication have been explored on other devices [55, 84, 95], research targeting smart glasses remains limited. In particular, the lack of dedicated hardware platforms and standardized datasets hinders systematic evaluation. To address this gap, we develop and release a DIY smart glasses prototype that enables comprehensive multi-modal acoustic data collection, facilitating benchmarking for voice liveness detection and authentication in this domain.

2.2 Liveness Detection and Authentication on Wearable Devices

Wearable devices such as earbuds and smartwatches are increasingly used as personal computing and interaction platforms [70]. As they handle sensitive user data and enable command execution, securing access has become critical [16]. Existing solutions can be broadly divided into *liveness detection* and *authentication*. Liveness detection distinguishes genuine human inputs from spoofed or adversarial signals, while authentication verifies user identity. Prior work has explored diverse sensing modalities, including audio [15, 26, 36], motion [22, 76, 92], and physiological signals such as PPG [100] and EMG [73]. Among these, voice-based approaches [15, 33, 74] are particularly attractive due to their low hardware requirements, seamless integration with existing voice interfaces, and hands-free usability. For example, Huang et al. [37] exploit the consistency between air- and bone-conducted signals for liveness detection, while Gao et al. [33] utilize in-ear and out-ear acoustic features for authentication.

Compared to other wearables, smart glasses present distinct opportunities and challenges for voice-based security. Commercial devices [7, 8, 11] typically integrate multiple microphones distributed along the frame, naturally enabling multi-channel audio capture with rich spatial and acoustic cues that can benefit liveness detection and authentication [90, 93]. Moreover, there is currently no unified hardware platform or publicly available dataset to systematically benchmark voice liveness detection and authentication in smart-glasses scenarios.

To bridge this gap, we develop a multi-microphone DIY smart-glasses platform to collect multi-channel acoustic data, and build a benchmark with multi-acoustic modalities. Based on this, we further design *AuthG-Live*, a sound-field-aware liveness detection method, and *AuthG-Net*, a multi-acoustic-modality authentication model that jointly leverages voiceprint, vibration, and spatial acoustic features, enabling systematic evaluation and advancing robust voice-based security on smart glasses.

2.3 Benchmark for Liveness Detection and Authentication on Smart Glasses

Modern smart glasses increasingly adopt multi-channel audio as a primary modality for voice interaction [7, 8, 11], making robust liveness detection and authentication essential for security-sensitive use. While these problems have been extensively studied in automatic speaker verification (ASV), existing datasets are not tailored to smart-glasses scenarios.

Widely used datasets such as ASVspoof [88], VoxCeleb [59], RedDots Replayed [47], and SITW [57] provide large-scale genuine and spoofed speech data for general ASV systems. However, they are primarily collected with single-channel or limited microphone setups and do not reflect the smart-glasses form factor. In particular, they fail to capture multi-channel acoustic characteristics induced by microphones distributed around the eyeglass frame, where spatial cues, channel-dependent energy patterns, and time delays are tightly coupled with device geometry. Complementary datasets such as REALMAN [91] and the Massive Distributed Microphone Array Dataset [25] provide realistic multi-channel recordings using wearable or distributed microphone arrays. Nevertheless, they are not designed for ASV tasks and do not include replay or synthetic attacks, limiting their applicability for evaluating liveness detection and authentication.

Table 1. Comparison of voice-based liveness detection or authentication systems on wearable devices.

Name	Hardware	Sampling Rate	Channels	Modalities	Open-Source
VOGUE[22]	Gyroscope embedded in existing device	44.1 kHz / 50 Hz	2	Air-conducted microphone + gyroscope	×
Vauth[29]	IMU manually mounted on glasses nose pad	44.1 kHz / 11 kHz	2	Air-conducted microphone + vibration sensor	×
Eve Said Yes[37]	BC microphone module manually mounted	48 kHz	2	Air-conducted microphone + bone-conducted microphone	×
Voshield[93]	Microphone array module	16 kHz	4	Air-conducted microphones	×
Ours	Multi-acoustic-modal microphone array on smart glasses frame	96 kHz	16	Multi air-conducted + bone-conducted microphones	✓

Meanwhile, prior work on wearable or head-mounted devices [29, 51, 98] has explored voice-based liveness detection and authentication, but typically relies on custom hardware with heterogeneous microphone layouts and acquisition settings. The lack of publicly available datasets and standardized platforms hinders reproducibility and fair comparison, as summarized in Table 1.

To address these gaps, we introduce a unified multi-microphone smart-glasses platform and collect a comprehensive dataset from 42 users, covering both genuine speech and representative spoofing attacks. By releasing both the hardware design and dataset, our work provides, to the best of our knowledge, the first benchmark dedicated to voice liveness detection and authentication on smart glasses.

3 AUTHGLASS DATASET: COMPREHENSIVE PASSIVE ACOUSTIC FEATURES FOR LIVENESS DETECTION AND AUTHENTICATION

This section introduce the AuthGlass Dataset that aims for two main objectives: **voice liveness detection** and **authentication** on smart glasses via passive acoustic sensing. Compared with existing datasets for authentication or liveness detection, our dataset offers four key advantages. **(a) Target on Smart Glasses**. To the best of our knowledge, AuthGlass is the first open-source acoustic dataset collected on a smart-glasses form factor for authentication and liveness detection. The dataset captures rich, multi-channel air- and bone-conducted signals from a glasses-mounted sensing platform that coupled with facial anatomy. **(b) Diverse Speakers and Spoofing Scenarios**. AuthGlass includes 42 participants, each providing 15 voice-command samples repeated at six volume levels, capturing substantial intra-speaker variability. The dataset further covers two spoofing categories with multiple attack scenarios, supporting systematic evaluation under diverse adversarial settings. **(c) 16-Channel Synchronized Multimodal Acoustic Capture**. The dataset captures both air- and bone-conducted acoustic signals, totaling 16 synchronized channels. This multimodal design enables the extraction of diverse acoustic features, including speech, vibration, and sound-field cues, and facilitates robust multimodal fusion for wearable authentication. **(d) High-Resolution Data**. Data are collected using a DIY smart-glasses prototype worn naturally

Table 2. Comparison of user-level authentication modalities on smart glasses or head wearables. **Task** indicates whether explicit user actions are required for authentication. **Effort** denotes additional interaction beyond normal usage. **Hands-free** indicates whether the modality supports fully hands-free interaction. **Continuous Auth.** indicates whether authentication can be performed continuously without user input. **Commodity HW** reflects common support on off-the-shelf smart glasses. **Energy** is a qualitative estimate of runtime cost.

Modality	Task	Effort	Hands-free	Continuous Auth.	Commodity HW	Energy
Touch (Password) [40]	✓	✓	×	No	✓	Low
Touch (Fingerprint) [87]	✓	×	×	Event-driven	×	Low
Gaze (Password) [30]	✓	✓	✓	No	×	High
Gaze (Implicit Task) [39]	×	×	✓	No	×	High
Iris recognition [14, 45]	✓	✓	✓	No	×	High
Facial Biometrics [50, 54]	×	×	✓	Continuous	×	High
Skull Biometrics [69, 72, 86]	×	×	✓	Continuous	×	Medium
Voice [37, 97]	×	×	✓	Event-driven	✓	Low

by users, with all audio recorded at a 96 kHz sampling rate to preserve fine-grained acoustic details. Replay attacks are generated using high-quality playback devices, resulting in realistic and challenging attack conditions.

While these advantages highlight the richness of AuthGlass, its design is driven by fundamental questions of motivation, threat models, and usable acoustic cues for smart-glasses-based authentication. Accordingly, we first formalize liveness detection and authentication under explicit threat models and usage scenarios to clearly define the objectives of the dataset (Section 3.1). We then identify key challenges introduced by the smart-glasses form factor (Section 3.2) and explore effective acoustic features across modalities to determine which signals should be collected (Section 3.3). Finally, we integrate these insights to construct the AuthGlass dataset, including hardware design, data collection, and statistical analysis (Section 3.4.1).

3.1 Motivation and Threat Model for AuthGlass Dataset Design

In this section, we motivate voice authentication on smart glasses and systematically define the authentication pipeline and threat models, laying the foundation for the AuthGlass dataset design objectives.

3.1.1 Motivation for Voice Liveness Detection and Authentication. Voice interaction has emerged as the primary input modality on modern smart glasses due to its hands-free and natural interaction paradigm [7, 8, 10, 11]. For example, users can issue voice commands to access information, control applications, or interact with AI assistants without interrupting their ongoing activities. Building authentication mechanisms directly on top of voice input therefore provides a seamless and intuitive solution that aligns with everyday usage.

Table 2 compares voice-based authentication with existing modalities for smart glasses. Compared to touch- and gaze-based approaches, voice authentication does not require explicit user tasks or additional interaction effort, enabling a frictionless user experience. Unlike touch-based methods, it naturally supports hands-free interaction, which is essential for wearable devices. In contrast to vision-based biometrics (e.g., iris or facial contour recognition), voice authentication can be implemented using commodity microphones already available on current smart glasses, while incurring relatively low energy consumption. Although voice authentication is inherently event-driven rather than continuous, it aligns well with command-based interaction patterns, allowing authentication to be performed opportunistically during normal usage.

Despite these advantages, voice-based systems remain vulnerable to spoofing attacks (e.g., replay or synthesis) and environmental noise. Therefore, it is critical to systematically understand the role of voice authentication

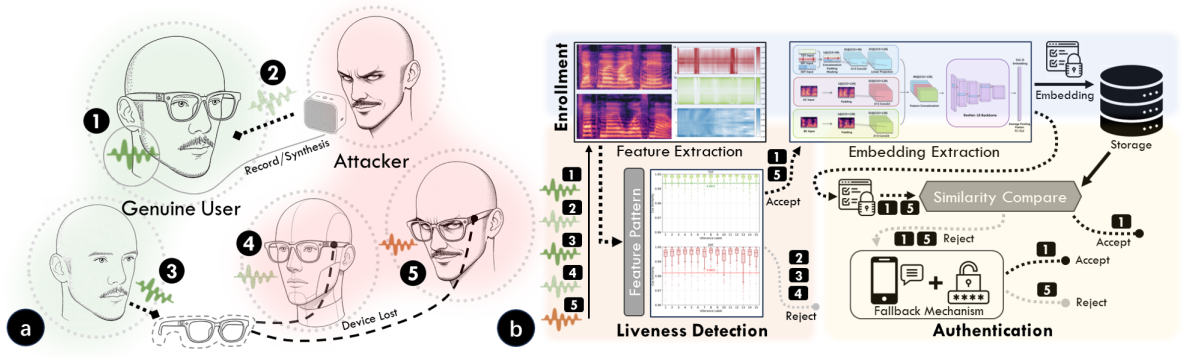


Fig. 2. Voice authentication pipeline and threat model. (a) Adversarial scenarios: (1) genuine user input, (2) environmental injection attacks, (3) non-wearer voice inputs, (4) advanced replay/spoofing attacks, and (5) device theft with impersonation. (b) System workflow: during enrollment, user voice embeddings are extracted and stored as templates. During operation, incoming speech first undergoes feature extraction and liveness detection, followed by embedding-based similarity comparison for authentication. Inputs from environmental injection attacks (2), non-wearer voice inputs (3), and advanced replay/spoofing attacks (4) are expected to be rejected at the liveness detection stage. In contrast, genuine user input (1) and device theft with impersonation (5) can pass liveness detection; the former is expected to be successfully authenticated, while the latter is rejected during the authentication stage. A fallback mechanism (e.g., smartphone-based verification) is triggered upon rejection to further prevent unauthorized access and to handle occasional false rejections of genuine users.

in practical systems. In particular, voice authentication is best suited for lightweight, event-driven verification without requiring additional user attention, while more secure fallback mechanisms (e.g., passwords on paired trusted devices) may still be needed for high-risk operations.

Motivated by these observations, we next formalize the authentication process for voice-based smart-glasses systems and define representative threat models and adversarial scenarios to guide the design of AuthGlass Dataset.

3.1.2 Voice Authentication Process and Threat Model. Voice Authentication Workflow. We consider an event-driven voice authentication pipeline on smart glasses (Fig. 2b). Prior to deployment, each user completes a lightweight **enrollment** procedure by recording a small set of voice samples (e.g., passphrases), from which acoustic embeddings are extracted and stored as user-specific templates.

During operation, authentication is implicitly triggered by voice commands. Upon receiving input, the system first performs **liveness detection** to verify that the signal originates from a live speaker wearing the device, thereby rejecting replayed or externally injected audio. If successful, the system proceeds to **authentication**, where extracted features are matched against the enrolled template to verify identity. Commands are executed only if both stages succeed.

To account for practical variability (e.g., environmental noise, intra-user differences, or device conditions), voice authentication may occasionally result in false rejections. To maintain usability, we incorporate a fallback authentication mechanism (e.g., password input on a paired trusted device such as a smartphone), serving as a secondary authentication channel when voice-based verification fails. Compared to voice interaction, this fallback is less hands-free and attention-free, but provides a more robust and secure backup in rare failure cases.

Adversarial Scenarios and Threat Model. We model realistic adversarial scenarios based on a representative smart-glasses usage setting (e.g., industrial environments such as factory workers). We define four attack scenarios and explicitly specify the stage (liveness detection or authentication) each targets (Fig. 2a).

(1) Environmental Injection Attacks (Liveness Failure Expected). An attacker injects malicious voice commands into the environment (e.g., via loudspeakers), attempting to trigger unintended operations. These commands may originate from text-to-speech (TTS), recorded user speech, or AI-generated voices. For example, an attacker replays a user's recorded command near the victim to unlock restricted resources. Such attacks do not originate from a live wearer and should be rejected by liveness detection.

(2) Non-wearer Voice Attacks (Liveness Failure Expected). Voice commands are issued when the legitimate user is not wearing the device (e.g., the glasses are placed on a table). Such external speech near the device may be captured. These inputs should also be rejected by liveness detection, as they do not correspond to a valid on-body speaking condition.

(3) Advanced Spoofing with User Voice (Liveness Failure Expected). An attacker obtains recordings of the target user's voice and attempts to reproduce them using advanced playback systems or humanoid simulators. While such attacks may mimic voice content, they cannot replicate the coupled acoustic characteristics of human speech production and anatomy. Therefore, they should be rejected by liveness detection.

(4) Device Theft and Impersonation (Authentication Failure Expected). An attacker gains physical access to the smart glasses and wears the device (e.g., a co-worker attempting to access unauthorized data). In this case, the attacker produces live speech while wearing the device, and thus may pass liveness detection. However, the attacker's voice characteristics differ from the legitimate user, and the system should reject the attempt during the authentication stage.

Attacker Capability Assumptions. We adopt the following assumptions regarding attacker capabilities: **(a)** The attacker cannot compromise or manipulate the enrollment process or stored templates. **(b)** The attacker may gain physical access to the device but cannot directly tamper with internal data streams or bypass the sensing pipeline. **(c)** The attacker can obtain recordings of the target user's voice (e.g., from public or environmental sources). **(d)** The attacker cannot faithfully reproduce the user's physiological acoustic characteristics, including vocal tract dynamics, bone-conducted signals, and head-related sound-field patterns.

Out-of-Scope and Limitations. We do not consider scenarios where an already authenticated device is later misused (e.g., after being left unattended). Due to the event-driven nature of voice authentication, identity is verified per command rather than continuously. Addressing this limitation requires complementary mechanisms (e.g., continuous authentication or wearing detection), which are beyond the scope of this work. More attack scenarios are discussed in Section 6.3.2.

3.2 Passive Acoustic Modality Selection and Challenges

In this section, we systematically analyze candidate passive acoustic features in terms of sensing feasibility, robustness, and device constraints, forming the basis for feature selection and the design of the AuthGlass dataset.

3.2.1 Human Articulatory System and Usable Passive Acoustic Modality. Voice-based authentication and liveness detection have been widely studied on conventional devices and various wearable platforms[83, 96]. However, the specific acoustic modalities that can be effectively exploited in this context remain unclear. To address this question, we begin with the mechanism of human speech production (Fig. 3). Airflow from the lungs excites the vocal folds, and the resulting sound source is subsequently shaped by the vocal tract into an acoustic signal with distinctive amplitude and spectral characteristics [17]. This process not only produces airborne speech but also induces coupled vibrations and propagation effects that are influenced by speaker-specific anatomical structures.

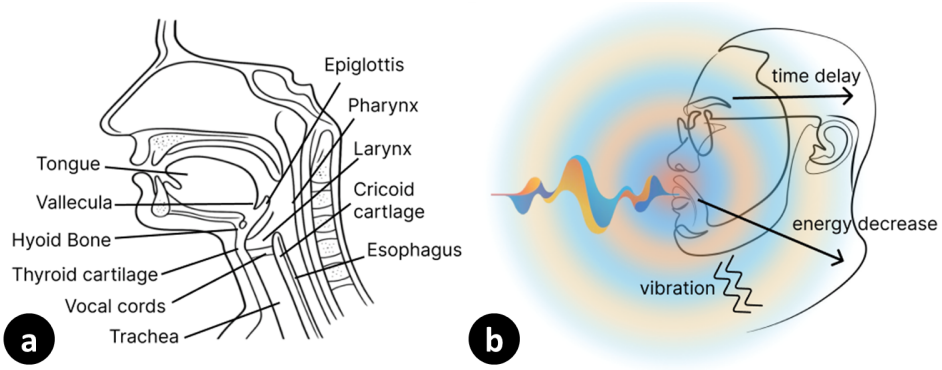


Fig. 3. Human speech production system diagram. (a) Structure of the vocal tract, including the vocal cords, tongue, and other articulatory organs. (b) Acoustic features related to speech production, including vibrations conducted through skin and bones, as well as the spatial propagation effects of the sound field such as time delays and energy attenuation.

Considering the sensing position and hardware constraints of smart glasses, a key question arises: which types of speech-related cues can be effectively leveraged for liveness detection and authentication on such devices?

3.2.2 Challenges on Leveraging Multi-Modal Acoustic Features. Research on acoustic liveness detection and speaker authentication for wearables or smart devices [15, 37, 84, 90] has primarily explored four categories of features: (i) voiceprint features extracted from airborne speech, (ii) vibration-related features arising from speech transmission through human skin, (iii) sound field features that characterize the spatial distribution of acoustic energy and phase, and (iv) near-mouth acoustic cues observed in the close-to-microphone region. Inspired by these prior feature categories, we systematically examine their applicability to smart-glasses-based sensing, and analyze the feasibility and associated challenges of transferring these modalities under the constraints of the smart glasses form factor.

First, prior studies have demonstrated that voiceprint features essential for speaker authentication can be reliably captured using a single airborne microphone [15]. This modality is naturally compatible with smart glasses, as most off-the-shelf devices already incorporate airborne microphones for voice interaction. Accordingly, a key question is where to place airborne microphones on smart glasses to effectively capture voiceprint features, which we investigate through pilot studies.

Second, prior work [37, 51] has shown that authentication systems relying solely on voiceprint features are vulnerable to spoofing attacks, motivating the use of speech-induced vibration signals as a complementary acoustic modality. Since smart glasses are in direct contact with the user's head, such vibrations can be naturally sensed. Existing studies [29, 71] have explored the use of head-worn IMUs to capture these signals; however, their limited sampling rates (typically below 200 Hz) are insufficient to cover the frequency band of speech production (up to 8 kHz). In contrast, bone-conduction microphones have been shown to capture speech-induced vibrations across the speech frequency range [37]. Prior systems such as SkullConduct [69] and NasoVoice [68] also employ bone-conduction sensing on head-worn devices with different sensor placements. Inspired by these observations, we conduct pilot studies to evaluate candidate placements of bone-conduction sensors on smart glasses, and analyze their effectiveness for capturing bone-conductive signals.

Third, as an alternative approach to strengthening voiceprint-based authentication, sound field features that capture the spatial distribution of acoustic energy and phase have been shown to be effective [90, 93]. The eyeglass frame spans a relatively large area around the head, enabling multiple microphones to be distributed

along the frame to sample the surrounding sound field. Indeed, many commercial smart glasses already integrate multiple microphones for audio capture and enhancement [7, 8], demonstrating the practical feasibility of sound field sensing. However, unlike prior work that typically assumes planar or matrix-structured microphone arrays, microphones on smart glasses are arranged along a curved, non-uniform geometry following the eyeglass frame. This structural difference challenges the applicability of existing sound-field analysis methods that rely on regular array layouts. Inspired by this observation, we conduct pilot studies to examine how such curved microphone configurations affect the extraction and usability of sound field features.

Finally, prior studies have shown that close-to-mic acoustic features, such as plosive-related pop sounds [83] and airflow-induced signals [84], can be leveraged for liveness detection and speaker authentication. However, these approaches typically rely on dedicated near-mouth microphones, and the use of such features has not been explored on eyeglass-mounted platforms. However, these approaches typically rely on dedicated near-mouth microphones, and the use of such features has not been explored on eyeglass-mounted platforms. The proximity of the lower rim of the eyeglass frame to the mouth may enable the capture of similar close-to-mic acoustic cues. Inspired by this observation, we investigate whether such features can be effectively captured using smart glasses.

To examine the usability of these features and guide the design of the AuthGlass dataset, we conduct a series of pilot studies to address the following questions:

- **Air-Conductive Feature (AC):** whether air-conduction microphones on smart glasses can reliably capture user-specific voiceprint information, and how their placement affects feature quality;
- **Vibration or Bone-Conductive Feature (BC):** whether bone-conduction microphones can be effectively integrated into smart glasses to acquire vibration signals, and how different placements influence their reliability;
- **Sound Field Feature (SF):** whether multiple air-conduction microphones on smart glasses can capture sound field information (e.g., spatial variations of acoustic energy and phase), and how such features can be effectively leveraged under non-uniform array configurations;
- **Close-to-mic Features:** whether close-to-microphone features can be captured using sensors positioned near the mouth on smart glasses.

3.3 Multi-Modal Acoustic Feature Exploration and Design

To address the questions raised in Section 3.2, we conduct four pilot studies to systematically investigate the feasibility of different passive acoustic modalities on smart glasses. Based on these studies, we identify three effective feature categories—air-conducted (AC), bone-conducted (BC), and sound-field (SF) features. We first built a proof-of-concept smart-glasses frame with microphones mounted at different positions to capture data from multiple acoustic modalities. The pilot study involved three participants, each producing six repetitions of randomly selected passphrases. After data collection, we analyze whether each modality can be used for liveness detection and authentication.

3.3.1 Study 1: AC Features. To capture air-conducted (AC) features, we mounted an air-conductive microphone at the center of the glasses frame (Fig. 4a). The microphone records audio at a sampling rate of 96 kHz. From the raw recordings, we computed Mel-spectrograms and segmented them at the phoneme level using MAUS [5]. The signal energy is primarily concentrated in the 0–8 kHz range; therefore, we applied an 8 kHz low-pass cutoff. The results indicate that the centrally placed AC microphone captures rich voiceprint-related information, with each phoneme establishing a unique time–frequency pattern. We further compared Mel-spectrograms from two users uttering the same phrase. As shown in Fig. 4b, the AC signals exhibit clear time–frequency differences across users, suggesting that AC sensing on smart glasses can capture discriminative voiceprint features for

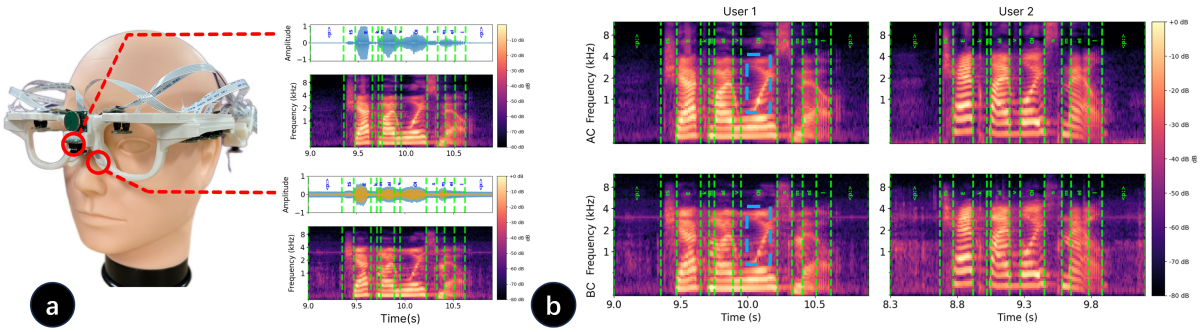


Fig. 4. This figure illustrates the air-conductive (AC) and bone-conductive (BC) microphones used for feature selection. (a) The raw audio signals and corresponding Mel-spectrograms captured by AC and BC microphones when a user speaks the phrase "check my voicemail." The utterance is segmented into phonemes. The results show that both AC and BC microphones capture rich time–frequency characteristics. (b) Comparison of air-conducted (AC) and bone-conducted (BC) signals from different users uttering "check my voicemail," visualized using a consistent color scheme. As highlighted by the blue boxes, AC and BC signals from the same user exhibit subtle frequency-related differences, whereas signals captured by either AC or BC microphones show pronounced inter-user variability.

authentication. In contrast to the central placement, microphones along the frame capture weaker and less discriminative signals.

3.3.2 Study 2: BC Features. Similar to Study 1, we utilized a voice picking unit (VPU) to capture bone-conducted (BC) signals. As the VPU requires direct contact with the human body for data acquisition, we initially placed the sensor on the nose pad [68] and on the inner side of the glasses leg [69], forming contact with the nose and head, respectively. The results show that when placed on the inner side of the glasses leg, the captured signals are significantly weaker and contain noticeable spike noise. We attribute this to poor contact conditions and potential friction with hair. Based on these observations, we selected the nose pad as the placement location for BC sensing.

After determining the sensor placement, we used the VPU (96kHz) mounted on the nose pad of the smart glasses to capture bone-conducted data, as shown in Fig. 4a. We analyzed the recorded signals using Mel-spectrograms and segmented them at the phoneme level with MAUS[5]. As shown in Fig. 4b, the VPU on smart glasses captures rich vibration data with time–frequency patterns similar to those of AC signals, while exhibiting subtle but consistent differences in the time–frequency domain, as highlighted by the blue boxes. We applied an 8 kHz cutoff to the BC signals and compared both AC and BC features across two users. Similar to AC signals, BC signals exhibit clear time–frequency differences across users, indicating that BC sensing provides complementary information that can enhance AC-based voiceprint authentication on smart glasses.

3.3.3 Study 3: SF Features. We further mounted multiple synchronized air-conducted (AC) microphones (96 kHz sampling rate) along the smart-glasses frame to capture sound field features during speech. As shown in Fig. 5a, both energy attenuation and time delays between selected AC channels and the centrally placed AC microphone vary within a short time window. The centrally placed microphone captures the highest energy, while microphones farther from the mouth exhibit reduced energy. These observations indicate that sound-energy decay and phase shifts during acoustic propagation can be captured using spatially distributed AC microphones.

To further extract these features, we segmented the recordings into short time frames. For each frame, we measured the time delay between each selected channel and the central microphone using cross-correlation, and

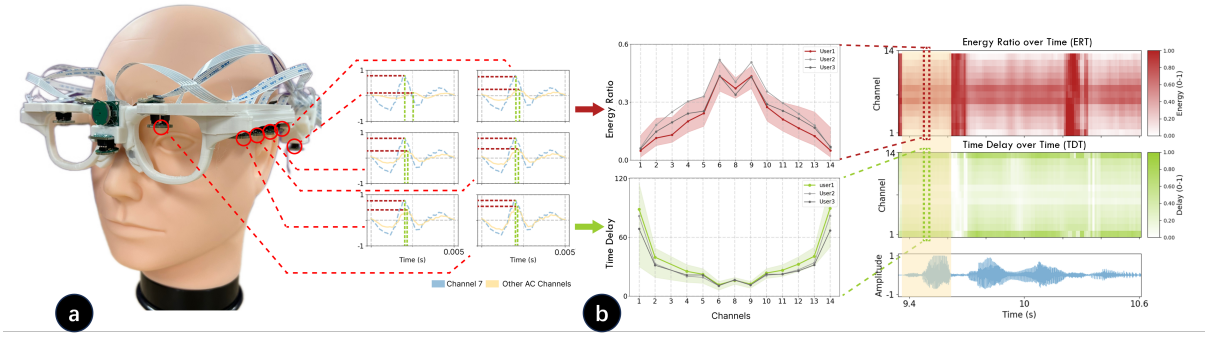


Fig. 5. This figure illustrates the computation of Energy Ratio over Time (ERT) and Time Delay over Time (TDT), along with the spatial acoustic information they capture. (a) Energy attenuation (red dashed line) and time delay (green dashed line) between a selected AC channel and the central AC channel within a short time frame. Both energy attenuation and time delay vary across different AC channels. (b) Normalized energy ratio and time delay distributions across all time frames when a user speaks the phrase "check my voicemail", forming the ERT and TDT. The left panel shows the averaged energy ratio and time delay distributions of different users, with inter-user variations highlighted by colored regions.

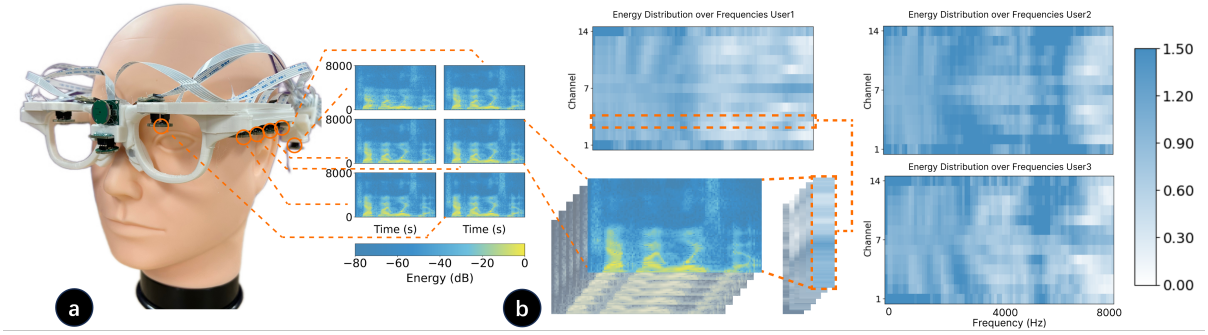


Fig. 6. This figure illustrates the computation of Energy Distribution over Frequencies (EDF) and the spatial acoustic information it captures. (a) STFT of signals recorded by microphones at different positions on the glasses. (b) For each microphone, the STFT is averaged along the time axis to obtain the energy–frequency representation. On the right, EDFs for different users uttering the phrase "check my voicemail" are shown, highlighting inter-user variations.

computed the corresponding energy attenuation, as shown in Fig. 5b. To normalize feature distributions across channels, both the time-delay and energy-ratio values were scaled and passed through a hyperbolic tangent function, yielding values in the range $[-1, 1]$. As the microphones are arranged in an array along the frame of the smart glasses, we organize them into a vector, resulting in a symmetrical pattern. Stacking consecutive time frames reveals consistent temporal patterns in both energy attenuation and time delay during speech. We denote these features as Channel Energy Ratio over Time (ERT) and Channel Time Delay over Time (TDT) respectively. As shown in Fig. 5b, averaging ERT and TDT along the time axis yields distinctive patterns across the three users, potentially related to individual head geometry.

In addition to time-domain features, we also examined sound-field characteristics in the frequency domain. Inspired by prior work on frequency-domain energy distributions [90, 93], we analyzed channel-wise energy

Table 3. Summary of dataset classes and configurations.

Category	Name	Description	Subjects	Utterances / Subject
Genuine User Data	Positive Sample	The genuine user data collected and segmented in section 3.4	42	15 (6 repeats)
Wearer-based Attack Data	Attack Setting 1 (Set 1)	The replay data generated by simulated wearers in section 3.4	42	15 (6 repeats)
Environment-based Attack Data	Attack Setting 2 (Set 2)	100 cm in front of the user.	42	15 (6 repeats)
	Attack Setting 3 (Set 3)	100 cm on right side.	42	15 (6 repeats)
	Attack Setting 4 (Set 4)	100 cm in back of the user.	42	15 (6 repeats)
	Attack Setting 5 (Set 5)	100 cm on left side.	42	15 (6 repeats)
	Attack Setting 6 (Set 6)	50 cm in front of the user.	42	15 (6 repeats)
	Attack Setting 7 (Set 7)	25 cm in front of the user.	42	15 (6 repeats)

attenuation across frequencies, as shown in Fig. 6 a. Specifically, we applied short-time Fourier transform (STFT) to the raw signals of each selected channel and averaged the resulting spectra along the time axis to obtain energy–frequency representations. We then computed the energy attenuation between each selected channel and the central microphone within specific frequency bands, and stacked the results across channels, as shown in Fig. 6. The resulting frequency-domain energy distribution exhibits user-specific patterns and inter-user variability, suggesting its potential utility for authentication. We refer to this feature as Channel Energy Distribution over Frequencies (EDF).

3.3.4 Study 4: Close-to-Mic Features. To examine whether close-to-microphone features can be captured using sensors mounted on smart glasses, we placed an airflow sensor and microphones along the lower rim of the eyeglass frame, adjacent to the lenses, to capture airflow and plosive-related pop sounds, following prior works[83, 84]. However, we did not observe discernible close-to-microphone features under this configuration.

3.4 AuthGlass Dataset

3.4.1 Design Overview. Dataset Design under Threat Models. Building upon the threat models and design objectives discussed in Section 3.1, the dataset is constructed to systematically cover four representative attack categories, including *(1) Environmental Injection Attacks*, *(2) Non-wearer Voice Attacks*, *(3) Advanced Spoofing with User Voice*, and *(4) Device Theft and Impersonation*. For attack scenarios (1) and (2), genuine voice inputs collected from the legitimate wearer are treated as positive samples, while adversarial samples are generated by replaying audio through external speakers in the environment, simulating injected or non-wearer speech. For scenario (3), adversarial samples are collected using humanoid simulators that reproduce the target user's voice, representing more advanced spoofing attempts. For scenario (4), data are collected from multiple real users, where each user may act as an attacker attempting to impersonate another enrolled user while wearing the device.

Accordingly, the dataset is organized into three major data categories: *(i) genuine user data*, *(ii) wearer-based attack data*, and *(iii) environment-based attack data*. Table 3 summarizes the detailed data composition and their correspondence to different attack scenarios. This threat-driven design enables systematic evaluation of both liveness detection and authentication under diverse and realistic adversarial conditions, as illustrated in Section 4.

Table 4. Summary of modalities and channel configurations.

Modality	Channel	Description
AC	Channel 7	AC channel mounted centrally and facing downward.
BC	Higher-energy channel (15, 16)	Bone-conduction channel with higher signal energy selected as the representative BC input.
SF	Channels [1–6], [8–14]	All remaining AC channels that can function as complementary channels in comparison with the central AC channel.

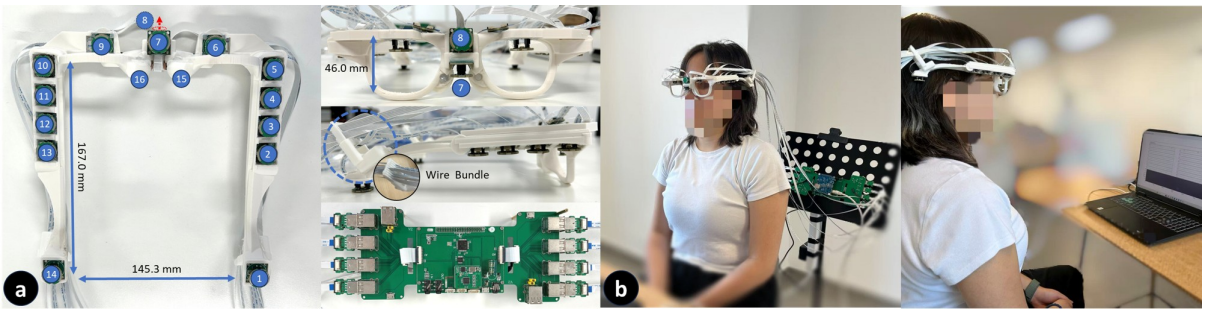


Fig. 7. (a) Structural and hardware overview of the prototype. The system comprises 14 air-conduction microphones and 2 bone-conduction microphones. Each microphone is connected to the data collection board via flexible flat cables, and the temple is designed with a wire bundle, which can reduce crosstalk between channels. b) Illustration of a user wearing the smart glasses prototype.

Multi-Modal Acoustic Data Collection. Based on the validation of acoustic modalities in Section 3.3, the dataset captures three complementary acoustic modalities, including *air-conducted (AC)*, *bone-conducted (BC)*, and *sound-field (SF)* signals. These modalities correspond to different physical propagation paths of speech, including airborne transmission, vibration through human tissues, and spatial acoustic distribution around the head. To enable synchronized acquisition of these modalities, we implement a DIY smart-glasses platform (Section 3.4.2) equipped with 16-channels microphone array consisting of both air-conduction and bone-conduction sensors. Table 4 summarizes the definition of each modality and its corresponding sensing channels.

3.4.2 Hardware Implementation. To support multi-modal data acquisition, we implement a custom DIY smart-glasses prototype, as illustrated in Fig. 7a. The prototype integrates 14 omnidirectional air-conduction (AC) microphones (Infineon IM73D122V01 [3]) and two bone-conduction (BC) voice picking units (VPU14AA01 [9]), with overall dimensions of 167.0 mm (length) and 145.3 mm (width).

To enable spatial acoustic sampling for sound-field analysis, the AC microphones are distributed across the eyeglass frame: five on each temple, one above each lens, and two at the center (facing downward and forward, respectively), providing diverse spatial coverage.

Guided by pilot study insights, one BC sensor is embedded in each nose pad to capture vibration signals while mitigating degradation caused by asymmetric skin contact. All sensors are connected via bundled flexible flat

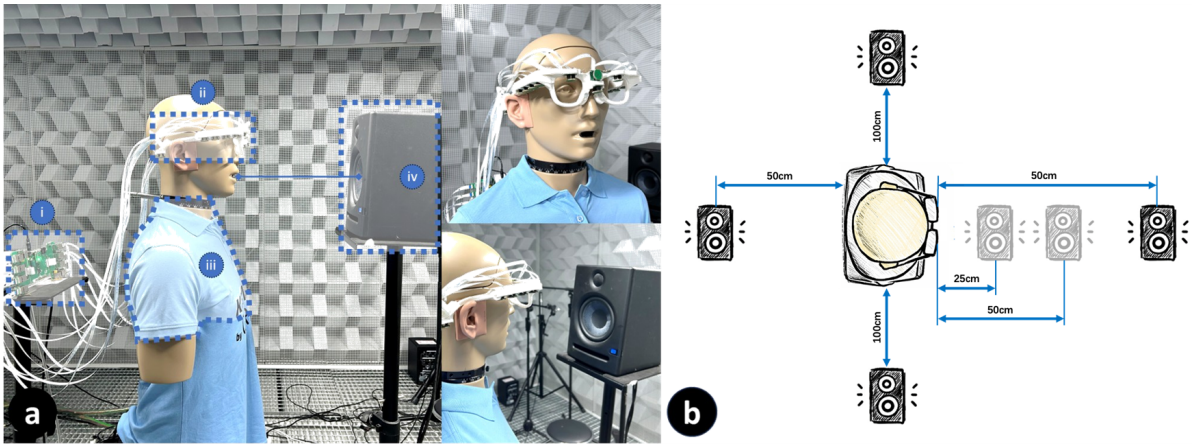


Fig. 8. (a) The environment setting of the voice attack. i is the data collection board and ii is the smart glasses frame. All microphones are connected via bundled flexible flat cables to a data collection board. iii is the torso–mouth simulator and iv is the loudspeaker. In the experiment, two types of replay sample were applied: loudspeaker replay and a torso–mouth simulator replay which replicates the exhalation process and torso vibrations during speaking. (b) In the replay attack, the loudspeaker was placed at different positions and distances from the smart glasses to simulate various attack situations.

cables (FFCs) to a data collection board [12], enabling synchronized acquisition of 16 channels at 96 kHz. The recorded data are streamed to a PC via TCP for subsequent processing.

3.4.3 Data Collection. We then collected data using our smart-glasses prototype for positive sample and two other attack categories, described in Table 3. The study protocol was approved by the Institutional Review Board (IRB). All data are anonymized for dataset publication.

Genuine User Data (Positive Sample). We recruited 42 participants (14 female and 28 male), aged 18–42 (mean 24.6 ± 4.3), from a university campus. Detailed participant demographics are provided in Appendix A.2.. All participants provided informed consent, permitting the use and publication of their anonymized data for research purposes. Data collection was conducted in a quiet conference room to minimize environmental noise. Participants were asked to wear our glasses prototype and were seated in a natural posture and instructed to avoid excessive head or body movements during recording. As illustrated in Fig. 7(b), each participant uttered a predefined set (a total of 15) of voice passphrases (see Appendix A.1) at three distinct volume levels, ranging from soft speech (e.g., speaking to a nearby listener) to loud speech (e.g., projecting to a listener approximately 5 meters away). Each passphrase–volume combination was repeated twice, resulting in six repetitions per sentence. Short rest intervals were provided between passphrases to reduce vocal fatigue. All utterances were captured by the glasses-mounted microphones and manually reviewed to confirm successful recording. The audio was subsequently segmented using MAUS [5], with the resulting segments manually verified for accuracy.

Wearer-based Attack Data (Attack Setting 1). To collect wearer-based attack data, we conduct experiments in a sound isolation room, as shown in Fig. 8a. The smart glasses are mounted on a GRAS 45BC KEMAR torso-mouth simulator [2], which replays genuine utterances through its built-in speaker. The playback volume is manually calibrated to match the sound pressure levels observed during genuine data collection. All genuine samples are replayed via the simulator's mouth and recorded by the glasses.

Environment-based Attack Data (Attack Setting 2–7). To collect environment-based attack data, we use the torso-mouth simulator as a mannequin to emulate a stationary user. Genuine utterances are replayed using a high-quality loudspeaker (PreSonus Eris E5 [1]) positioned at the same height as the mannequin, with playback volume manually calibrated. To model diverse attack source locations, the speaker is placed at six distinct positions around the mannequin, corresponding to different directions and distances (Fig. 8b). For each position, all genuine samples are replayed and recorded by the glasses.

3.4.4 Statistics Analysis. After data collection, we obtained AuthGlass datasets that consists of three categories of data. Each sample consisting of 16 synchronized audio channels at 96 kHz. For *Genuine User Data*, the dataset contains a total of $42 \times 6 \times 15 = 3{,}780$ utterances, corresponding to 15 passphrases (length from 2 to 5 words) spoken at three volume levels by each participant. For *Wearer-based Attack Data (Attack Setting 1)*, we collected 3,780 attack samples, matching the size and content of the genuine dataset. For *Environment-based Attack Data (Attack Setting 2–7)*, each genuine utterance was replayed and recorded at six distinct speaker locations, yielding $3{,}780 \times 6 = 22{,}680$ attack recordings. All recordings were segmented using MAUS [5] and temporally aligned and segmented, followed by manual verification of alignment and segmentation quality.

While the dataset provides comprehensive multi-modal acoustic recordings, we note that the participant pool is still relatively limited in size and demographic diversity (e.g., primarily recruited from a university population), which may not fully capture broader variations in speech characteristics. A more detailed discussion of these limitations is provided in Section 6.3.1. The complete dataset, including raw multichannel recordings, segmentation annotations, and the design files of the smart-glasses prototype, has been publicly released, enabling future work to reproduce, extend, and enrich the dataset with more diverse participants and scenarios.

4 BENCHMARKING VOICE LIVENESS DETECTION AND AUTHENTICATION ON SMART GLASSES

In this section, we first design a set of benchmark tasks based on the collected data to evaluate voice liveness detection and authentication on smart glasses (Section 4.1). Then, we select seven liveness detection methods and three authentication methods to cover a range of evaluation scenarios (Section 4.2). We further present our proposed lightweight liveness detection and authentication system (Section 4.3). Finally, we report experimental results across all benchmark tasks, showing that our method achieves state-of-the-art performance (Section 4.4) with conclusion (Section 4.5)

4.1 Benchmark Tasks Design

We define four benchmark tasks (two for liveness detection and two for authentication) with task description and evaluation metrics, following the authentication and threat model in section 3.1.

Liveness Detection. aims to determine whether the input audio originates from a genuine glasses wearer. Ideally, it should reject all inputs that do not correspond to a live user wearing the device. In this work, we focus on three representative adversarial scenarios introduced in Section 3.1, including *Environmental Injection Attacks*, *Non-wearer Voice Attacks*, and *Advanced Spoofing with User Voice*. These scenarios are instantiated in our dataset through *wearer-based attack data* and *environment-based attack data* (Table 3). In addition, liveness detection approaches that rely on identifying specific attack devices or synthesized signals may fail to generalize to unseen attacks. This limitation is particularly critical in real-world scenarios, where attacks may originate from arbitrary and previously unknown devices. To investigate different methods' performance on this issue, we further design a task to examine the generalization of Liveness Detection to unseen attacks. For each task, we use accuracy at the EER operating point, i.e., $1 - \text{EER}$ as evaluation metric.

4.1.1 Task 1: Liveness Detection Accuracy. (a) Data Split. We construct seven evaluation sets by pairing positive samples with attack samples from each attack type (Attack Setting 1–7), where each set contains an equal number of positive and attack samples with one-to-one correspondence. A subject-independent split is adopted, such that samples from a subset of subjects are used for training and the remaining subjects for testing, with no subject overlap. Positive and attack samples are labeled as 0 and 1, respectively.

(b) Method. For each split, the liveness detection model is trained on the training set and evaluated on the test set. The trained model predicts the labels of both positive and attack samples in the test data. Performance is measured using overall classification accuracy, and results are averaged across multiple splits.

4.1.2 Task 2: Generalization of Liveness Detection to Unseen Attacks. (a) Data Split. To evaluate cross-attack generalization, we construct seven datasets by pairing positive samples with attack samples from each attack type (Attack Setting 1–7), following the same protocol as in Task 1. A subject-independent split is adopted, where samples from a subset of subjects are used for training and the remaining subjects for testing. The same subject split is consistently applied across all datasets to ensure fair comparison.

(b) Method. For each split, the model is trained on one dataset (corresponding to a specific attack type) using the training subjects, and evaluated on the test portions of the remaining datasets (corresponding to unseen attack types) using the test subjects. For example, the model can be trained on (Positive, Set 1) samples from the training subjects, and then tested on (Positive, Set 2) to (Positive, Set 7) samples from the test subjects. The model predicts labels for both positive and attack samples, and performance is measured by classification accuracy. The final result is obtained by averaging accuracy across all training–testing combinations for each split.

Authentication. aims to distinguish voice inputs from different genuine users by accepting the enrolled user while rejecting all others. Accordingly, we first define a primary authentication task to evaluate the baseline performance of different methods. In practical scenarios, an authentication system should support fast enrollment with minimal user effort, as requiring repeated utterances for each new voice command is impractical. To reflect this constraint, we further design a task to evaluate authentication performance for cross-utterance scenarios.

4.1.3 Task 3: Authentication Accuracy. (a) Data Split. To evaluate authentication performance, we adopt a subject-independent split, where samples from a subset of subjects are used for training and the remaining subjects for testing. Only positive samples are used, as the task focuses on distinguishing among genuine users. Each sample is labeled with a subject ID (1–42) and an utterance ID (1–15).

(b) Method. For each split, the authentication model is trained using the training set with corresponding subject and utterance labels. During testing, one sample per utterance is randomly selected from the test subjects to form the enrollment set, resulting in 15 enrollment samples per subject. This selection is kept consistent across all methods. The remaining samples are used as the test set. The model predicts the subject identity for each test sample, and performance is measured using classification accuracy.

4.1.4 Task 4: Cross-Utterance Authentication Performance. (a) Data Split. To evaluate cross-utterance authentication performance, we adopt the same subject-independent split as in Task 3. Only positive samples are used, with each sample annotated by a subject ID (1–42) and an utterance ID (1–15).

(b) Method. For each split, the authentication model is trained on the training set with corresponding subject and utterance labels. During testing, we simulate a cross-utterance enrollment scenario by selecting a subset of utterances per subject to form the enrollment set. Specifically, the 15 utterances are partitioned into multiple non-overlapping groups, and each group is used as enrollment while the remaining utterances form the test set. The selection of utterance groups is fixed and consistent across all methods to ensure fair comparison. For each configuration, the model predicts the subject identity of test samples, and authentication accuracy is computed. The final performance is obtained by averaging results across all configurations for each split.

Table 5. Comparison of voice liveness detection methods.

Method	AC	BC	SF	Feature	Model / Method
VOID [15]	✓	-	-	Manually Defined Features	Traditional ML
He et al.(2024) [36]	✓	-	-	Manually Defined Features	CNN
Boneauth [51]	✓	✓	-	MFCC	Cosine Similarity
Eve Said Yes [37]	✓	✓	-	STFT with pooling	Temporal Similarity
Li et al.(2021) [53]	✓	-	✓	STFT with phase	CNN
Cafield [90]	-	-	✓	Fieldprint	GMM
Voshield [93]	-	-	✓	Sound Field Dynamics	CNN
Ours	-	-	✓	TDT and ERT	Cosine Similarity

Table 6. Comparison of voice authentication methods.

Method	AC	BC	SF	Feature	Model / Method
Park et al. (2025) [62]	✓	-	-	MFCC	LSTM
Eve Said Yes [37]	-	✓	-	CQT	CNN with AL [1]
Cafield [90]	-	-	✓	Fieldprint	GMM
Voshield[2][93]	-	-	✓	Sound Field Dynamics	CNN
Ours	✓	✓	✓	MFCC, TDT, ERT and EDF	CNN with AL [1]

[1] indicates methods employing adversarial learning.
[2] the method originally do not support authentication.

4.2 Existing Method and Adjustments to Multi-channel Acoustic Data

We selected seven representative methods for voice liveness detection and four methods for authentication, spanning AC, BC, and SF acoustic modalities. We further modified these methods to accommodate multiple features, aligning their input formats for systematic comparison.

Voice Liveness Detection Methods. We selected seven representative voice liveness detection methods, as summarized in Table 5. These methods span AC, BC, and SF acoustic modalities and employ diverse feature representations and detection models. To enable a fair comparison, we align all methods in terms of input modalities. Methods and our corresponding modifications are introduced in Appendix A.3.

Voice Authentication Methods. We selected four voice authentication methods, as summarized in Table 6. Similar to the liveness detection methods, all authentication methods are modified to accommodate multi-channel inputs for a fair comparison. Methods and our corresponding modifications are introduced in Appendix A.3.

4.3 Our Method: AuthG-Live & AuthG-Net

In this section, we briefly introduce our liveness detection method, AuthG-Live, and our authentication model, AuthG-Net.

4.3.1 Feature Extraction. For feature extraction, log-Mel spectrograms were computed for both air-conduction (AC) and bone-conduction (BC) channels. BC signals were preprocessed to suppress low-frequency interference,

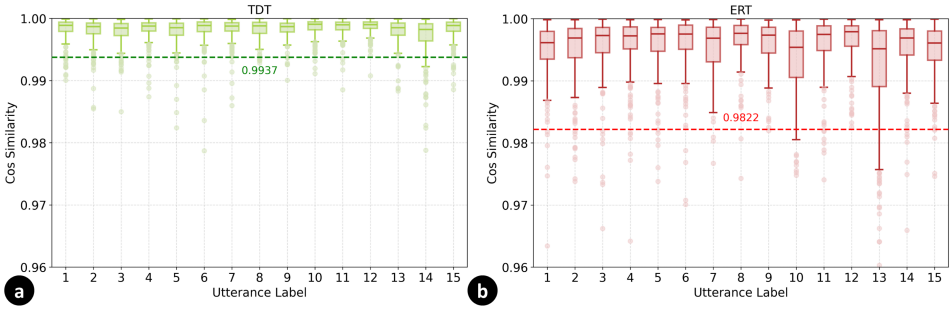


Fig. 9. This figure shows the distribution of cosine similarities between the unified patterns and the corresponding aTDT and aERT across all genuine samples for the 15 utterances. Each box in the boxplot represents the interquartile range (25th to 75th percentile), with the line inside indicating the median. Scattered points denote outliers, and the dashed line marks the 3rd percentile of the genuine similarity distributions. In (a), the TDT cosine similarity between aTDT and the unified patterns defines a threshold of 0.9937, whereas in (b), the ERT yields a threshold of 0.9822.

while all signals were resampled to a unified sampling rate prior to transformation. The Mel spectrograms were generated using a fixed set of Mel filter banks within a predefined frequency range.

To characterize sound field (SF) properties, we extracted three complementary features: Energy Ratio over Time (ERT), Time Delay over Time (TDT), and Energy Distribution over Frequency (EDF). These features capture spatial sound energy attenuation, inter-channel propagation delays, and spectral energy distribution across microphones, following the design principles described in Section 3.3.

ERT quantifies relative energy differences across channels by computing band-limited spectral energy and normalizing it with respect to reference AC channel (Channel 7), followed by scaling and normalization to ensure numerical stability. TDT estimates relative time delays between each channel and the reference channel via cross-correlation. A confidence-aware weighting mechanism is applied to suppress unreliable delay estimates, and the resulting values are further normalized for stability. EDF characterizes the frequency-wise energy distribution for each channel by aggregating spectral representations over time. The distributions are normalized with respect to the reference channel to preserve relative spectral characteristics while reducing sensitivity to overall amplitude variations. Detailed implementation and parameter settings are provided in Appendix B.1.

4.3.2 AuthG-Live: Attack Device Independent Voice Liveness Detection. Most existing voice liveness detection methods learn discriminative patterns from attack samples observed during training [15, 36]. As a result, their generalization performance strongly depends on the diversity of attack types seen during training. When confronted with unseen spoofing techniques or replay configurations, detection accuracy often degrades, leading to reduced robustness in real-world deployments.

In contrast, AuthG-Live is a generalizable voice liveness detection framework that leverages physical sound propagation characteristics rather than attack-specific artifacts. It models the near-source spatiotemporal sound field by extracting Energy Ratio over Time (ERT) and Time Delay over Time (TDT) from multi-channel air-conducted signals, capturing intrinsic source–vocalization relationships that are difficult for replay or synthesized attacks to replicate. Notably, AuthG-Live is trained solely on genuine speech data, avoiding the need to enumerate diverse attack scenarios and enabling robustness to unseen spoofing methods. Furthermore, these features can be extracted in a streaming manner, supporting rapid attack detection (Section 5.5).

To this end, we incorporate TDT and ERT, as these features reflect the spatial origin of the received signals. Specifically, TDT and ERT are averaged along the time axis for each channel to form mean templates, denoted

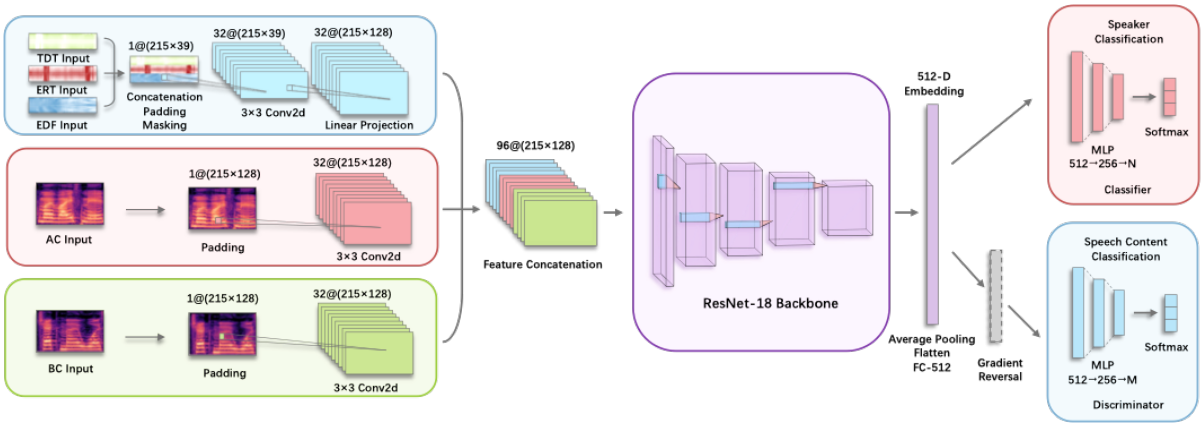


Fig. 10. The figure demonstrate the architecture of AuthG-Net. The blue/red/green blocks on the left denote the SF/AC/BC inputs, which are encoded by lightweight CNN heads and then stacked and fed into a ResNet-18 backbone (purple block) to produce a 512-D embedding. In the domain-adversarial learning (DAL) branch, the model is optimized with two objectives: speaker classification (red) and speech-content classification (blue). The speaker classifier is trained directly on the embedding, whereas the passphrase classifier is connected via a Gradient Reversal Layer (GRL), which reverses gradients during backpropagation to encourage content-invariant embeddings.

as aTDT and aERT, which summarize sound propagation delay and energy attenuation patterns. For liveness detection, the aTDT and aERT from genuine user samples are averaged to obtain unified aTDT and aERT patterns. These two unified patterns represents the typical acoustic propagation behavior of genuine speech. Given an input signal, its corresponding aTDT and aERT are computed and compared with the unified patterns using cosine similarity. A similarity score exceeding a predefined threshold indicates that the signal is likely emitted from a real human speaker.

To determine the decision threshold, we compute the cosine similarity between the unified patterns and the corresponding aTDT and aERT extracted from all genuine samples. The results exhibit consistently high similarity across utterances and subjects. Accordingly, the thresholds are set based on the 3rd percentile of the genuine similarity distributions, allowing 97% of genuine samples to be accepted. This data-driven criterion ensures a high true acceptance rate for genuine speech while effectively rejecting most spoofing attacks. As illustrated in Fig. 9, the resulting similarity thresholds are 0.9937 for aTDT and 0.9822 for aERT. An input signal is classified as live only if both similarity scores exceed their respective thresholds.

4.3.3 AuthG-Net: Passphrase-Independent Voice Authentication. AuthG-Net is a lightweight multi-modal neural network designed for user authentication on smart glasses, as illustrated in Fig. 10. It takes air-conducted (AC), bone-conducted (BC), and sound-field (SF) features as input, and learns a unified user representation through multi-modal fusion.

Specifically, each modality is first processed by a dedicated encoder to extract modality-specific features. These features are then fused and fed into a shared backbone network to produce a compact embedding representing the user. To enable generalization across different utterances, we adopt a domain-adversarial learning (DAL) framework, where the model is jointly optimized for speaker classification while suppressing speech-content information via an adversarial objective. This design encourages the learned embedding to capture user-specific characteristics that are invariant to the spoken content.

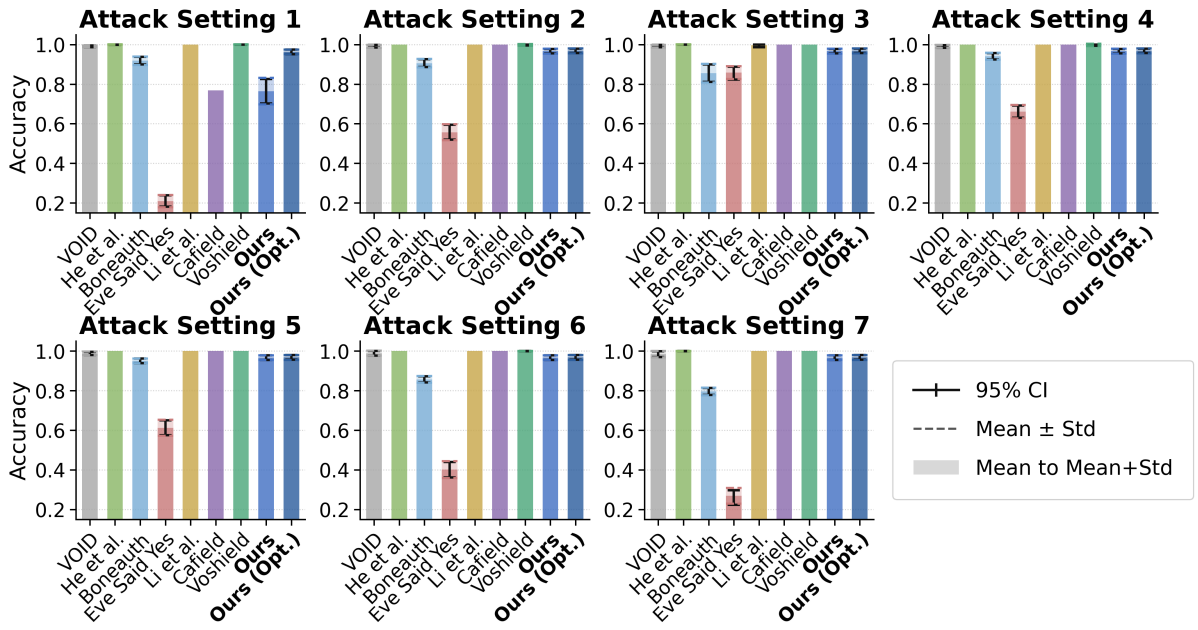


Fig. 11. Result for benchmark task 1. Note: (1) optimization (Opt.) for our method is described in section 5.3; (2) Since the other methods are trained and evaluated using known attack types, they may be prone to overfitting to specific attacks. In contrast, our method determines the decision threshold using only genuine samples and demonstrates superior performance under unseen attack scenarios, as discussed in Section 4.4.2.

During training, the model is supervised with both subject and utterance labels to learn discriminative and utterance-independent representations. During inference, the extracted embedding is used as the user template, and authentication is performed by comparing embeddings using cosine similarity. Detailed network architecture, feature processing, and training configurations are provided in Appendix B.2.

4.4 Results and Analysis on Benchmark Tasks

In this section, we provide result and analysis on the benchmark tasks defined in section 4.1

4.4.1 Task 1: Liveness Detection Accuracy. We first evaluate all selected liveness detection methods on Task 1 using 7-fold cross-validation, with results summarized in Fig. 11. The result for original baseline methods are also included for reference (Appendix C.1). Our method, leveraging all SF modalities, achieves high accuracy for Attack Setting 2 to 7, demonstrating strong robustness against environment-based replay attacks.

However, for Attack Setting 1 (wearer-based attack), the accuracy drops to 76%, indicating that the attacker can partially mimic the spatial characteristics of a genuine user. As analyzed in Section 5.3, this degradation arises from channels that are insensitive to sound source position when relying on time delay and energy ratio cues. By selecting only spatially sensitive channels, the accuracy improves significantly to over 96%.

Although several baselines report near-perfect accuracy, this performance likely stems from learning device-specific characteristics of the attacking loudspeaker rather than intrinsic liveness cues. Since each experiment uses the same attack type for both training and testing, models can exploit device-related features. As shown in later generalization experiments, such methods degrade substantially when evaluated on unseen attacks.

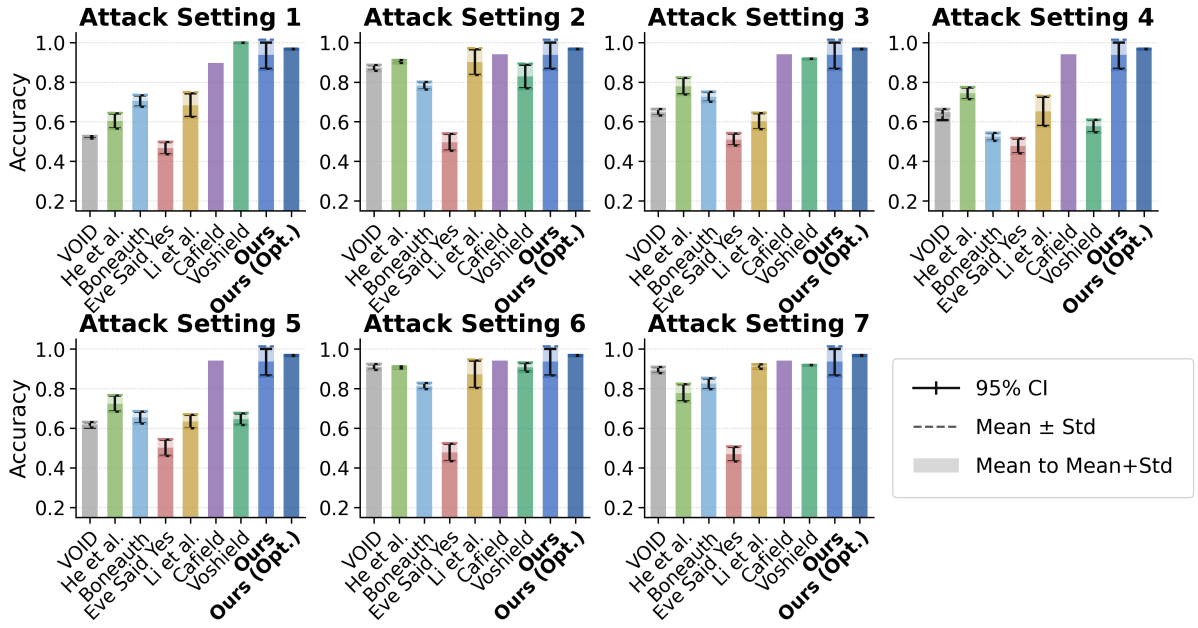


Fig. 12. Result for benchmark task 2. Note: (1) optimization for our method is described in section 5.3; (2) As our method do not require attack samples for training, we average the liveness detection accuracy on all attacks.

Finally, *Eve Said Yes*, which relies on loose AC–BC similarity, fails to distinguish genuine and attack samples. This is because airborne signals can also induce correlated vibrations on BC sensors, and similar AC–BC patterns can be reproduced in simulated wearer attacks, making reliable discrimination difficult.

4.4.2 Task 2: Generalization of Liveness Detection to Unseen Attacks. We further evaluate generalizability through Task 2 by training on one attack setting and testing on others with a 7-fold cross-validation. Results are shown in Fig. 12. Noticeably, all baseline methods suffer performance degradation when facing unseen attacks, with those relying on attacker-device-specific features (e.g., VOID and He et al.) degrading the most.

By enforcing a strict AC–BC similarity metric, Boneauth retains partial attack detection capability, but still drops from around 90% to 75%, as its thresholds (determined by EER) vary across attack settings and become inconsistent. *Eve Said Yes* completely fails on unseen attacks, yielding around 50% accuracy. In contrast, SF-based methods (Li et al., Cafield, and Voshield) exhibit less degradation, suggesting that sound field features support cross-attack detection since they are independent of attack-device characteristics.

Among Li et al., Cafield, Voshield, and our method, ours (with optimization) achieves the highest and most stable accuracy across all seven attack types, while the others degrade under one or more attacks. Li et al., which uses original STFT spectrograms, inevitably captures device-specific features and shows the largest degradation (maximum over 30%) compared to Cafield and Voshield. Although the latter two rely only on SF-related features, their CNN-based models still require both genuine and attack samples during training, limiting generalization. In contrast, our method requires no attack samples and relies on static similarity-based comparison, enabling robust performance on unseen attacks.

4.4.3 Task 3: Authentication Accuracy. We evaluate authentication performance in Task 3 using 7-fold cross-validation, with results shown in Fig. 13a. Our method achieves the best performance among all compared

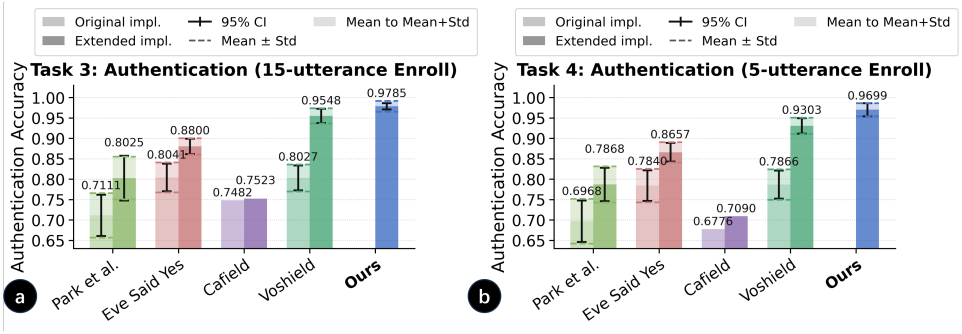


Fig. 13. (a) Result for benchmark task 3. (b) Result for benchmark task 4.

approaches, demonstrating that the proposed feature extraction strategy and AuthG-Net can effectively capture user-related characteristics.

Park et al. employ an LSTM with randomized embedding generation, but their method is designed for single-user scenarios and fails to learn discriminative embeddings across multiple users, resulting in low accuracy (80.25%). Although *Eve Said Yes* supports embedding extraction, it struggles to effectively fuse multi-modal features, yielding only limited improvement (from 80.41% to 88.00%). Cafield utilizes fieldprint SF features but performs poorly under both original and extended implementations, likely due to the limited capacity of its GMM-based classifier. In contrast, Voshield, based on SFD features, achieves performance close to ours, improving by approximately 13% over its original implementation, which indicates the effectiveness of CNN-based architectures in extracting discriminative acoustic representations.

4.4.4 Task 4: Cross-Utterance Authentication Performance. We evaluate cross-utterance authentication performance in Task 4 using 7-fold cross-validation. As shown in Fig. 13 b, our method achieves the best performance, with only 1% degradation compared to Task 3, indicating its ability to learn utterance-independent user features, largely due to DAL.

In contrast, the other methods exhibit varying degradation, with Cafield and Voshield being most affected (5% and 3%, respectively). This is likely because their sound field features (fieldprint and SFD) are closely tied to utterance content, as different phonemes produce distinct sound field characteristics. By separately modeling sound field features in both frequency (EDF) and time domains (ERT and TDT), our method achieves stronger generalization across utterances. For Park et al. and *Eve Said Yes*, although their overall authentication performance is relatively limited, the degradation under cross-utterance settings is smaller (around 1%–2%), suggesting partial robustness in extracting utterance-independent features.

4.5 Conclusion

In conclusion, our method achieves state-of-the-art performance in both voice liveness detection and authentication. It demonstrates strong generalization capability for liveness detection against unseen attacks, as well as robust performance in cross-utterance authentication scenarios.

5 PRACTICAL EVALUATIONS

Since our method is developed using high-quality equipment for theoretical analysis, it is necessary to validate its effectiveness under practical constraints. To this end, we conduct a series of ablation studies. Specifically, we evaluate the impact of reduced sampling rates compared to the original 96,kHz setting (Section 5.1), limited

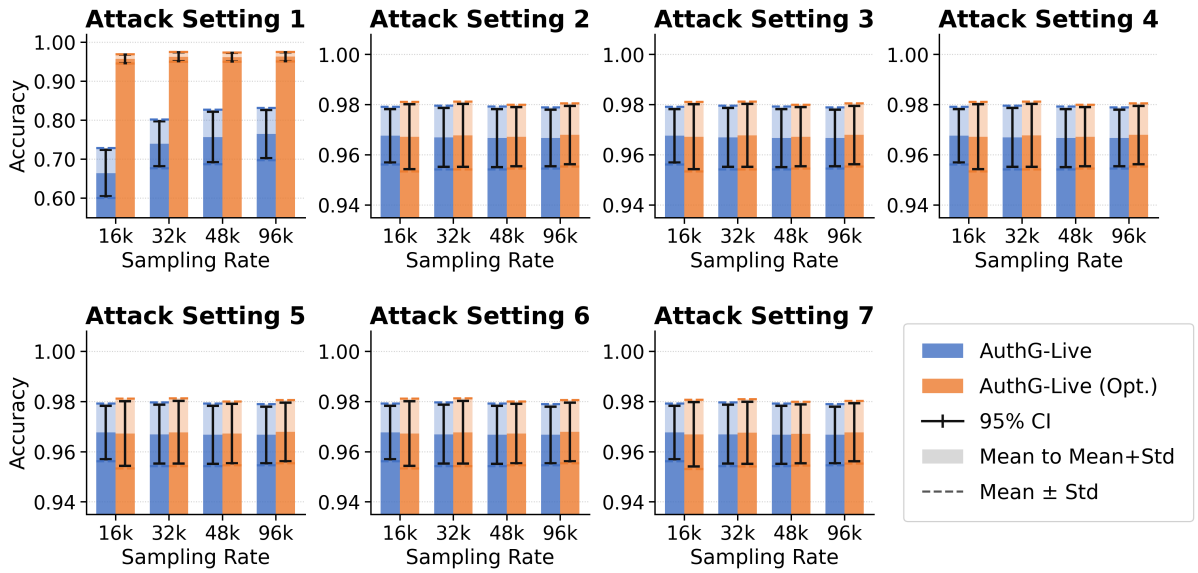


Fig. 14. Liveness detection accuracy under different sampling rates. Note: Optimization is described in Section 5.3.

acoustic modalities instead of jointly using AC, BC, and SF (Section 5.2), and a reduced number of SF channels given the impracticality of deploying 14 AC microphones on smart glasses (Section 5.3). We further simulate microphone layouts of off-the-shelf products, including Rokid [8] and Meta Ray-Ban [7], to evaluate both liveness detection and authentication under realistic settings (Section 5.4). Finally, we conduct a simulated live study to assess liveness detection and authentication under two levels of system simplification, focusing on latency, energy consumption, and performance (Section 5.5).

5.1 Limited Sampling Rate

Our prototype records audio at 96 kHz, whereas commercial smart glasses typically operate at around 16 kHz. To evaluate the impact of sampling rate, we test four settings: 16 kHz, 32 kHz, 48 kHz, and 96 kHz. Lower-rate recordings are simulated by applying FIR-based low-pass filtering followed by uniform downsampling on each channel. While reduced sampling rates have minimal effect on Mel-spectrogram computation, they can degrade sound field modeling due to reduced temporal resolution. To mitigate this, the downsampled signals are linearly interpolated to recover a denser temporal representation while preserving bandwidth constraints. The processed signals are then used for feature extraction and evaluation. We follow Benchmark Task 1 (liveness detection) and Task 3 (authentication) in Section 4.1, using 7-fold cross-validation repeated twice. All AC, BC, and SF modalities are included. Results are shown in Fig. 14 (liveness detection) and Fig. 15a (authentication).

5.1.1 Results and Discussion. As shown in Fig. 14, the original AuthG-Live with all SF channels achieves consistently high liveness detection accuracy on attack settings 2–7 (all above 96%), with no noticeable degradation as the sampling rate decreases. However, it performs poorly on attack setting 1, where accuracy drops from 76% to 66% as the sampling rate decreases from 96 kHz to 16 kHz. Similar with analysis in Section 4.4.1, this issue can be addressed by the channel optimization introduced in Section 5.3, which improves robustness by focusing on position-sensitive SF channels. With this optimization, AuthG-Live achieves consistently high accuracy (above

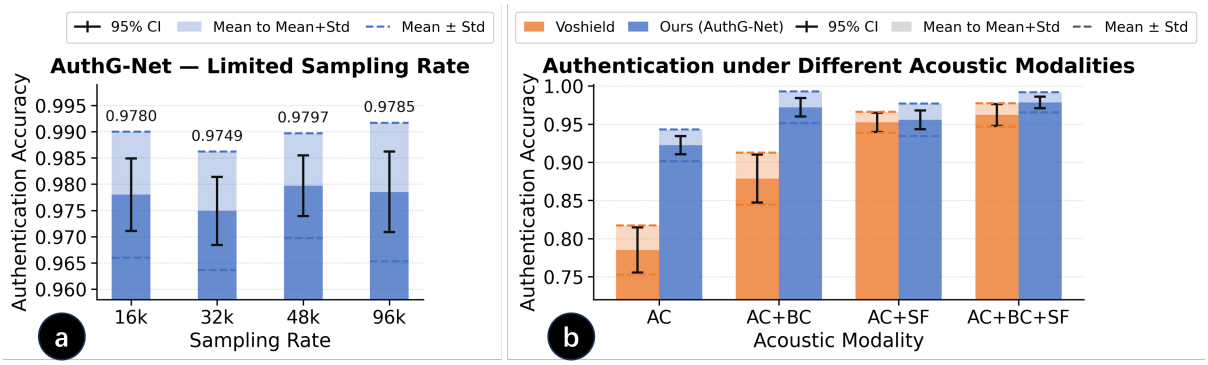


Fig. 15. (a) Authentication accuracy under different sampling rates. (b) Authentication accuracy under different input modalities.

95%) across all seven attack settings under different sampling rates; For authentication, Fig. 15a shows that AuthG-Net maintains strong performance despite reduced sampling rates, achieving accuracy above 97.93%.

5.2 Limited Acoustic Modalities

Although our prototype supports AC, BC, and SF modalities, many commercial smart glasses lack BC sensors or sufficient microphones for SF capture. It is therefore important to evaluate whether our method remains effective under limited modality configurations.

We consider four representative combinations: AC only, AC+BC, AC+SF, and AC+BC+SF. Since AuthG-Live relies solely on SF and cannot operate without SF inputs, we evaluate only authentication performance using AuthG-Net, following Benchmark Task 3 (Section 4.1.3) with 7-fold cross-validation repeated twice. The sampling rate is fixed at 96 kHz. To support different modality inputs, we implement three additional variants of AuthG-Net by removing the corresponding modality encoders before feature fusion, leveraging its modular design. For example, the AC+SF variant retains only AC and SF encoders, while the AC-only variant reduces to a residual CNN with DAL.

We also implement Voshield for comparison, as it achieves the second-best performance in Section 4.4.3. Similarly, three variants of Voshield are constructed to support reduced modality inputs following the same modification principle (Section 4.2).

5.2.1 Results and Discussion. As shown in Fig. 15b, reducing available acoustic modalities degrades authentication performance for both our method and the baseline, confirming that BC and SF provide effective complementary information to AC. Our method consistently outperforms the baseline when combining AC with either SF or BC, demonstrating that AuthG-Net can effectively extract and fuse user-related features across modalities. Notably, even with AC-only input, our method achieves 91.04% accuracy, still surpassing the baseline, indicating strong robustness under limited modality conditions.

5.3 Limited Input Channels and Optimization on Channel Selection

After validating generalization across modality combinations, we further evaluate performance under limited input-channel settings, motivated by the practical constraints of microphone availability in smart glasses (e.g., power, weight, and comfort). We investigate whether reduced microphone configurations can achieve competitive or improved performance in both liveness detection and authentication. Specifically, we examine the number and

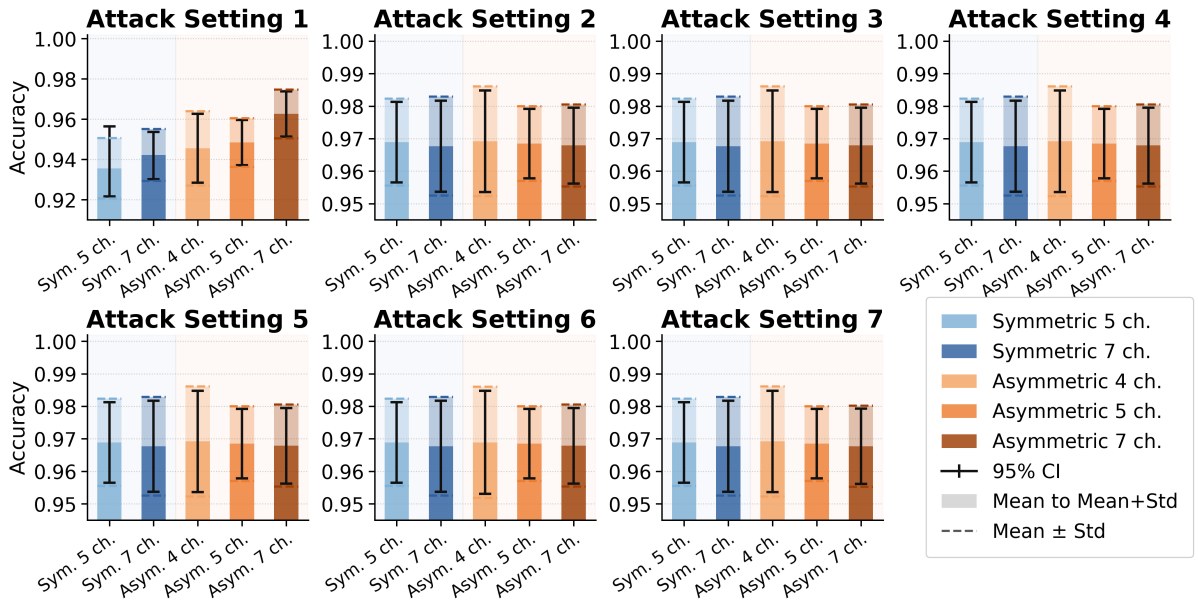


Fig. 16. Liveness detection accuracy under different channel inputs.

placement of air-conduction (AC) microphones, as well as the impact of incorporating bone-conduction (BC) microphones.

We consider three representative AC configurations reflecting real devices: 4-channel (Rokid [8]), 5-channel (Ray-Ban Meta [7]), and a feasible 7-channel setup based on our prototype. For each configuration, we optimize microphone placement for liveness detection using AuthG-Live, focusing on symmetric layouts commonly used in practice while also exploring asymmetric alternatives. During optimization, Channel 7 is fixed as the AC input, and all combinations of the remaining AC channels are evaluated following Benchmark Task 1 (Section 4.1). The best-performing symmetric and asymmetric layouts are identified for each configuration.

Using these optimized layouts, we then evaluate authentication performance with and without BC inputs following Benchmark Task 3 (Section 4.1) with 7-fold cross-validation repeated twice. To support varying SF channel numbers, we adjust the input dimension of the SF encoder projection layer in AuthG-Net, while optional BC inputs are incorporated as described in Section 5.2. The sampling rate is fixed at 96 kHz.

5.3.1 Results and Discussion. The optimized channel configurations are illustrated in Fig. 18. Across all setups, air-conduction microphones behind the ear are consistently selected, indicating higher sensitivity to time-delay and energy-ratio variations due to their distance from the mouth.

Fig. 16 shows the liveness detection accuracy of all configurations. For all setups, accuracy on attack settings 2 to 7 exceeds 96%, demonstrating robust detection of environmental replay attacks. For attack setting 1, only the 7-channel asymmetric configuration achieves accuracy above 96%, while others remain around 94%, suggesting that both sufficient channel count and optimized placement are necessary to detect wearer-based attacks. Moreover, for the same number of channels, asymmetric layouts consistently outperform symmetric ones (e.g., 4-channel asymmetric vs. 5-channel symmetric), implying that asymmetric configurations capture richer spatial information given the approximate symmetry of human sound fields. These findings provide practical guidance for microphone deployment in smart glasses.

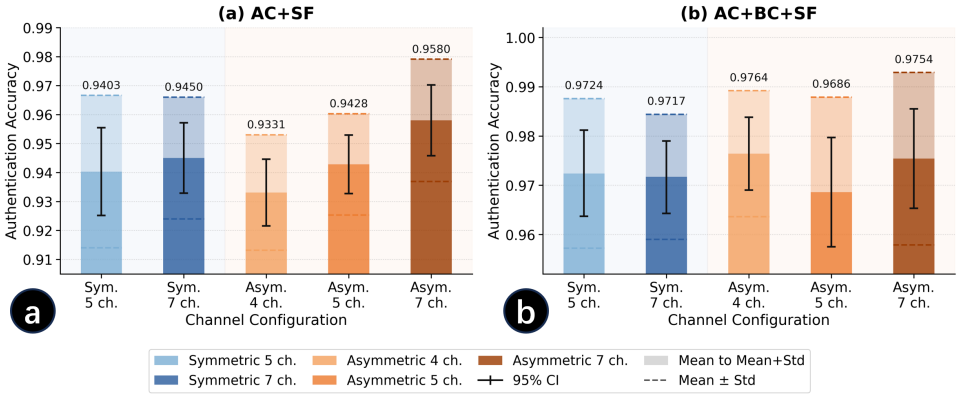


Fig. 17. Authentication accuracy under different channel inputs (1) without and (2) with BC modality.



Fig. 18. Liveness detection and authentication accuracy were evaluated under limited input-channel settings. (a) Top row shows the original microphone layout of the prototype and the selected microphones for the 5- and 7-channel symmetric configurations. Bottom row shows the selected microphones for the 4-, 5-, and 7-channel asymmetric configurations. Selected microphones are highlighted in orange. (b) demonstrates the microphone layouts of Rokid Glasses and Ray-Ban Meta Glasses on the left, and the microphones used in our experiments to simulate their recording characteristics on the right, with selected microphones highlighted in orange.

Fig. 17 demonstrates authentication accuracy with and without the BC modality. SF features from additional AC channels provide complementary information, with more channels generally improving performance. However, BC features offer a stronger complementary effect: when BC is included, accuracy remains stable across different SF configurations at around 97%, consistently higher than configurations without BC. This suggests that incorporating BC microphones is more beneficial than relying solely on AC channels, offering further design insights for future smart glasses.

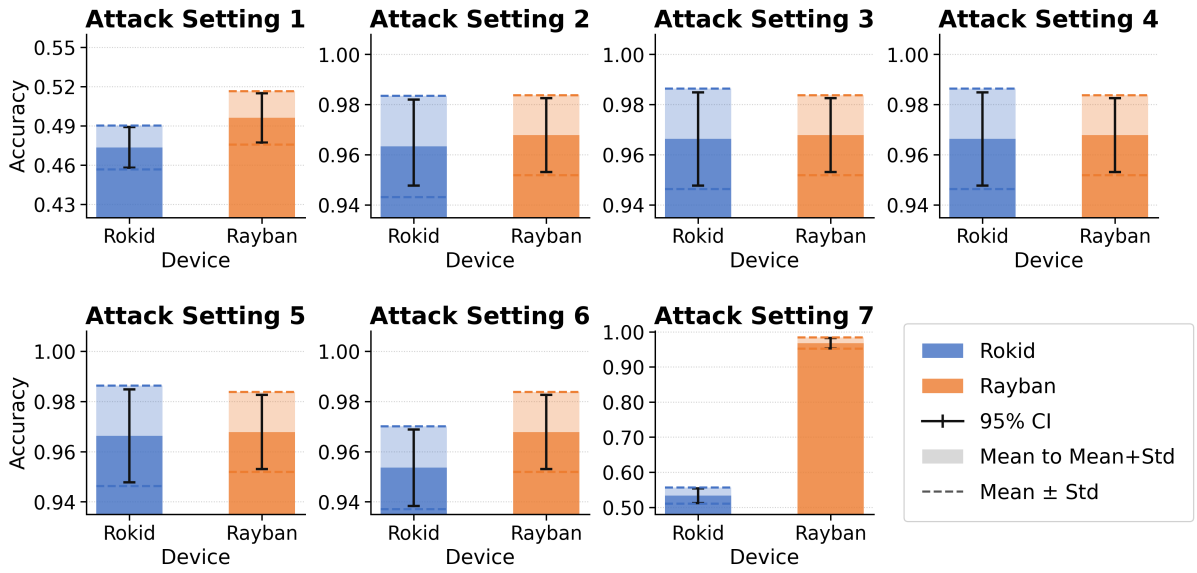


Fig. 19. Liveness detection accuracy for Rokid and Ray-ban Meta configurations.

5.4 Evaluation with Real-world Products' Configuration

5.4.1 Rokid Glass. Rokid Glass employs four symmetrically placed air-conduction microphones at the front and middle positions on both temples (Fig. 18 b), operating at 16 kHz. To emulate this setup, we select Channels 2, 5, 10, and 13 in our prototype to match the spatial layout, with Channel 5 as the reference (substituted AC modality). Corresponding AC and SF features are recomputed following the modifications in Sections 5.1 and 5.3. We evaluate liveness detection and authentication using Benchmark Task 1 and Task 3 (Section 4.1), with 7-fold cross-validation repeated twice.

5.4.2 Results and Discussion. Under the Rokid configuration, liveness detection accuracy (Fig. 19) remains high for attack settings 2–6 (above 96%), consistent with Section 5.3. However, performance drops significantly for attack settings 1 and 7 (i.e., closest or wearer-based attacks), with accuracy around 50%. This is likely due to the symmetric microphone layout introducing spatial redundancy and the compact distribution limiting sound field coverage: for sound sources close to the 4-channel array and near the symmetry plane, microphones on both sides receive highly similar signals, causing the aTDT and aERT features to collapse into a single three-point spike (upward or downward) that fails to capture the full voice propagation pattern; adopting asymmetric layouts, as illustrated in Section 5.3, can mitigate this issue.

For authentication (Fig. 20), AuthG-Net achieves performance consistent with Section 5.3, reaching 97% with BC and 94% without BC. This demonstrates robustness to different microphone layouts, even when the AC channel is off-center, and confirms that AuthG-Net maintains high accuracy under reduced sampling rates (16 kHz vs. 96 kHz).

5.4.3 Ray-Ban Meta. Ray-Ban Meta integrates five air-conduction microphones at 16 kHz (Fig. 18 b), including four similar to Rokid and one at the right nose pad. We emulate this setup using Channels 2, 5, 10, 13, and 7, with Channel 7 as the AC reference for SF feature extraction. Evaluation procedures and model modifications follow those of the Rokid configuration.

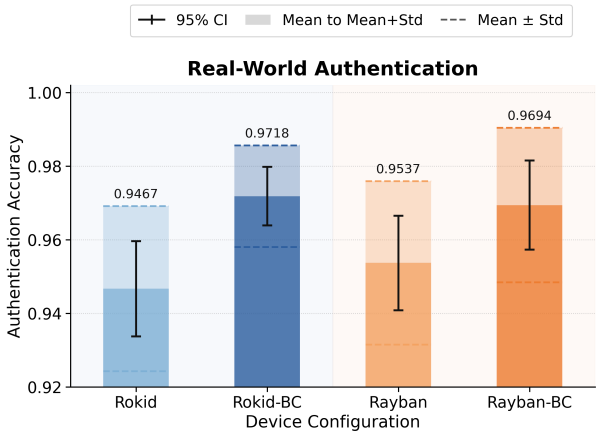


Fig. 20. Authentication accuracy for Rokid and Ray-ban Meta configurations.

5.4.4 Results and Discussion. Under the Ray-Ban Meta configuration, liveness detection accuracy (Fig. 19) remains high for attack settings 2–6 (above 95%), consistent with Rokid. Unlike Rokid (50%), performance on attack setting 7 is also high and comparable to Section 5.3. This is because, with five channels, the aTDT and aERT features form a four-point trapezoidal pattern, which can partially capture the user's voice propagation pattern. However, performance on attack setting 1 remains around 50%, indicating that when the sound source is near the mouth and exhibits head-like geometry, the symmetric 5-channel layout still fails to distinguish genuine speech from wearer-based attacks. Similar to Rokid, asymmetric microphone layouts may alleviate this limitation. Notably, the nose-pad microphone in Ray-Ban Meta introduces inherent asymmetry, which may enhance sound field feature capture; this effect warrants further investigation.

For authentication (Fig. 20), the simulated Ray-Ban Meta configuration at 16 kHz achieves performance comparable to both Rokid and Section 5.3, further confirming that AuthG-Net remains effective under limited sampling rates.

5.5 Live Evaluation

5.5.1 Implementation and Experiment Procedure. We evaluate the feasibility of *AuthG-Live* and *AuthG-Net* under live deployment via a simulated setup. Since our acquisition board only supports synchronized recording, multi-channel audio is streamed and processed on a Raspberry Pi[6] to emulate a resource-constrained edge device (Fig. 21c). The system operates in a streaming manner according to the sampling rate, mimicking real-time processing.

5.5.2 Live Simplification. Preliminary results show that latency is dominated by TDT and ERT, which require frame-wise cross-correlation and energy computation. To reduce overhead, we design two simplification levels: **(a) Level 1:** compute TDT/ERT at 24 kHz; **(b) Level 2:** use odd frames with padding to reconstruct full samples, further reducing computation while preserving time-delay accuracy. All other features are processed with 96kHz. We adopt a multi-threaded pipeline, where AC, BC, TDT, ERT, and EDF features are processed in parallel streams, followed by a unified inference thread for liveness detection and authentication.

5.5.3 System Performance on Latency and Energy. Latency is evaluated under simulated live conditions with 5 repeats for each length of input (Fig. 21a). For Level 1, processing latency remains below 4 s for inputs shorter than 3 s, but grows with input length. In contrast, Level 2 achieves a nearly constant latency of around 1.5–1.7 s,

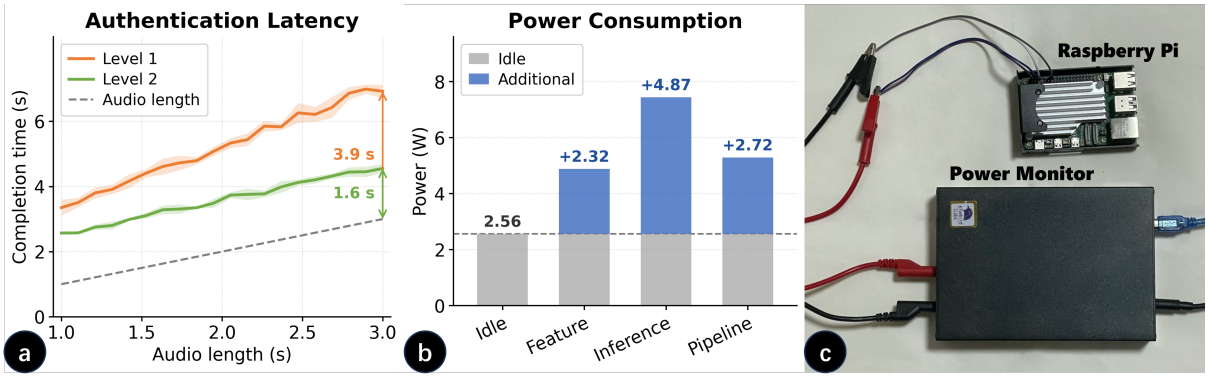


Fig. 21. Live system overhead. (a) Authentication latency vs. audio length for the two simplification levels (shaded: ±1 std over 5 repeats); the gap to the dashed *audio length* line is the processing latency (3.9 s for Level 1 and 1.6 s for Level 2 at 3 s input). (b) Power on the Raspberry Pi: idle baseline (2.56 W, dashed) plus the additional power per 3-second segment; the memory footprint stays below 267 MB. (c) The power-measurement setup.

independent of input duration, enabling near real-time processing. Energy and memory overheads are modest. Idle power (no audio streaming) is 2.56 W; for each 3-second segment, feature extraction and inference add 2.32 W and 4.87 W respectively, averaging 2.72 W over the full pipeline (Fig. 21b). This average overhead is of the same order as common continuous smartphone workloads such as a voice call (~1.05 W) or video playback (~0.45 W) [20]. Crucially, since authentication is brief and on-demand (~1.5 s per attempt), its per-use energy is only a few joules; even at 100 authentications per day this adds under 0.1 Wh—below 1% of a typical smartphone battery. The memory footprint stays below 267 MB for both levels, comparable to a single everyday mobile application. As these figures are measured on a Raspberry Pi standing in for a resource-constrained edge device, a dedicated glasses SoC would incur even lower cost.

5.5.4 Liveness Detection and Authentication Accuracy. We evaluate accuracy on Benchmark Task 1 and Task 3 using 7-fold cross-validation. For liveness detection (Fig. 22), both levels maintain performance close to the original model, with only a slight drop in attack setting 1 due to reduced TDT/ERT precision. For authentication (Fig. 23), no noticeable degradation is observed under either level, with or without BC features, demonstrating strong robustness under simplified processing.

6 DISCUSSION

In this section, we first summarize the key findings of this work (Section 6.1) and then highlight the potential applications of the proposed system (Section 6.2). Finally, we discuss its limitations and outline possible directions for future research (Section 6.3).

6.1 Conclusion of Findings

In this work, we systematically investigate voice for liveness detection and authentication on smart glasses, leading to the following insights.

Physics-driven liveness robustness. On-body propagation features (AuthG-Live) generalize to unseen attacks (Section 4.4.2), as replayed audio fails to reproduce spatial and anatomical constraints.

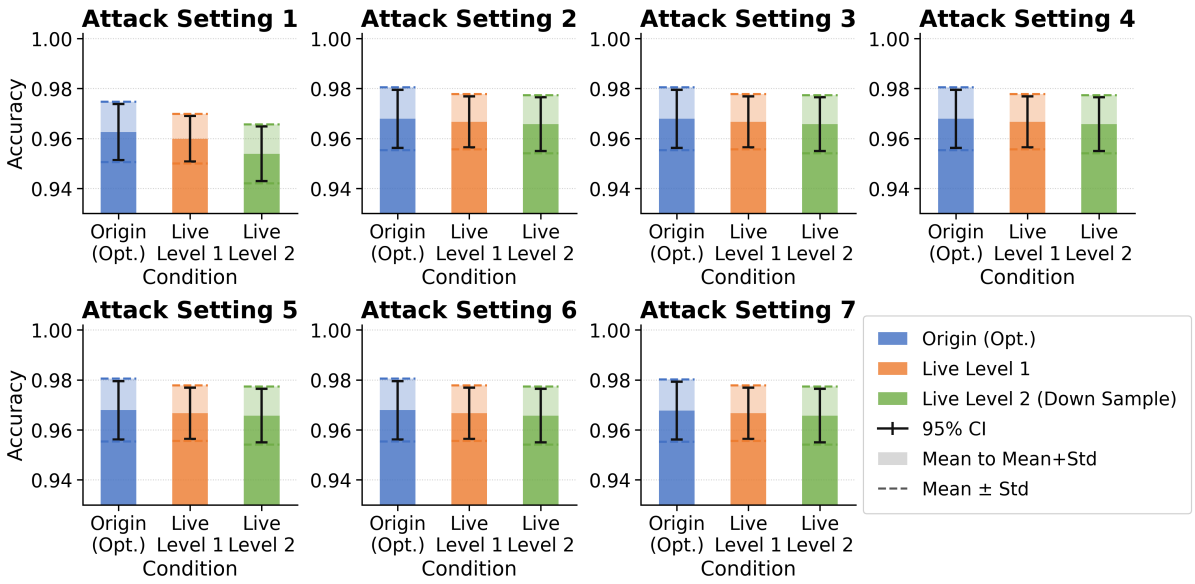


Fig. 22. Liveness detection accuracy under two levels of simplification.

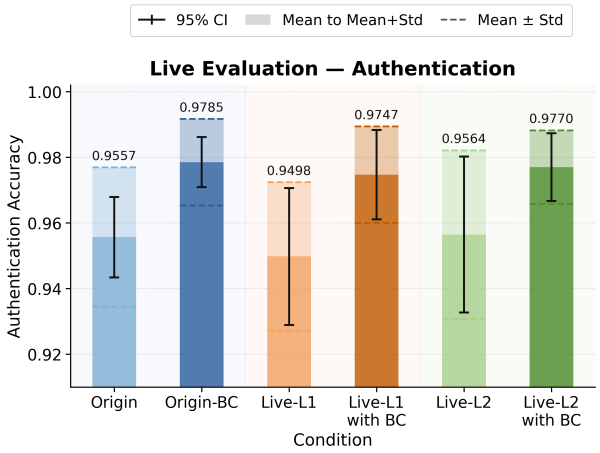


Fig. 23. Authentication accuracy under two levels of simplification.

Multi-modal necessity for authentication. Fusing AC, BC, and SF consistently outperforms unimodal/bimodal baselines across standard and cross-utterance settings (Section 4.4), improving robustness to spoofing and intra-speaker variability.

Deployment insights under real constraints. Performance remains strong under reduced microphone configurations, validating feasibility on commodity devices (Section 5); meanwhile, we observe that symmetric microphone layouts limit spatial feature capture, highlighting the importance of placement-aware sensor design.

Alignment with implicit authentication. Passive sensing (no extra user effort), on-body constraints (BC/SF tied to wearing condition), and strong cross-utterance and unseen-attack generalization (Section 4.4) together indicate that the learned representations are compatible with implicit and opportunistic authentication pipelines, where authentication can be triggered seamlessly during natural voice interaction.

6.2 Potential Applications

6.2.1 AuthG-Live and AuthG-Net. AuthG-Live and AuthG-Net enable secure voice-based interaction on smart glasses by leveraging multi-acoustic cues for joint liveness detection and authentication. The system can handle diverse real-world voice inputs, including those from the wearer, surrounding users, or when the device is not worn. By integrating authentication into passive multi-acoustic sensing, the system naturally supports seamless and implicit verification during normal voice interaction, without requiring additional user effort. In combination with other on-device sensors, such as a camera, this enables intuitive and secure interaction scenarios. For example, a user may simply look at a payment QR code and issue a voice command such as 'Pay the bill,' with authentication performed transparently as part of the interaction.

6.2.2 AuthGlass Dataset. As demonstrated in Section 4.1, the AuthGlass Dataset is designed to facilitate comprehensive evaluation of voice liveness detection and authentication on smart glasses. By providing synchronized multi-channel audio inputs, the dataset enables systematic ablation studies across different acoustic modalities and input channel configurations, allowing researchers to investigate the contribution of each modality and channel to system performance. Moreover, the dataset includes both genuine samples and attack data collected from multiple users, which supports rigorous evaluation of authentication and liveness detection under diverse adversarial scenarios.

6.3 Limitation and Future Works

6.3.1 Dataset Bias and Generalizability. Although AuthGlass includes 42 participants and multiple attack settings, it may still exhibit dataset bias due to the relatively homogeneous participant pool at university campus and controlled collection conditions. This may limit generalization to broader populations (e.g., diverse demographics, accents, and speaking styles) and real-world scenarios. In addition, the predefined attack settings may not cover all adversarial strategies in practice. To address this limitation, we open-source both the dataset and the prototype to facilitate future data collection and enrichment across more diverse users, environments, and attack conditions.

6.3.2 More Adversarial Scenarios. While our dataset covers representative attack settings derived from the defined threat model, it does not exhaust all possible adversarial strategies. More advanced attacks, such as adaptive replay, multi-source injection, or high-fidelity voice synthesis, may better approximate genuine acoustic patterns and are not fully captured in our current evaluation. In addition, variations in wearing conditions may introduce ambiguity in on-body acoustic cues. To address this limitation, we release the AuthGlass dataset and hardware design to enable future work to explore more diverse and challenging adversarial scenarios and improve system robustness.

6.3.3 In-the-Wild Data and Noise. Although the AuthGlass dataset is relatively large, it is collected under controlled conditions and does not fully capture real-world acoustic variability. In practical deployment, environmental noise may degrade performance. While fallback mechanisms can mitigate usability issues caused by false rejections, we do not explicitly evaluate or optimize robustness under noisy conditions, nor do we include noise in data collection. To address this limitation, we open-source our prototype and dataset to support future data collection in real-world environments. Additionally, noise augmentation techniques (e.g., using multi-channel noise datasets [78]) can be explored to improve robustness.

6.3.4 More Configurations of AC Microphones. Our study validates the effectiveness of incorporating sound-field features for liveness detection and authentication. However, we only explore a limited set of AC microphone configurations and do not fully optimize microphone placement. While our results suggest that spatial distribution plays a critical role, a systematic exploration of layout design remains missing. By releasing our dataset and prototype, we aim to facilitate future studies on microphone placement and configuration strategies, which may further improve performance and provide deeper insights into sound-field sensing on smart glasses.

6.3.5 Deployment to Commercial Devices. Despite demonstrating strong performance, several challenges remain for real-world deployment. The computation of sound-field features (e.g., TDT, ERT, EDF) introduces additional processing overhead and requires accurate synchronization across microphones. While our simulated real-time implementation demonstrates that input-independent processing can be achieved with a constant latency of approximately 1.5 seconds, further optimization is still needed for efficient deployment on resource-constrained commercial devices. In addition, integrating bone-conduction sensors into compact and wearable designs remains an open hardware challenge. Nevertheless, ongoing advances in embedded systems and sensor integration are likely to improve the feasibility of deploying such systems in practice.

6.3.6 Implicit Authentication. Voice interaction is inherently passive, and our approach further leverages on-body constraints through BC and SF features. Combined with strong generalization in both liveness detection and authentication, as well as low-latency processing in the live setting, these properties align with key requirements of implicit authentication systems. However, our current design and evaluation follow an event-driven pipeline, and both AuthG-Live and AuthG-Net are formulated accordingly. By releasing our dataset and prototype, we aim to enable future work to extend this framework toward continuous and implicit authentication under real-time audio streams.

6.3.7 Sound Field Based Interaction on Smart Glasses. Our findings indicate that sound-field features are closely related to human speech production and propagation, suggesting broader interaction possibilities. Prior work such as Proximic [65] and NasoVoce [68] uses inter-channel differences to detect mouth-related gestures. Building on this insight, sound-field features on smart glasses may enable gesture recognition (e.g., hand-to-mouth interactions) during speech without requiring active acoustic sensing [34], opening new directions for passive and multimodal interaction.

7 CONCLUSION

This work presents AuthGlass, along with two methods, AuthG-Live and AuthG-Net, for voice liveness detection and authentication on smart glasses. Our results show that combining sound-field and multimodal acoustic features enables robust defense against diverse spoofing and impersonation attacks while maintaining high authentication accuracy. Beyond explicit verification, our results suggest the potential of leveraging voice interaction as an implicit, event-driven authentication signal, enabling seamless and hands-free security without additional user effort. We further contribute AuthGlass, the first multi-acoustic-modal, multi-channel dataset tailored to smart glasses, providing a unified benchmark for evaluating liveness detection and authentication under realistic threat scenarios. By releasing this dataset and benchmark, we aim to facilitate reproducible research and future exploration of practical voice-based security systems, including continuous or hybrid authentication, cross-device generalization, and resource-aware deployment on commodity smart glasses.

8 USE OF AI TOOLS

We use GPT-5 for grammar correction and sentence refinement. We also use GPT-5 and Claude Code to beautify picture. To be specific, sketches in Fig. 1, Fig. 2a, Fig. 3b, Fig. 8b, are generated by GPT-5. Plots Fig. 11 to 17, Fig. 19, 20, 22, 23, 24, 25 are polished by Claude Code.

Acknowledgments

This work is supported by the National Key R&D Program of China under Grant No. 2024YFB2808800 & 2024YFB2808803 and Ant Group Research Fund.

References

[1] 2025. Eris® E5 Studio Monitor. https://intl.presonus.com/products/eris-e5-studio-monitor?srsltid=AfmBOopFGpJKhO6xHwJ_iRWlSQZsIOf9Kiw8Wv5WjZ0mKTitWmFcpOnO.
[2] 2025. GRAS 45BC KEMAR Head and Torso with Mouth Simulato. https://www.grasacoustics.com/products/head-torso-simulators-kemar/kemar-non-configured/product/749-45bc.
[3] 2025. Infineon IM73D122V01: Ultra-low noise high sensitivity digital XENSIVTM MEMS microphone. https://www.infineon.com/assets/row/public/documents/24/49/infineon-im73d122-datasheet-en.pdf?fileId=8ac78c8c83cd3081018409cafdfb46db.
[4] 2025. Microsoft Hololens. https://learn.microsoft.com/en-us/hololens/.
[5] 2025. The Munich Automatic Segmentation System MAUS. https://www.bas.uni-muenchen.de/Bas/BasMAUS.html.
[6] 2025. Raspbarry Pi 5. https://item.taobao.com/item.htm?id=848554796006.
[7] 2025. Ray-ban Meta. https://www.ray-ban.com/usa/ray-ban-meta-ai-glasses.
[8] 2025. Rokid Glasses. https://glasses.rokid.com.
[9] 2025. Sonion Voice Picking Unit VPU14aa01. https://www.ariat-tech.com/parts/VPU14AA01?srsltid=AfmBOoqPZF0ySUfKmsHCBWudGqWET3zyMNi78FVCvMtjfjzo8pOYW5_n.
[10] 2025. Xiaomi Smart Audio Glasses. https://www.mi.com/global/product/xiaomi-smart-audio-glasses.
[11] 2025. Xreal One Pro. https://www.xreal.com/one-pro.
[12] 2026. 16-Element Discrete Microphone Array (Free-Form Layout) Kit. https://item.taobao.com/item.htm?id=665645262286&mi_id=0000_37rOJdF150WubGFdgNkOLMTKH1gYxqfX44ihDN0cXo&spm=tbpc.boughtlist.suborder_itemtitle.1.24992e8dOxApXT&skuId=4792669972340.
[13] 2026. Amazon Echo Frames Shop. https://www.amazon.com/b/?node=117191383011&ref_=mars_byline_echo_frames_branded.
[14] 2026. Smart meets style: Power your payments with Alipay+ GlassPay. https://www.alipayplus.com/glasspay/.
[15] Muhammad Ejaz Ahmed, Il-Youp Kwak, Jun Ho Huh, 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.
[16] Orlando Arias, Jacob Wurm, Khoa Hoang, and Yier Jin. 2015. Privacy and security in internet of things and wearable devices. *IEEE transactions on multi-scale computing systems* 1, 2 (2015), 99–109.
[17] Pierre Badin and Antoine Serrurier. 2006. Three-dimensional modeling of speech organs: Articulatory data and models. In *Technical Committee of Psychological and Physiological Acoustics*, Vol. 36. The Acoustical Society of Japan, 421–426.
[18] Logan Blue, Hadi Abdullah, Luis Vargas, and Patrick Traynor. 2018. 2ma: Verifying voice commands via two microphone authentication. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*. 89–100.
[19] Bruno Bouchard, Kevin Bouchard, and Abdenour Bouzouane. 2020. A smart cooking device for assisting cognitively impaired users. *Journal of reliable intelligent environments* 6, 2 (2020), 107–125.
[20] Aaron Carroll and Gernot Heiser. 2010. An Analysis of Power Consumption in a Smartphone. In *Proceedings of the 2010 USENIX Annual Technical Conference (USENIX ATC'10)*. USENIX Association, 271–284.
[21] Pan Chan, Tzipora Halevi, and Nasir Memon. 2015. Glass otp: Secure and convenient user authentication on google glass. In *International Conference on Financial Cryptography and Data Security*. Springer, 298–308.
[22] Shan Chang, Xinggan Hu, Hongzi Zhu, Wei Liu, and Lei Yang. 2022. Vogue: Secure user voice authentication on wearable devices using gyroscope. In *2022 19th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 361–369.
[23] Wan-Jung Chang, Liang-Bi Chen, and Yu-Zung Chiou. 2018. Design and implementation of a drowsiness-fatigue-detection system based on wearable smart glasses to increase road safety. *IEEE Transactions on Consumer Electronics* 64, 4 (2018), 461–469.
[24] Tao Chen, Yongjie Yang, Chonghao Qiu, Xiaoran Fan, Xiuzhen Guo, and Longfei Shangguan. 2024. Enabling hands-free voice assistant activation on earphones. In *Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services*. 155–168.
[25] Ryan M. Corey, Matthew D. Skarha, and Andrew C. Singer. 2019. Massive Distributed Microphone Array Dataset. doi:10.13012/B2IDB-6216881_V1
[26] Phillip L. De Leon, Michael Pucher, Junichi Yamagishi, Inma Hernaez, and Ibon 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. doi:10.1109/TASL.2012.2201472
[27] Yuemeng Du, Jingyan Qin, Shujing Zhang, Sha Cao, and Jinhua Dou. 2018. Voice user interface interaction design research based on user mental model in autonomous vehicle. In *International Conference on Human-Computer Interaction*. Springer, 117–132.

[28] Serife Kucur Ergünay, Elie Khoury, Alexandros Lazaridis, and Sébastien Marcel. 2015. On the vulnerability of speaker verification to realistic voice spoofing. In *2015 IEEE 7th international conference on biometrics theory, applications and systems (BTAS)*. IEEE, 1–6.

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

[30] Eira Friström, Elias Lius, Niki Ulmanen, Paavo Hietala, Pauliina Kärkkäinen, Tommi Mäkinen, Stephan Sigg, and Rainhard Dieter Findling. 2019. Free-form gaze passwords from cameras embedded in smart glasses. In *Proceedings of the 17th International Conference on Advances in Mobile Computing & Multimedia*. 136–144.

[31] Bhanuka Gamage. 2024. AI-Enabled Smart Glasses for People with Severe Vision Impairments. *ACM SIGACCESS Accessibility and Computing* 137 (2024), 1–1.

[32] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*. PMLR, 1180–1189.

[33] Yang Gao, Yincheng Jin, Jagmohan Chauhan, Seokmin Choi, Jiyang Li, and Zhanpeng Jin. 2021. Voice in ear: Spoofing-resistant and passphrase-independent body sound authentication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–25.

[34] Kaiyi Guo, Tianyu Wu, Yang Gao, Qian Zhang, and Dong Wang. 2025. EchoTouch: Low-power Face-touching Behavior Recognition Using Active Acoustic Sensing on Glasses. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 2, Article 31 (June 2025), 33 pages. doi:10.1145/3729481

[35] Jibo He, Jake Ellis, William Choi, and Pingfeng Wang. 2015. Driving While Reading Using Google Glass Versus Using a Smartphone: Which is More Distracting to Driving Performance?. In *Driving Assessment Conference*, Vol. 8. University of Iowa.

[36] Ruiwen He, Yushi Cheng, Zhicong Zheng, Xiaoyu Ji, and Wenyuan Xu. 2024. Fast and lightweight voice replay attack detection via time-frequency spectrum difference. *IEEE Internet of Things Journal* 11, 18 (2024), 29798–29810.

[37] Chenpei Huang, Hui Zhong, Pavana Prakash, Dian Shi, Xu Yuan, and Miao Pan. 2025. Eve said yes: Airbone authentication for head-wearable smart voice assistant. *IEEE Transactions on Mobile Computing* (2025).

[38] Yuliy Iliev and Galina Ilieva. 2022. A framework for smart home system with voice control using NLP methods. *Electronics* 12, 1 (2022), 116.

[39] Masaya Inoue and Kazuya Murao. 2024. User Authentication Method for Smart Glasses using Gaze Information of Registered Known Images and AI-generated Unknown Images. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 525–530.

[40] MD Rasel Islam, Doyoung Lee, Liza Suraiya Jahan, and Ian Oakley. 2018. Glasspass: Tapping gestures to unlock smart glasses. In *Proceedings of the 9th Augmented Human International Conference*. 1–8.

[41] Jingun Jung, Sangyoon Lee, Jiwoo Hong, Eunhye Youn, and Geehyuk Lee. 2020. Voice+ tactile: Augmenting in-vehicle voice user interface with tactile touchpad interaction. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[42] Madhu R Kamble, Hardik B Sailor, Hemant A Patil, and Haizhou Li. 2020. Advances in anti-spoofing: from the perspective of ASVspoof challenges. *APSIPA Transactions on Signal and Information Processing* 9 (2020), e2.

[43] Navdeep Kaur and Parminder Singh. 2023. Conventional and contemporary approaches used in text to speech synthesis: A review. *Artificial Intelligence Review* 56, 7 (2023), 5837–5880.

[44] Lawrence George Kersta. 1962. Voiceprint identification. *The Journal of the Acoustical Society of America* 34, 5_Supplement (1962), 725–725.

[45] Smita Khade, Swati Ahirrao, Shraddha Phansalkar, Ketan Kotecha, Shilpa Gite, and Sudeep D Thepade. 2021. Iris liveness detection for biometric authentication: A systematic literature review and future directions. *Inventions* 6, 4 (2021), 65.

[46] Sunwook Kim, Maury A Nussbaum, and Joseph L Gabbard. 2016. Augmented reality "smart glasses" in the workplace: industry perspectives and challenges for worker safety and health. *IIE transactions on occupational ergonomics and human factors* 4, 4 (2016), 253–258.

[47] Tomi Kinnunen, Md Sahidullah, Mauro Falcone, Luca Costantini, Rosa González Hautamäki, Dennis Thomsen, Achintya Sarkar, Zheng-Hua Tan, Héctor Delgado, Massimiliano Todisco, et al. 2017. Reddots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research. In *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 5395–5399.

[48] Huining Li, Chenhan Xu, Aditya Singh Rathore, Zhengxiong Li, Hanbin Zhang, Chen Song, Kun Wang, Lu Su, Feng Lin, Kui Ren, et al. 2020. VocalPrint: Exploring a resilient and secure voice authentication via mmWave biometric interrogation. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 312–325.

[49] Jingjin Li, Chao Chen, Mostafa Rahimi Azghadi, Hossein Ghodosi, Lei Pan, and Jun Zhang. 2023. Security and privacy problems in voice assistant applications: A survey. *Computers & Security* 134 (2023), 103448.

[50] Ke Li, Devansh Agarwal, Ruidong Zhang, Vipin Gunda, Tianjun Mo, Saif Mahmud, Boao Chen, François Guimbretière, and Cheng Zhang. 2024. SonicID: User Identification on Smart Glasses with Acoustic Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 4 (2024), 1–27.

[51] Yue Li, Xueru Gao, Qipeng Song, Yao Wang, Pin Lyu, and Haibin Zhang. 2024. BoneAuth: A Bone Conduction-Based Voice Liveness Authentication for Voice Assistants. *IEEE Internet of Things Journal* (2024).

[52] Yung-Hui Li and Po-Jen Huang. 2017. An accurate and efficient user authentication mechanism on smart glasses based on iris recognition. *Mobile Information Systems* 2017, 1 (2017), 1281020.

[53] Zhuohang Li, Cong Shi, Tianfang Zhang, Yi Xie, Jian Liu, Bo Yuan, and Yingying Chen. 2021. Robust detection of machine-induced audio attacks in intelligent audio systems with microphone array. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 1884–1899.

[54] Hyunchul Lim, Guilin Hu, Richard Jin, Hao Chen, Ryan Mao, Ruidong Zhang, and Cheng Zhang. 2023. C-Auth: Exploring the Feasibility of Using Egocentric View of Face Contour for User Authentication on Glasses. In *Proceedings of the 2023 ACM International Symposium on Wearable Computers*. 6–10.

[55] 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.

[56] Isaias Majil, Mau-Tsuen Yang, and Sophia Yang. 2022. Augmented reality based interactive cooking guide. *Sensors* 22, 21 (2022), 8290.

[57] Mitchell McLaren, Luciana Ferrer, Diego Castan, and Aaron Lawson. 2016. The speakers in the wild (SITW) speaker recognition database.. In *Interspeech*. 818–822.

[58] Yan Meng, Jiachun Li, Matthew Pillari, Arjun Deopujari, Liam Brennan, Hafsah Shamsie, Haojin Zhu, and Yuan Tian. 2022. Your microphone array retains your identity: A robust voice liveness detection system for smart speakers. In *31st USENIX Security Symposium (USENIX Security 22)*. 1077–1094.

[59] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612* (2017).

[60] AN Nithyaa, R Premkumar, N Sudhandirapriya, and R Durga. 2023. Eye Glasses for the Visually Challenged Using Artificial Intelligence Application. In *International Conference on Soft Computing and Signal Processing*. Springer, 347–360.

[61] Francesca Palermo, Luca Casciano, Lokmane Demagh, Aurelio Teliti, Niccolò Antonello, Giacomo Gervasoni, Hazem Hesham Yousef Shalby, Marco Brando Paracchini, Simone Mentasti, Hao Quan, et al. 2025. Advancements in Context Recognition for Edge Devices and Smart Eyewear: Sensors and Applications. *IEEE Access* (2025).

[62] Seung-A Park, Sun-Beom Kwon, Yong Woo Lee, Jejin Jo, Jun-Seob Kim, Mee Lan Han, and Dooho Choi. 2025. Toward almost-zero fault acceptance of deep learning-based voice authentication using small training dataset. *Soft Computing* (2025), 1–12.

[63] Ge Peng, Gang Zhou, David T Nguyen, Xin Qi, Qing Yang, and Shuangquan Wang. 2016. Continuous authentication with touch behavioral biometrics and voice on wearable glasses. *IEEE transactions on human-machine systems* 47, 3 (2016), 404–416.

[64] Lam Pham, Phat Lam, Dat Tran, Hieu Tang, Tin Nguyen, Alexander Schindler, Florian Skopik, Alexander Polonsky, and Hai Canh Vu. 2025. A comprehensive survey with critical analysis for deepfake speech detection. *Computer Science Review* 57 (2025), 100757.

[65] Yue Qin, Chun Yu, Zhaoheng Li, Mingyuan Zhong, Yukang Yan, and Yuanchun Shi. 2021. Proximic: Convenient voice activation via close-to-mic speech detected by a single microphone. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12.

[66] P Selvi Rajendran, Padmaveni Krishnan, and D John Aravindhar. 2020. Design and implementation of voice assisted smart glasses for visually impaired people using google vision api. In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. IEEE, 1221–1224.

[67] P. Selvi Rajendran, Padmaveni Krishnan, and D. John Aravindhar. 2020. Design and Implementation of Voice Assisted Smart Glasses for Visually Impaired People Using Google Vision API. In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. 1221–1224. doi:10.1109/ICECA49313.2020.9297553

[68] Jun Rekimoto, Yu Nishimura, and Bojian Yang. 2026. NasoVoce: A Nose-Mounted Low-Audibility Speech Interface for Always-Available Speech Interaction. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–11.

[69] Stefan Schneegass, Youssef Oualil, and Andreas Bulling. 2016. SkullConduct: Biometric user identification on eyewear computers using bone conduction through the skull. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 1379–1384.

[70] Suranga Seneviratne, Yining Hu, Tham Nguyen, Guohao Lan, Sara Khalifa, Kanchana Thilakarathna, Mahbub Hassan, and Aruna Seneviratne. 2017. A survey of wearable devices and challenges. *IEEE Communications Surveys & Tutorials* 19, 4 (2017), 2573–2620.

[71] Cong Shi, Yan Wang, Yingying Chen, Nitesh Saxena, and Chen Wang. 2020. Wearid: Low-effort wearable-assisted authentication of voice commands via cross-domain comparison without training. In *Proceedings of the 36th Annual Computer Security Applications Conference*. 829–842.

[72] Hyejin Shin, Jun Ho Huh, Bum Jun Kwon, Iljoo Kim, Eunyong Cheon, HongMin Kim, Choong-Hoon Lee, and Ian Oakley. 2024. SkullID: Through-skull sound conduction based authentication for smartglasses. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–19.

[73] Ryohei Shioji, Shin-ichi Ito, Momoyo Ito, and Minoru Fukumi. 2018. Personal authentication and hand motion recognition based on wrist EMG analysis by a convolutional neural network. In *2018 IEEE International Conference on Internet of Things and Intelligence System (IOTAIS)*. IEEE, 184–188.

[74] Sayaka Shiota, Fernando Villavicencio, Junichi Yamagishi, Nobutaka Ono, Isao Echizen, and Tomoko Matsui. 2015. Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification. In *INTERSPEECH 2015 16th Annual Conference of the International Speech Communication Association*. International Speech Communication Association, France, 239–243.

[75] Yao Song, Yanpu Yang, and Peiyao Cheng. 2022. The investigation of adoption of voice-user interface (VUI) in smart home systems among Chinese older adults. *Sensors* 22, 4 (2022), 1614.

[76] Yu Su, Yongjiao Li, and Zhu Cao. 2023. Gait-based privacy protection for smart wearable devices. *IEEE Internet of Things Journal* 11, 2 (2023), 3497–3509.

[77] Ashique T P, Hridya Girish, Ivine P Sabu, Thomas Raju, and Anoop V. 2025. AI-Powered Smart Glasses for Real-Time Speech-to-Text Transcription. In *2025 4th International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS)*. 442–445. doi:10.1109/ACCESS65134.2025.11135871

[78] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. 2013. DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments. *(No Title)* (2013).

[79] Mohamed A Torad, Belgacem Bouallegue, and Abdelmoty M Ahmed. 2022. A voice controlled smart home automation system using artificial intelligent and internet of things. *TELKOMNIKA (Telecommunication Computing Electronics and Control)* 20, 4 (2022), 808–816.

[80] Andreas Triantafyllopoulos, Björn W Schuller, Gökçe İymen, Metin Sezgin, Xiangheng He, Zijiang Yang, Panagiotis Tzirakis, Shuo Liu, Silvan Mertes, Elisabeth André, et al. 2023. An overview of affective speech synthesis and conversion in the deep learning era. *Proc. IEEE* 111, 10 (2023), 1355–1381.

[81] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7167–7176.

[82] Suresh Veesa and Madhusudan Singh. 2025. Deep learning countermeasures for detecting replay speech attacks: a review. *International Journal of Speech Technology* 28, 1 (2025), 39–51.

[83] Qian Wang, Xiu Lin, Man Zhou, Yanjiao Chen, Cong Wang, Qi Li, and Xiangyang Luo. 2019. VoicePop: A Pop Noise based Anti-spoofing System for Voice Authentication on Smartphones. In *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*. 2062–2070. doi:10.1109/INFOCOM.2019.8737422

[84] Yao Wang, Wandong Cai, Tao Gu, Wei Shao, Yannan Li, and Yong Yu. 2019. Secure your voice: An oral airflow-based continuous liveness detection for voice assistants. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 3, 4 (2019), 1–28.

[85] Emily Wenger, Max Bronckers, Christian Cianfarani, Jenna Cryan, Angela Sha, Haitao Zheng, and Ben Y Zhao. 2021. " Hello, It's Me": Deep Learning-based Speech Synthesis Attacks in the Real World. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 235–251.

[86] Adrian Wong, Thad Eugene Starner, and Joshua Weaver. 2017. Wearable computing device authentication using bone conduction. U.S. Patent 9,740,842. Assignee: Google Inc. Appl. No. 14/976,353, filed Dec. 21, 2015. Issued Aug. 22, 2017.

[87] Zhanwei Xu, Haoxiang Pei, Jianjiang Feng, and Jie Zhou. 2025. FingerGlass: Enhancing Smart Glasses Interaction via Fingerprint Sensing. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.

[88] Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, et al. 2021. ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. *arXiv preprint arXiv:2109.00537* (2021).

[89] Chen Yan, Xiaoyu Ji, Kai Wang, Qinhong Jiang, Zizhi Jin, and Wenyuan Xu. 2022. A survey on voice assistant security: Attacks and countermeasures. *Comput. Surveys* 55, 4 (2022), 1–36.

[90] 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 2019 ACM SIGSAC Conference on Computer and Communications Security*. 1215–1229.

[91] Bing Yang, Changsheng Quan, Yabo Wang, Pengyu Wang, Yujie Yang, Ying Fang, Nian Shao, Hui Bu, Xin Xu, and Xiaofei Li. 2024. RealMAN: A real-recorded and annotated microphone array dataset for dynamic speech enhancement and localization. *Advances in Neural Information Processing Systems* 37 (2024), 105997–106019.

[92] Junshuang Yang, Yanyan Li, and Mengjun Xie. 2015. Motionauth: Motion-based authentication for wrist worn smart devices. In *2015 IEEE International conference on pervasive computing and communication workshops (PerCom Workshops)*. IEEE, 550–555.

[93] Qiang Yang, Kaiyan Cui, and Yuanqing Zheng. 2023. VoShield: Voice liveness detection with sound field dynamics. In *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*. IEEE, 1–10.

[94] Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao. 2023. Audio deepfake detection: A survey. *arXiv preprint arXiv:2308.14970* (2023).

[95] Linghan Zhang, Sheng Tan, Yingying Chen, and Jie Yang. 2022. A continuous articulatory-gesture-based liveness detection for voice authentication on smart devices. *IEEE Internet of Things Journal* 9, 23 (2022), 23320–23331.
[96] 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.
[97] Linghan Zhang and Jie Yang. 2021. A continuous liveness detection for voice authentication on smart devices. *arXiv preprint arXiv:2106.00859* (2021).
[98] Tianfang Zhang, Qiufan Ji, Zhengkun Ye, Md Mojibur Rahman Redoy Akanda, Ahmed Tanvir Mahdad, Cong Shi, Yan Wang, Nitesh Saxena, and Yingying Chen. 2024. SAFARI: Speech-Associated Facial Authentication for AR/VR Settings via Robust VIbration Signatures. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*. 153–167.
[99] Zhan Zhang, Enze Bai, Aram Stepanian, Swathi Jagannath, and Sun Young Park. 2025. Touchless interaction for smart glasses in emergency medical services: user needs and experiences. *International Journal of Human–Computer Interaction* 41, 5 (2025), 2984–3003.
[100] Tianming Zhao, Yan Wang, Jian Liu, Yingying Chen, Jerry Cheng, and Jiadi Yu. 2020. Trueheart: Continuous authentication on wrist-worn wearables using ppg-based biometrics. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 30–39.
[101] Chu-Xiao Zuo, Zhi-Jun Jia, and Wu-Jun Li. 2024. AdvTTS: Adversarial Text-to-Speech Synthesis Attack on Speaker Identification Systems. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4840–4844.

A APPENDIX ON DATASET AND METHODS FOR BENCHMARKING

A.1 Command Set and Rationale

Table 7 lists the 15 commands used in our experiments. Each phrase contains no more than five words, enabling the evaluation of liveness detection and speaker authentication. They were selected from common daily expressions that could serve as practical voice commands, ensuring the scenarios are realistic for everyday use. In addition, the set was chosen to collectively cover most English vowels and consonants, providing representative phonetic content so that the model's performance is not biased by the specific phonemes of individual phrases.

A.2 Additional Participants Demography

We recruited a total of 42 participants from a university campus community (Fig. 24). The age of participants ranged from 18 to 42 years, with a mean of 24.6 ± 4.3 years. Among them, 28 identified as male (66.7%) and 14 as female (33.3%).

The participant pool primarily consisted of students (n=41), with one staff member. Students were drawn from diverse academic backgrounds, including Engineering and Computer Science (n=33), Arts (n=3), interdisciplinary Colleges (n=3), Medicine (n=1), and a Research Institute (n=1), reflecting a typical campus population.

In terms of linguistic background, the majority of participants were non-native English speakers using Chinese-accented English (n=37). Additional accents included South Asian (n=2) and Arabic-accented English (n=2), along with one native speaker of American English (n=1). While the dataset is dominated by a single accent group, it still captures moderate linguistic variability within a realistic university setting.

All participants reported being proficient English users in daily communication. During data collection, they were instructed to speak at a natural conversational pace, maintain clear pronunciation, and avoid exaggerated articulation. This setup ensures that the recorded speech reflects realistic voice-command usage rather than carefully controlled or artificially slowed speech.

Data collection was conducted in a quiet indoor environment. Participants were asked to remain in a natural seated posture while wearing the prototype device. No strict constraints were imposed on individual speaking styles beyond clarity and naturalness, allowing natural inter-speaker variability in speech patterns.

Table 7. Passphrase set used in the experiments.

(1) Hi siri	(2) Hi google	(3) Hi alexa
(4) Pay the bill	(5) Call tom	(6) Turn on bluetooth
(7) Take a photo	(8) Play some music	(9) Show me my first message
(10) Send an email to john	(11) Check my voicemail	(12) What's my appointment
(13) Set an alarm	(14) Thank you	(15) A hundred dollars

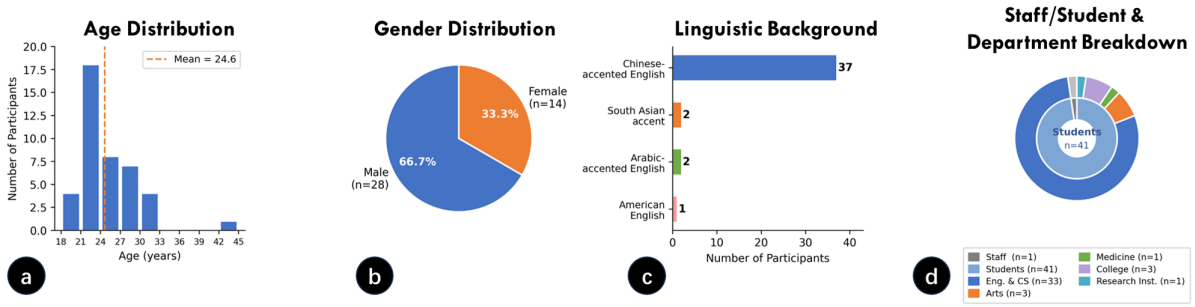


Fig. 24. Demographic overview of the 42 participants in the AuthGlass dataset. (a) Age distribution (range: 18–42, mean 24.6 ± 4.3 years). (b) Gender distribution: 28 male (66.7%) and 14 female (33.3%). (c) Linguistic background: the majority are non-native English speakers with Chinese-accented English (n = 37), alongside speakers with South Asian (n = 2), Arabic-accented (n = 2), and American English (n = 1) backgrounds, reflecting moderate linguistic diversity within a primarily university-campus population. (d) Affiliation breakdown: 41 students and 1 staff member, with students drawn predominantly from Engineering & Computer Science (n = 33), and smaller representation from Arts (n = 3), interdisciplinary Colleges (n = 3), Medicine (n = 1), and a Research Institute (n = 1).

A.3 Selected Methods for Benchmarking and Our Modifications

Liveness Detection. For liveness detection, we selected 7 methods for benchmarking.

VOID [15] VOID extracts four types of features from a single-channel AC signal, including three handcrafted features (low-frequency power, signal power linearity, and high-frequency power) and LPCCs. These features are fed into several traditional machine learning classifiers, among which an SVM with an RBF kernel achieves the best performance. To extend VOID to multi-channel inputs, we reapply the entire feature extraction pipeline to each input channel and concatenate all channel-wise features as the input to the SVM classifier.

He et al. [36] Similar to VOID, this method extracts five types of features from a single-channel AC signal, consisting of four handcrafted features and LPCCs. The extracted features are input to a self-designed SE-ResNet, a residual convolutional neural network (CNN). To support multi-channel inputs, we perform feature extraction independently on each channel and stack the resulting features across channels to form a multi-channel input. We modified the input layer of the SE-ResNet for the multi-channel input.

Boneauth [51] Boneauth computes MFCCs from both BC and AC signals and performs liveness detection by measuring the cosine similarity between the two modalities. An input is accepted only if the similarity exceeds a predefined threshold. The threshold is determined by minimizing the Equal Error Rate (EER). Since the original method operates on a single BC–AC pair, we extend it to the multi-channel setting by pairing the BC channel with each AC channel individually and applying majority voting to determine the final liveness decision.

Eve Said Yes [37] Eve Said Yes uses pooled STFT energy spectrograms as input features and defines a temporal similarity metric between AC and BC signals for liveness detection. An utterance is accepted if the similarity

exceeds a threshold, determined by minimizing EER. As this method also relies on a single AC–BC channel pair, we adopt the same multi-channel voting strategy as used for Boneauth.

Li et al. [53] This method is inherently designed for multi-channel voice liveness detection. It leverages both energy and phase spectra derived from STFT as input features and employs a CNN for classification. Since the model naturally accommodates multi-channel inputs, we directly feed features from all selected channels only with the modification on the CNN input layer.

Cafield [90] Cafield proposes Fieldprint, a feature computed as the ratio of the energy–frequency relationship vectors between two channels. A GMM is trained to model genuine samples and distinguish them from attack samples. We extend this method by computing Fieldprints between each channel and the reference AC channel (Channel 7), and concatenating all resulting features for GMM-based classification

Voshield [93] Voshield utilizes sound field dynamics (SFD), defined as the ratio of STFT spectra between two channels, and employs a residual CNN for binary classification. Similar to Cafield, we compute SFD features between each channel and the reference AC channel (Channel 7), and stack all resulting features as the input to the CNN, with its' input layer modified for all the input channels.

Authentication. For authentication, we selected 4 methods for benchmarking.

Park et al. [62] This method originally operates on a single AC channel and uses MFCCs as input features. A self-designed LSTM network is employed as the authentication model. During enrollment, embeddings extracted from an intermediate layer are stored as user templates, and cosine similarity is used to compare test embeddings against the enrolled template. Authentication is accepted if the similarity exceeds a predefined threshold. Following this procedure, we compute MFCC features for all selected channels, stack them together, and modify the model input layer accordingly to support multi-channel inputs.

Eve Said Yes [37] Eve Said Yes uses CQT features extracted from a single BC channel as input to a residual CNN trained with adversarial learning to produce speaker embeddings. During enrollment, embeddings are extracted from an intermediate fully connected layer and compared with stored templates using cosine similarity. Similar to Park et al., we extend this method by stacking features extracted from all channels and adapting the model input layer to accommodate the expanded multi-feature input.

Cafield [90] As discussed previously, Cafield employs a GMM-based framework for both liveness detection and authentication. We apply the same multi-channel feature construction and model modification strategy as used in the liveness detection setting.

Voshield [93] Voshield does not originally support authentication, as it is designed for binary liveness classification. However, since it is based on a CNN architecture, it can be readily adapted for authentication by replacing the binary output layer with a fully connected layer followed by a softmax output. We extract embeddings from the fully connected layer immediately following the residual block and perform authentication using cosine similarity, following the same protocol as Park et al. and Eve Said Yes.

B DETAILED IMPLEMENTATION OF FEATURE EXTRACTION AND AUTHG-NET

B.1 Feature Extraction Implementation

This appendix provides a detailed description of the feature extraction procedures used for both air-conduction (AC) and bone-conduction (BC) signals, as well as the derivation of sound field (SF) features.

B.1.1 AC and BC Channels. Log-Mel spectrograms were computed for both AC and BC channels. For AC signals, raw waveforms were directly transformed into Mel spectrograms. In contrast, BC signals were first processed using a Butterworth high-pass filter with a cutoff frequency of 100 Hz to suppress motion-induced low-frequency artifacts and ambient noise.

Table 8. Feature extraction parameters.

Parameters for Mel Spectrogram		
Sampling rate: 96,000 Hz	Maximum frequency: 8,000 Hz	
Hop length: 1024	N_{FFT}: 2048	N_{MELS}: 128
Parameters for ERT and TDT		
Sample rate: 96,000 Hz	Frequency range: 100–8,000 Hz	
Step length: 0.005 s	Scaling factor: 2000	
Parameters for EDF		
Sample rate: 96,000 Hz	Maximum frequency: 8,000 Hz	
Hop length: 1024	N_{FFT}: 1024	

All signals were subsequently resampled to 96 kHz via linear interpolation to ensure a consistent temporal resolution. Mel spectrograms were then generated using 128 Mel filter banks with an upper frequency limit of 8 kHz. The resulting representations capture perceptually relevant spectral characteristics while maintaining a uniform feature dimensionality across modalities.

B.1.2 Sound Field Feature Construction. To characterize spatial acoustic properties, three complementary features were extracted from the AC SF channels: Energy Ratio over Time (ERT), Time Delay over Time (TDT), and Energy Distribution over Frequency (EDF). These features follow the design principles described in Section 3.3.

Energy Ratio over Time (ERT). ERT was computed by first applying the short-time Fourier transform (STFT) to each voiced segment. Spectral energy within the 100–8,000 Hz band was then summed for each AC channel (see Section 3.4.1). To reduce variability caused by absolute sound intensity, channel-wise energy was normalized with respect to a reference channel (Channel 7). The resulting ratios were linearly scaled and further normalized using a hyperbolic tangent (tanh) function to improve numerical stability and bound the feature range.

Time Delay over Time (TDT). TDT was derived by estimating relative time delays between each AC channel and the reference channel (Channel 7) using cross-correlation. For each channel, the time lag corresponding to the maximum correlation peak was selected as the delay estimate. A confidence score was computed by normalizing the peak correlation with respect to signal energies and applying a logit-like transformation. The final delay value was obtained by weighting the estimated lag with this confidence score, which suppresses unreliable estimates caused by noise or absent signals while preserving robust temporal alignment cues. As with ERT, the resulting values were linearly scaled and normalized using a tanh function.

Energy Distribution over Frequency (EDF). EDF was computed by applying STFT to each voiced segment across all channels. For each AC channel, the resulting spectrogram was averaged along the temporal dimension to obtain a frequency-wise energy profile. Channel 7 was again used as the reference, and all other channels were normalized by the maximum STFT energy of the reference channel. This normalization preserves relative spectral distributions across frequency bands while mitigating variations in overall signal amplitude. A summary of all feature-related parameters is provided in Table 8.

B.2 AuthG-Net Implementation

AuthG-Net is a lightweight multimodal neural architecture for user authentication (see Fig. 10). It jointly exploits AC, BC, and SF modalities to learn discriminative and passphrase-independent user representations. The model follows a three-stage pipeline: modality-specific encoding, feature fusion, and embedding learning with adversarial supervision.

Each modality is first processed by an independent encoder. For the AC and BC inputs, mel-log spectrogram features (Section 4.3.1) are used. All samples are zero-padded to a fixed size of $(215, 128)$ to ensure consistent input dimensions across the dataset. Each encoder consists of a single 3×3 2D convolution layer (stride 1, 32 output channels) followed by a ReLU activation. Since AC and BC share identical feature extraction settings, the same encoder architecture is adopted for both.

For the SF modality, three complementary features: TDT, ERT, and EDF are jointly utilized. TDT and ERT are naturally aligned with the temporal resolution of AC/BC features and are padded to $(215, 13)$. To mitigate instability in TDT when the signal energy is low, a masking strategy is applied based on the AC signal energy, zeroing out frames below a predefined threshold. EDF, originally of size $(86, 13)$, is temporally rescaled via linear interpolation to $(215, 13)$. The three SF features are concatenated along the temporal dimension and passed through a 3×3 2D convolution layer (stride 1, 32 channels), followed by a linear projection that expands the feature width to 128. ReLU activations are applied throughout.

The encoded features from AC, BC, and SF branches are concatenated along the channel dimension, producing a fused representation with 96 channels. This fused tensor is processed by a modified ResNet-18 backbone, where the first convolution layer is adapted to accept the increased number of input channels. The backbone retains all four residual blocks and outputs a 512-dimensional embedding after global average pooling and flattening.

To enforce passphrase-invariant representation learning, a domain adversarial learning (DAL) framework [32, 81] is employed. Two classifiers are attached to the embedding: a speaker classifier and a passphrase (content) classifier, each implemented as a three-layer fully connected network with a softmax output. The speaker classifier is optimized directly, while the passphrase classifier is connected via a Gradient Reversal Layer (GRL), which inverts gradients during backpropagation to suppress content-related information. The overall training objective is given by:

$$L(w_f, w_s, w_p) = L_s(w_f, w_s) - \lambda \cdot L_p(w_f, w_p),$$

where L_s and L_p denote cross-entropy losses for speaker and passphrase classification, respectively, and $\lambda = 1$. The parameters w_f, w_s, and w_p correspond to the feature extractor, speaker classifier, and passphrase classifier.

During training, each sample is annotated with both user identity and utterance labels, and the model is optimized using Adam. After convergence, the 512-dimensional embedding (highlighted in Fig. 10) is used as the user representation. For enrollment, embeddings are extracted and stored as templates. During authentication, cosine similarity is computed between the stored template and the query embedding; access is granted if the similarity exceeds a predefined threshold.

C RESULTS FOR EXPERIMENTS

C.1 Performance of the Original Implementation of Liveness Detection Methods

In Section 4.5, we evaluated the accuracy of previous methods on multi-channel microphone data in Benchmark Task 1. In some of the original papers, only a small number of microphones were used. Based on the descriptions in the original works, we also implemented the original versions of the selected methods. The liveness detection accuracy results are presented in Fig. 25.

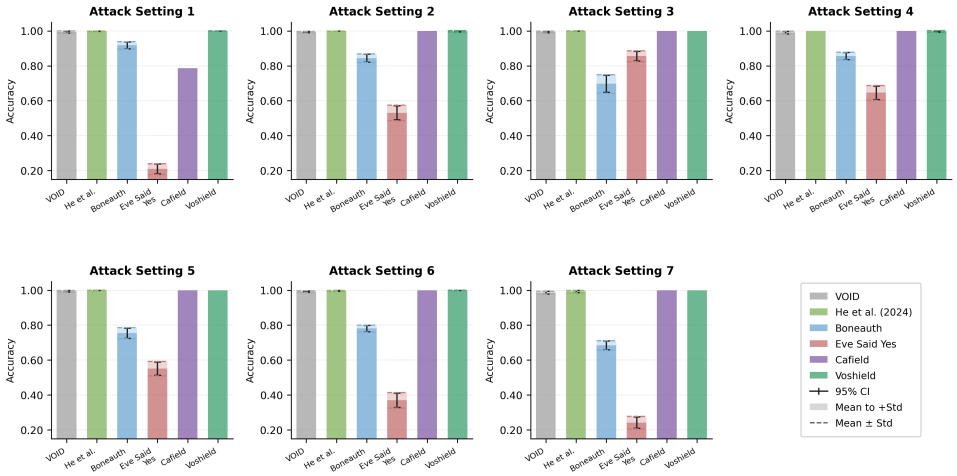


Fig. 25. Liveness detection accuracy for original baseline implementation.