SkullID: Through-Skull Sound Conduction based Authentication for Smartglasses

Hyejin Shin
Samsung Research
Seoul, Republic of Korea
hyejin1.shin@samsung.com

Jun Ho Huh
Samsung Research
Seoul, Republic of Korea
junho.huh@samsung.com

Bum Jun Kwon
Samsung Research
Seoul, Republic of Korea
bjun.kwon@samsung.com

Iljoo Kim
Samsung Research
Seoul, Republic of Korea
ij00.kim@samsung.com

Eunyong Cheon
Ulsan National Institute of Science and Technology
Ulsan, Republic of Korea
beer@unist.ac.kr

HongMin Kim
Ulsan National Institute of Science and Technology
Ulsan, Republic of Korea
khm489@unist.ac.kr

Choong-Hoon Lee
Samsung Research
Seoul, Republic of Korea
choonghoon.lee@samsung.com

Ian Oakley
KAIST
Daejeon, Republic of Korea
ian.r.oakley@gmail.com

ABSTRACT

This paper investigates the use of through-skull sound conduction to authenticate smartglass users. We mount a surface transducer on the right mastoid process to play cue signals and capture skull-transformed audio responses through contact microphones on various skull locations. We use the resultant bio-acoustic information as classification features. In an initial single-session study (N=25), we achieved mean Equal Error Rates (EERs) of 5.68% and 7.95% with microphones on the brow and left mastoid process. Combining the two signals substantially improves performance (to 2.35% EER). A subsequent multi-session study (N=30) demonstrates EERs are maintained over three recalls and, additionally, shows robustness to donning variations and background noise (achieving 2.72% EER). In a follow-up usability study over one week, participants report high levels of usability (as expressed by SUS scores) and that only modest workload is required to authenticate. Finally, a security analysis demonstrates the system's robustness to spoofing and imitation attacks.

CCS CONCEPTS

• **Security and privacy → Usability in security and privacy**; **Biometrics**.

KEYWORDS

Smartglass authentication, Bone conduction, Acoustic response, Biometric

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642506

ACM Reference Format:
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 CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3613904.3642506

1 INTRODUCTION

While the most widely deployed consumer offerings for smartglasses serve as interfaces to voice assistants [2], applications and devices that target other market segments [49] such as online collaboration [61], retail and shopping [51] and media consumption [6, 33] are rapidly developing. Similarly, enterprise use cases for smartglasses in domains such as health care [57, 58] and military operations [10, 46] are attracting considerable attention. These applications inherently rely on storing and accessing private data such as contacts, messages, financial credentials, photographs, videos, and patient records. Consequently, it is imperative to secure access to smartglasses. However, adapting existing authentication techniques to smartglasses is challenging due to their unique form factor. For example, passwords require precise, discreet entry of complex codes, something that is hard to achieve on smartglasses' small and prominently located input surfaces. Google Glass is a case in point: it incorporated a passcode based on entering a sequence of four taps and/or swipes on its small (74mm by 10mm) temple-mounted touchpad [67]. Users report that this type of code would be hard to enter, and express concerns about their codes being susceptible to observation attacks [60].

Biometrics, such as fingerprint and face or iris recognition, are a potential solution—users do not have to manually recall or enter passcodes, and they are robust to observation attacks. However, it is challenging to integrate traditional biometric channels into smartglasses as they typically require dedicated physical space for sensors. Additionally, they impose tight constraints on the situations in which signals can be captured—fingers, faces, or eyes must

be appropriately placed on or in front of sensors. In smartglasses, users may experience challenges in achieving such tasks. For example, users may find it difficult to accurately place their finger on an out-of-sight fingerprint reader mounted on the arms of a pair of glasses. These problems mean it is difficult to design and deploy effective authentication sensors in smartglasses. For example, the iris scanners proposed for the Apple Vision Pro [6] suit a fully enclosed head mounted display rather than a more lightweight form-factor, such as that typically used for voice-assistant frames [2].

The use of acoustic signals for authentication offers an appealing solution to this problem by combining the merits of a biometric approach without the requirement to integrate dedicated hardware—current smartglasses (and other wearables such as earbuds) are already typically equipped with advanced speakers (based on open-ear [2] and bone-conduction [19, 64] technologies) and sophisticated microphone arrays [2, 19, 64] sampling both in-air and in-body signals [59, 63]. To leverage such functionality, this paper builds on prior work highlighting the potential for developing implicit authentication systems for smartglasses based on the acoustic signals that pass through a user's skull [55]. Our goals are to extend prior work in four key ways: by capturing audio responses from various signals on contact microphones mounted directly on a user's skull (rather than from air-gapped in-air microphones); by exploring signals (alone and in combination) simultaneously recorded from microphones mounted on multiple (between two and four) skull sites; by conducting multi-session studies that measure usability data and explore the stability of the acoustic signals over time and; by performing a thorough security analysis characterizing the susceptibility of the system to physical imitation attacks (e.g., among anthropometrically similar participants), and signal replay attacks.

We realize these objectives by conducting a series of studies using smartglass prototypes of increasing maturity and robustness. These studies address four key research questions:

- Can audio signals transmitted directly through the skull serve as a reliable biometric measurement?
- What are the optimal signals to transmit and skull locations from which to sample biometric audio signals?
- Can authentication be comfortably achieved and reliable performance maintained over prolonged periods?
- Is the authentication system robust in response to typical attack vectors?

Major results include a robust demonstration of the viability of bio-acoustic signals transmitted through the skull to serve as a biometric (achieving best-in-class Equal Error Rates (EERs) of as low as 2.35%), and a recommendation to simultaneously sample audio responses from microphones on both the brow and mastoid process, prominent and readily accessible bony regions of the skull that promise high energy transfer intensity and consistent contact quality, in order to achieve this. In addition, longitudinal studies of multiple authentication sessions distributed both over a single day and spread more widely over a week, show modest drops in authentication performance: EERs increase to just 2.72% and 2.94%, respectively. Noise-condition experiments demonstrate the performance can be maintained in the presence of background audio noise and vibration interference. In comparison, a state-of-the-art system that uses an in-air microphone to capture air-transmitted signals [55] is heavily affected by audio noise, recording an elevated EER of 65.61% in the same experiments. Further, our cue signal type analysis reveals that consistent authentication performance can be achieved with a "speech cue" (a synthetic voice), and its performance is least affected by the training set size. We also demonstrate the feasibility of our system for integration into real world devices: on a low-resource Raspberry Pi, under-powered compared to current smartglass platforms, our optimized implementation takes just 30 seconds to train a full support vector machine (SVM) classifier. Finally, signal replay attacks showed 4 (out of 17) users could be spoofed. However, by adding specific attack training samples, we were able to build robust classifiers for all participants. In addition, imitation attacks performed using signals from physically similar individuals affected just one participant. Taken together, these results are the major contributions of this paper: detailed guidance for constructing practical bio-acoustic authentication systems for smartglasses, and a thorough report on their performance.

2 RELATED WORK

2.1 Biometric Authentication on SmartGlasses

Biometric authentication, such as fingerprint and facial recognition, is widely used on mobile devices [43]. While these techniques are well-established, there are inherent challenges integrating these methods into smartglasses. Fingerprint readers, for example, would be placed out of direct sight of the user, which may reduce reliability. A recent study of authentication methods [13] reported approximately half of the participants expressed concerns about using fingerprint readers placed out of sight on the back of a phone. Facial recognition approaches would likely face further challenges—cameras located on smartglasses would likely need extreme fields of view to capture the whole face. Alternative approaches based on capturing partial face images may be more promising. For example, Lim et al. [41] show that nose and cheek contour images, captured by a downward facing camera mounted on the nose bridge of a pair of glasses, can achieve 2% false rejection rate (FRR) and 4.97% false acceptance rate (FAR) in a three day study (N=20). However, camera based systems are inherently susceptible to variations in user appearance (e.g., wearing cosmetics, jewelry or a mask) and ambient light conditions (e.g., day vs night), which might affect performance. Finally, we note that while other biometric authentication techniques, such as iris scanning, may be a more natural fit to the glasses form factor, deploying rear-facing infrared cameras to capture the eyes may be expensive and require significant engineering efforts (e.g., Apple Vision Pro [6]). Literature suggests performance of this technology may also be less than ideal: Boutros et al. [9] explore the feasibility of verifying iris information through the use of two synchronized eye-facing infrared cameras, and report 6.35% EER. In addition, iris scanning biometrics are reported to suffer from usability problems [13].

Reflecting these hardware challenges, there has been considerable research interest in developing behavior-based biometrics for head mounted wearables. One approach has been to require users make simple head gestures in response to, for example, a specific request (to shake "no" or nod "yes" to answer questions) [68] or the timing and rhythm in a musical sample [40]. These approaches

show promise: both report EERs (or half total error rates) of approximately 4.5%. However both also incur usability costs. The workload involved in answering questions posed on a near-eye display is unknown and the accuracy of classifying head gestures remains imperfect (an additional 4% error rate in [68]), while the time required for rhythmic motions to yield sufficient information to authenticate users is prolonged at approximately 10 seconds. Such effort may be too high for populations used to rapid authentication systems—entering a PIN or pattern, for example, takes less than two seconds [47]. An alternative approach combines touch behavioral biometrics with voice biometrics to improve smartglass authentication performance [50]: this approach achieves a promising result (2.86% FRR and 1.27% FAR) in five different usage scenarios involving Google Glass. Despite this strong performance, we note voice biometrics are known to be sensitive to background audio noise [39], and prior work has suggested numerous usability issues with the small touchpad available on Google Glass [60]. Such practical factors may ultimately impact performance but have not been studied in depth.

2.2 Bio-Acoustic Authentication

Bio-acoustics based authentication has been studied extensively in the context of head-mounted wearables. A major focus has been on authentication with user generated sounds, specifically, speech. Liu et al. [42], for example, study the feasibility of recording speech-induced sounds through piezo-microphones in contact with a user's head, neck, or chest, and learning acoustic response patterns to authenticate users. They achieve 96.1% balanced accuracy, or 3.9% half total error rate (HTER), in a two session lab study conducted over two weeks (N=29). Feng et al. [15] suggest a continuous authentication system for voice assistants using a bone conduction microphone mounted on smartglasses, wireless earphones, or necklaces to continuously authenticate speakers. They achieve 3% FRR and 0.1% FAR. Gao et al. [17] explore the feasibility of using speech-induced in-ear acoustic responses to authenticate users on earphones, achieving 3.64% EER in a single-day lab study (N=23). This body of work highlights the uniqueness of self-produced sounds captured via bio-acoustic channels. However, despite their promising performance, these systems require a users' explicit vocalization, which may make users reluctant to authenticate in public spaces.

A smaller body of work has examined the potential for device-produced bio-acoustic sound to serve as a biometric. This scenario has the advantage that users are not required to take any action in order to authenticate: the system can trigger a sound and record a response independently. An early example of this approach is Skull-Conduct [55]. In this system, implemented on a stock Google Glass device, white noise is played through a bone conductance speaker behind the ear and transformed sounds are recorded on an in-air microphone in front of the brow. This system achieved 6.9% EER through a single session study (N=10). However, we note that as the microphone is air-gapped from the skull, the audio transmission paths in the system are unclear. More recent work tackles this limitation. For example, Isobe and Murao [32] augment the nose pads of a pair of glasses with piezoelectric disks. They use one as a speaker and the other as a microphone and study the feasibility of using through-nose acoustic response to a chirp signal for smartglass authentication. They achieve 9% EER in a single-day study (N=11). More promising performance is reported using earphones. Gao et al. [18] examine the use of in-ear sound reflections to authenticate users on wireless earphones, achieving 94.52% balanced accuracy (equivalent to 5.48% HTER) in a single-day lab study (N=20). Wang et al. [65] also use in-ear sound reflections for user authentication but, unlike Gao et al. [18], they use reflections caused by ear canal deformation while speaking. They achieve 97.38% recall and 95.02% precision in an immediate recall session (N=24), and 95% accuracy in a subsequent session performed after 4 months (the participant return rate is unknown). Finally, Irwansyah et al. [31] examine through-skull sound for user authentication by transmitting an impulse from a bone conduction transducer located behind the ear, and capturing responses on an in-ear microphone: they report 2.6% EER (N=10).

This paper extends this prior work. Specifically, we recognize the benefits of bio-acoustic solutions that rely on system-produced sound for authentication (and hence do not require explicit user action) and conduct an extensive evaluation of their potential for a smartglass form factor device. Specifically, we extend prior work on smartglasses by 1) using a bone conduction speaker and microphones to explore signals transmitted directly through the skull (rather than over an air gap [55]), 2) examining the potential of multiple skull sites for capturing bio-acoustic responses, 3) experimenting with different cue types, and 4) studying the effects of microphone contact variability and background noise on authentication performance. These practical concerns are, by and large, omitted in prior work. We provide a more detailed description of the key differences between our work and prior research with respect to the sensors used, required authentication actions, studied (experimental) conditions, and reported accuracy in Appendix A.1 Table 2.

3 DESIGN AND IMPLEMENTATION

We describe the hardware implementation, overall data processing pipeline, and feature engineering details.

3.1 Theory of Operation

The human face is our most distinct and distinguishable feature. One factor that contributes to the uniqueness of each human face is variations in the underlying bone structures—human skulls differ substantially from one another. Some differences are readily observable. Large scale surveys in the US, for example, indicate male mean head circumference is 57.44 centimeters (SD=1.6) while for women it is 56.11 centimeters (SD=1.94) [22]. In addition, head shapes also differ considerably, both in terms of the simple ratio between length and breadth [37], and also through many other aspects of their form (e.g., in terms of the protrusion and angle of the brow and brow ridge) [71]. Other differences are harder to observe: the skull is composed of 22 separate bones [3], and size and thickness of each varies from individual to individual [4]. For example, the frontal, parietal and occipital bone of female skulls are reported to be, on average, between 4.2% and 12% thicker than those in male skulls [38].

These natural anatomical and physiological differences impact the bio-acoustic responses that can be captured from (or heard in)

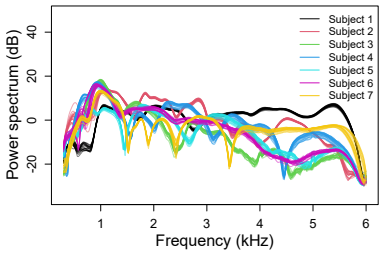

Figure 1: Power spectrum of 7 subjects. 20 samples collected on the mastoid when the chirp cue (with a frequency range from 50 Hz to 6 KHz) was used.

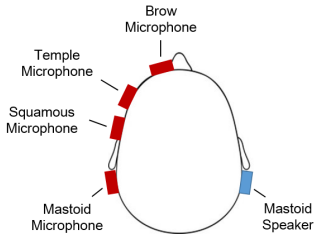

Figure 2: Speaker and microphone locations explored in the first study.

the skull. They are reported to exert wide ranging effects from influencing medical imaging procedures [35] to affecting musical preferences [62]. In a representative and detailed analysis of the spectral responses generated from skull conducted sounds captured from 30 individuals, Gordon et al. [23] concluded that transmitted energy levels differed substantially between each individual across a broad range of frequencies (25 Hz to 5 kHz). Figure 1 illustrates this point. It shows the power spectrum of twenty response signals collected from seven individuals during playback of a chirp sound (rising from 50 Hz to 6 kHz) on a bone conductance speaker placed on right mastoid process. Signals were recorded with a contact microphone placed on the left mastoid process. Within subject power spectrum patterns were consistent while the patterns among the seven individuals were highly diverse. These two characteristics indicate that through-skull acoustic responses can serve as an effective biometric for identifying smartglass users.

3.2 Hardware Design and Implementation

We created three different prototypes in this work in order to support different study objectives. Each shared common elements in terms of the hardware used, the sounds played, and the speaker and microphone locations targeted on the skull. We describe these common aspects here and the specifics of each prototype in the individual study sections.

In terms of the hardware, we used the same microphones and speakers throughout. Specifically, we sensed responses with Knowles BU-21771 contact microphones, a device previously deployed in a range of closely related prior work studying through body audio transmission [25, 53]. To boost signals to line levels, we used a modified version of the amplifier design proposed for this device by Zhang et al. [69]. Modifications including removal of post-amplification filters (as these showed few improvements to signal quality in pilot tests) and tests with a variety of different gains (ultimately customized for each prototype). For actuation, we used a commercially available surface transducer[1] also deployed in prior work on active in-body audio transmission [70]. We drove the speaker using a breakout for the TPA2012 class D audio amplifier[2] configured for 24 dB gain and powered by a 350mAh 3.7V li-ion battery, a typical specification for a wearable device. In terms of design, amplifiers were always located immediately adjacent to microphones or transducers in order to minimize the impact of RF noise. Additionally, sensors and speakers were always encased in 3D printed in skin-safe Thermoplastic polyurethane (TPU), specifically NinjaTek SemiFlex[3], in such a way that a thin film of TPU (0.35 millimeters) covered actuator/sensor surfaces. For audio cables running external to our prototypes we used 3.5 millimeter audio jack connectors and shielded cables.

In terms of sounds, we used three cue signals in total; two were dropped in our final study. The sounds were (i) a chirp from 50 Hz to 6 kHz, (ii) a mobile phone wake-up melody, and (iii) speech, in the form of a short spoken phrase. The melody is the start-up sound for a polyphonic Samsung feature phone. The speech was a computer-generated voice (Google's WaveNet voice G) uttering the words *"Welcome. Authenticating User."* Each individual cue sound was two seconds in duration and normalized, over its full length, to an amplitude of -15 dBFS.

Finally, we considered a single speaker location and four different microphone locations in this work. We retained only two of these microphone locations in our second and third studies. The speaker location was always the right mastoid process—the bony protrusion immediately behind the ear. We selected this location as it is a common location for the bone conduction speaker in numerous prior skull response studies [12, 14, 23, 44] and also in commercial smartglasses (e.g., Google Glass). Its proximity to the ear, and relative lack of interference/obstructions (e.g., hair) make it a suitable location for delivering audible in-skull sound. The four microphones location were: the left mastoid process; the left temporal squamous (the area immediately above and in front of the ear), the left temple and the brow (either offset to the left, or centrally located). We selected these sites as they are reasonably aligned with the physical structures of many commercial brands of smartglasses such as Google Glass, Bose Frames, Vuzix Blade, Microsoft HoloLens, Lenovo ThinkReality, and Apple Vision Pro. These speaker and microphone locations are illustrated in Figure 2.

3.3 System Overview

Enrollment. This initial activity involves collecting users' own reference response signals, pre-processing collected signals, extracting features, and training user-specific authentication classifiers. During enrollment, users are required to put on their smartglasses several times, and remain still while a cue signal (e.g., a system wake up tone) is played through a surface transducer mounted on the right mastoid (located just behind the ear). Microphones resting on the skull are used to measure bone conduction responses. These response signals are processed into the signal processing features

[1] https://www.adafruit.com/product/1674
[2] https://www.adafruit.com/product/1552
[3] https://ninjatek.com/

described in Section 3.4. Multiple samples are collected each time the glasses are worn (hereafter referred to as a *donning*). When data collection is complete, a user-specific binary classifier is trained using the features extracted from those reference signals and a pre-deployed imposter/other-user train set.

Authentication. After users don their glasses, a cue signal is automatically played through the transducer. Response signals are collected through the microphones, and processed into features. These features are submitted to the trained classifiers to generate authentication (probability) scores, which in turn, are compared against a pre-defined threshold value: users are successfully authenticated if a score is higher than the threshold.

3.4 Pre-processing and Feature Extraction

We initially process data from each microphone separately. We first apply a bandpass filter between 300 Hz and 19 kHz to reflect the frequency response range of our surface transducer [1]. We used power spectral density (PSD) and Mel-frequency cepstral coefficients (MFCCs), both of which are popular features in speech recognition and sound processing, as our feature set. Based on the data collected through the first study, we experimented with a wide range of signal processing features, including PSD, MFCCs, short time Fourier transform (STFT), and correlation between two pairwise signals (e.g., coherence). The "MFCC and PSD concatenation" feature set showed peak performance. Our PSD computation method involves calculating the periodogram of the entire signal using FFT, and smoothing that periodogram with 50 Hz windows [11]. We then sample PSD at every 100 Hz in the range from 300 Hz to 6 kHz, resulting in a log-PSD feature vector containing 58 features. We selected PSD features up to 6 kHz based on examining energy transfer measurements which suggested that minimal energy was transferred through the skull above 6 kHz. This observation is inline with the prior findings [26]. As for MFCCs, we computed 13 cepstral coefficients using a Hamming sliding window (window length of 25ms and overlap of 10ms). To compress the time-series data, we computed statistical features (mean, standard deviation, skewness, and kurtosis) to create an MFCC set consisting of 52 features. We then concatenate PSD and MFCC features into a final one-dimensional vector of 110 features.

3.5 Classification Algorithms

We built our authentication classifiers using a binary support vector machine (SVM) with radial basis function kernel. Based on the first study data, we evaluated the performance of several algorithms, including SVM, random forest, and XGBoost. SVM showed the best performance. While examining the effects of combining microphone locations, we observed that a feature concatenation technique achieves optimal performance on the first two prototypes but a weight-adjusted ensemble model shows superiority in the third prototype. We attribute this to different prototypes resulting in various microphone contact quality profiles. We report the optimal performance accordingly in each study.

3.6 Threat Model

We consider two possible threats: signal replay attacks, and imitation attacks. Signal replay attacks would involve several steps. First, the attacker needs to compromise a victim's response signal (authentication secret), either using a piece of malware installed on the victim's glasses or using a separate recording device. Second, the attacker needs to inject this signal on their own head (e.g., by mounting a separate surface transducer) at the exact moment they seek to bypass authentication on the stolen glasses. We imagine this kind of replay attack is inherently challenging to perform: injected signals would be deformed through a new signal passage (different skull and skin arrangements), and likely result in a signal pattern unrecognizable by the authentication system. We conduct a separate attack study to demonstrate the robustness against such signal replay attacks (see Section 7.1). A more feasible attack involves recruiting imitators, or individuals who are anthropometrically similar to the victim (e.g., with respect to head sizes, weight, and height), and simply asking them to don the stolen glasses and go through the authentication steps. If the signal transformations of the imitators closely match the victim's signal transformations, the authentication system may be spoofed. In Section 7.2, we study the performance of such imitation attacks.

4 STUDY 1: EXPLORING MICROPHONE LOCATIONS

To answer the first two research questions (see Section 1), we conducted a single-session lab study designed to experiment with various cue signal compositions, and collect bone conduction signals from four different microphone locations. This section reports the study methods, comparative accuracy results, and our final recommendations for selecting multiple microphone locations.

4.1 Headband Prototype

Figure 3 shows the headband prototype used in this study. It was equipped with a single bone conduction speaker (surface transducer) and four bone conduction contact microphones. To mount these devices we designed simple plastic clips that slid onto a GoPro headband [21]. This allowed us to locate the units firmly on the head (by selecting an appropriate band tightness) and to adjust the position of each device by simply sliding it along the band. The bone conduction speaker was mounted on the right mastoid process and, as shown in Figure 2, the microphones were located on the left side of the head at all four locations: the brow, temple, squamous, and mastoid. The brow microphone was located directly above left eye. In order to simultaneously record signals from all four microphones, we used a commercial audio interface: the Focusrite Scarlett 18i8 [16]. We configured this device to simultaneously record four continuous 44.1 kHz mono audio samples using Pro Tools [7], a commercial audio recording software package with all post-processing disabled. We powered the input amplifiers using an externally located 9V battery. We drove the speaker from a dedicated audio playback device (a Teensy 4.1[4] with its audio board expansion module[5]) configured to play back the full set of audio samples needed in each study session at the press of a physical button. This ensured that audio playback levels were consistent throughout the study.

[4]https://www.pjrc.com/store/teensy41.html
[5]https://www.pjrc.com/store/teensy3_audio.html

Figure 3: Headband prototype on a participant's head (left), and connections to audio interface (right). Right shows four microphones (center of image) connected via an external power supply (9V battery) to a Focusrite Scarlett 18i8 audio interface (in red). This device connects via USB to a PC to deliver multi-channel audio signals.

4.2 Procedure and Measures

Each study participant wore the headband prototype five times. To ensure similarity between each donning, study moderators maintained the band tightness (i.e., length) across donnings and sought to minimize variations in the speaker and microphone locations. To record the consistency of these efforts, moderators took pictures of participants in every donning session while obscuring their face with a paper screen (see Figure 3). During each donning the three two-second cue signals (chirp, melody, and speech) were played on the bone conduction speaker mounted on the right mastoid 20 times and the resultant acoustic responses were captured from the four microphone locations. We left a 0.25 second gap between each cue and prefixed playback of each set of three cues with a 0.5-second 14 kHz sinusoidal tone. This was included to facilitate accurate segmentation of the recorded sounds during analysis. As such, we logged 1,200 samples (5 donnings×3 cues×4 locations×20 samples) from each participant. After the data collection phase was complete, participants were asked to complete a short demographics questionnaire. The study took approximately one and a half hours to complete, and the participants were compensated with the equivalent of 13 USD in local currency. The study protocols presented in this section and in all subsequent sections were approved by the university's IRB.

4.3 Demographics

We recruited a total of 25 participants: 21 were male, and four were female (see Table 3 in Appendix A.2). The average age was 25.3 years (SD=3.0). The difference between the smallest and largest head circumferences was 5.6 centimeters. All four female participants had shoulder length hair while 86% of male participants had short to medium length hair, and 14% had very short hair. We asked female participants to tie their hair back.

4.4 Authentication Evaluation Setup

As a preliminary feasibility check, we first evaluated the performance of multi-class SVM classifiers (treating it as an user identification problem) using the MFCC/PSD concatenation feature set. User identification accuracy was between 93–98% when information gathered through multiple microphones was used together; full details are explained in Appendix A.3.1.

Next, we evaluated how the user authentication (binary classification) performance differs between the four microphone locations and three cue signals. We divided the 25 participants into two groups: 15 participants were used for imposter training and the remaining 10 participants were set as genuine users. For each genuine user, the imposter set was used to train a binary classifier and the other 9 genuine users served for unknown user (imposter) testing. We repeated the performance evaluation of trained classifiers over 100 random permutations of 10 genuine users and 15 imposters. We report all results as the mean and standard deviation over 100 permutations.

For each genuine user, microphone location and cue type, we selected all samples from the first four donnings to create an 80 sample genuine training set. Twenty samples from the last donning session were used for genuine user testing (to measure FRRs). To create a balanced train set, we randomly selected 5 samples from each of the 15 imposters to form a 75 sample imposter train set. For unknown user testing (measuring FARs), we used all samples from the remaining 9 users in the genuine user set, creating an imposter test set consisting of 900 samples. In addition, when combining the features from multiple microphones, we used a simple feature concatenation approach. Finally, we used equal error rate (EER) as an evaluation metric to measure a single (overall) error rate across all genuine users. It is defined as the error rate at which FAR and FRR are equal.

Table 1: Authentication accuracy for each cue type by all possible combinations of one to four microphone locations expressed as mean EER and SD (in brackets) measured over 100 random permutations.

Location	EER (%)		
	Chirp	Melody	Speech
Brow	7.03 (3.42)	5.68 (2.69)	6.87 (3.54)
Mastoid	7.57 (4.18)	7.95 (3.29)	8.19 (3.66)
Squamous	9.24 (3.83)	9.93 (4.37)	8.88 (3.50)
Temple	10.50 (9.32)	9.69 (3.69)	8.94 (4.06)
Brow/Mastoid	2.84 (1.97)	2.35 (1.45)	2.58 (1.75)
Brow/Squamous	1.92 (1.53)	2.43 (1.37)	1.66 (1.06)
Brow/Temple	4.53 (3.04)	2.27 (1.42)	4.39 (3.19)
Mastoid/Squamous	4.01 (2.38)	4.90 (2.78)	3.81 (2.31)
Mastoid/Temple	4.14 (3.08)	4.36 (2.67)	2.56 (1.91)
Squamous/Temple	3.52 (2.68)	6.63 (3.65)	4.50 (3.80)
Brow/Mastoid/Squamous	0.95 (0.93)	1.59 (1.17)	0.90 (1.01)
Brow/Mastoid/Temple	1.14 (0.85)	1.37 (1.06)	0.95 (0.98)
Brow/Squamous/Temple	2.08 (2.12)	2.15 (1.28)	2.33 (2.46)
Mastoid/Squamous/Temple	1.85 (1.68)	4.21 (3.35)	1.37 (1.18)
Brow/Mastoid/Squamous/Temple	0.83 (0.77)	1.44 (0.96)	0.64 (0.73)

4.5 Authentication Accuracy

Table 1 shows the mean EER for each of the four microphones and three cue types computed over 100 permutations. Pairwise comparison results that show statistically significant differences (with respect to pairwise t-tests) are summarized in Table 4 (see Appendix A.3.2). These results suggest that the brow and mastoid locations outperform the squamous and temple locations with meaningful differences (Bonferroni corrected t-tests $p < 0.05$).

To investigate the effects of combining multiple microphone locations on authentication accuracy, we experimented with all possible combinations. Table 1 presents the mean EER for two, three, and four microphone combinations, computed over 100 permutations: in general, using more than one microphone has a positive impact on accuracy, implying that each location offers unique

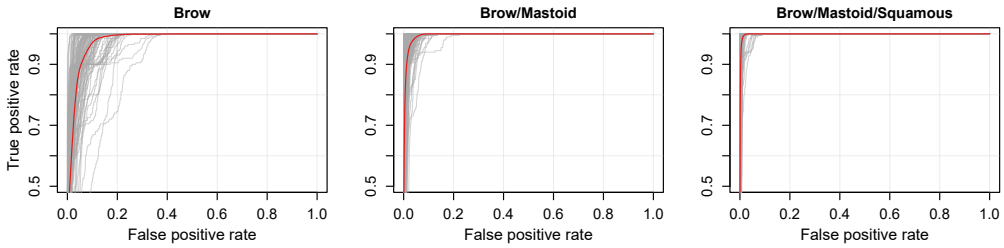

Figure 4: ROC curves over 100 permutations with the average performance (red line) for the selected locations when the speech cue was used.

bio-acoustic information about a user. The brow and mastoid combination achieved significantly lower error rates (2.35–2.84% EERs) compared to when the brow was used alone. These results are shown in Figure 4. The brow, mastoid, and squamous combination showed further reduction in the error rates (0.90–1.59% EERs). We achieved the optimal performance by combining all four locations (0.64% EER) while using the speech cue.

Next, we evaluated how the performance differs between the three cue signals. Table 5 (see Appendix A.3.3) summarizes comparisons that were statistically significant. Noticeably, on the brow alone, the melody cue demonstrated superiority over the chirp and speech cues. The speech cue demonstrated superiority over the other cues in the majority of cases using multiple microphones.

4.6 Microphone Location Analysis

The study results indicate that authentication accuracy varies substantially by microphone location. This section explores the factors that contribute to such variations. First, we examined energy transfer intensity from each microphone. A transfer function measures how the cue signal s transforms into the response signal x. At each frequency f, it is computed as $H(f) = P_{xs}(f)/P_{ss}(f)$, where P_{xs} is the cross-PSD between x and s, and P_{ss} is the PSD computed on s. Figure 5 compares the mean energy transfer intensity from each of the four different microphones. The brow, an unobstructed bony prominence, showed the highest energy transfer intensity: the largest area under the curve. The squamous, which can feature obstructions such as hair, showed the lowest intensity [48]. Examining the relationship between energy transfer intensity and EER via Pearson correlation showed a clear negative trend (Pearson's $r = -0.61$, $p = 0.034$): improved EERs are associated with higher intensity signals.

Second, we examined the consistency of PSD data over each participant's donning sessions using Wilks' Lambda. Specifically, we computed $\Lambda = |W|/|T|$, where $|W|$ and $|T|$ are, respectively, the within-subject and the total variances. Small Wilks' Lambda implies small within-subject variation (i.e., high consistency in signals within subject). The brow and mastoid locations show high consistency, with a notable drop in performance on the temple and, in particular, on the squamous (see Appendix A.3.4 Figure 14). We suggest that consistency may be difficult to achieve on these latter two locations due to the complexity of the shape of the skull and interference due to factors such as hair.

Third, we inspected how using microphone combinations improves performance. Sixteen participants showed good FRRs on all four microphones consistently over 100 permutations while the remaining nine participants showed poor authentication performance on at least one microphone. Figure 6 shows the per-user FRR (averaged over 100 permutations) at each of the four locations for these nine participants: note, all achieved a low FRR (below 2%) on at least one location, implying that with an adequate strategy for combining microphones, all users would experience consistently low FRRs. Importantly, in the two participants who failed sessions for the brow, the mastoid microphone contributed reliable data: in consequence, the brow and mastoid concatenation results showed stable FRRs. Similarly, in all cases in which the mastoid signals failed, the brow microphone showed good performance, also leading to stable, low FRRs on concatenated features. A candidate explanation for these trends is that, in some participants, misplacement of one microphone led to a reduction in the reliability of the responses it captured. However, such misplacements did not propagate to the other microphone, meaning that the poor quality signal from one did impact the reliable signals captured from the other. Thus, the two microphones complemented each other in all failed cases to stabilize error rates. Taken together, these observations emphasize the importance of designing a system that supports multiple microphones: even if one microphone is poorly positioned, the other microphone may be in good contact, and provide sufficient information for the authentication to succeed.

Based on these analyses, we recommend using microphones on the brow and mastoid process for bio-acoustic authentication. These two locations are relatively unobstructed, lead to high signal amplitudes, show consistent performance over time and are relatively independent of one another—variability or poor signal quality occurring on one of these locations can often be compensated for by stable and high quality signals captured on the other.

5 STUDY 2: MULTI-RECALL SESSION PERFORMANCE

To answer our third research question, we constructed a more stable frame-based prototype (based on the Study 1 findings), and conducted a multi-session recall study with 30 new participants. This study was designed to investigate extended recall performance over a single day, and the interfering effects of background noise on recall performance. We also collected response signals through an in-air microphone, and used this data to compare the performance against the SkullConduct system [55].

5.1 Glass Frame Prototype

The new prototype developed for this study took the form of a pair of glasses that, following recommendations from the first study,

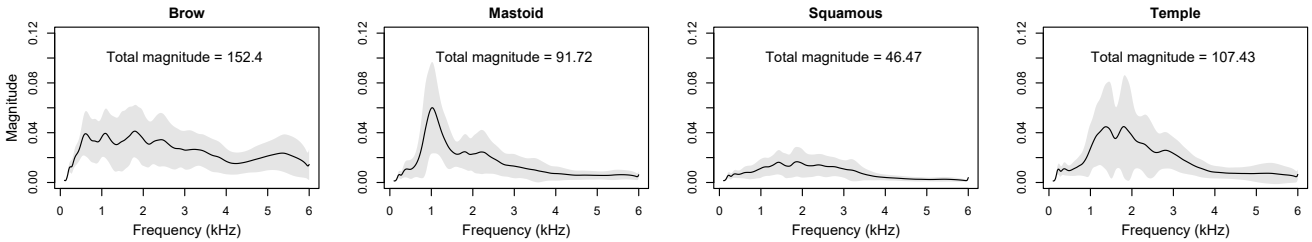

Figure 5: Magnitude of the transfer function averaged across 25 participants with standard deviation for four microphone locations when a chirp from 50 Hz to 6 kHz was used.

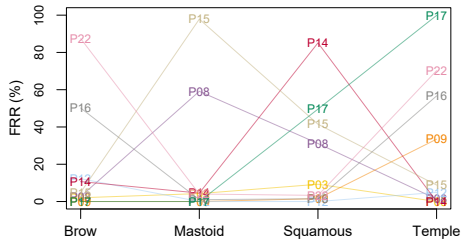

Figure 6: Average FRRs of 9 participants whose FRRs were larger than 5% on at least one microphone.

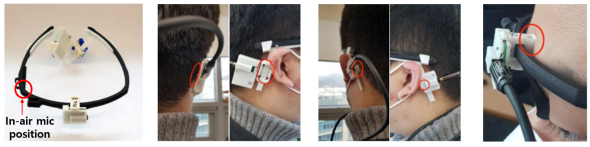

Figure 7: Smartglasses prototype. Left shows the unworn assembled prototype, while center-left, center-right and right show fit, respectively, for the mastoid speaker, mastoid microphone, and brow microphone. In each of these panels the location of the speaker or microphone is highlighted with a red ellipse.

located microphones on the brow and left mastoid and maintained the speaker on the right mastoid. The prototype was built on a commercially available open-source glasses frame (part of the open-source Pupil core eye tracking system [34]). This frame is robust, flexible and designed to mount stably and firmly on the heads of a very wide range of users. To adapt this frame for our purposes we designed and fabricated mounts for the speaker and microphones. Mastoid mounts clipped to the arms of the frame and featured an angled strut that extended downwards and inwards from the tip of the arms. The mastoid speaker and microphone modules, which were functionally identical to those used in the first study, were located on these struts. We fixed a location for these units through pilot studies with a range of representative local participants: although head size naturally varies, we selected a location which achieves good contact with the mastoid for a broad range of typical participants. The brow microphone simply clipped to the center of the front arc of the frame: directly in the center of the brow. Locating the microphone centrally ensured better contact with the brow for all participants, as the left brow regions (used in the first study) exhibited a range of curvatures that rendered reliable placement impossible. The brow microphone made contact with the brow simply by ensuring the frame was tightly pulled back on the head. This prototype is shown in Figure 7, including close up depictions of the microphone and speaker placements on the head of a typical participant. In addition to these components, we added an in-air microphone to the system. Specifically, we placed a RODE Lavalier GO microphone [36] in a clip on the right hinge of the glass frame, pointed down towards the wearer's mouth. This is a typical location for an in-air microphone in a pair of smartglasses, as it affords capture of the wearer's speech. We included this microphone to support comparison of the performance of bone-conductance microphones with the in-air microphone studied in SkullConduct [55].

To operate this prototype, we selected the Teensy audio system, used for playback in the first study, to also record all sounds. This simplified executing the study, as moderators had fewer devices to manage. Given the prolonged structure planned for this study, we expected this simpler platform to reduce protocol errors and increase reliability. In addition, the Teensy platform can supply power to input amplifiers, simplifying the physical set up. It is also able to accurately synchronize simultaneous recording and playback of 44.1 kHz 16 bit audio, removing the need for a cue segmentation step during analysis. To enable recordings from the three microphones simultaneously, we configured the Teensy with two audio expansion boards. To ensure reliable signal recording, all samples used in the study were played and recorded individually, with all data read from and stored in RAM. This made sure that latencies (e.g., due to read/write delays) were minimal. After each sample had been played, and the responses recorded, the data was immediately transmitted to a host PC for analysis. The three two-second audio samples used as cues during this study (chirp, melody and speech, as in the first study) were stored on an SD card and loaded into RAM as needed to support the study protocol.

5.2 User Study Methodology

In this study, we recruited participants to be either genuine users or imposters. Genuine user participants completed four separate study sessions: an initial enrollment session followed by three separate recall sessions, each a minimum of 30 minutes apart. These recall sessions emulate users who might don and re-don their glasses fairly frequently—to clean them, to rub their eyes, or to complete various personal hygiene routines. The study took place in a normal office environment with 30–40 dB noise. The enrollment sessions involved participants putting on and taking off the glasses frames six times. In each donning, we played 20 examples of the three cue sounds (chirp, melody, and speech) used in the first study. In each of the three

recall sessions, participants experienced 20 of each of the three cue sounds in normal conditions (akin to those during enrollment) and a subsequent 20 of each cue while a distractor sound clip was played in the air as a background noise. We used a different distractor in each recall session: music, speech and finally ambient noise. Each distractor was adjusted to be between 50 and 60 dB in volume. In total, we collected data from 3 microphones by 20 examples by 3 audio cues by 12 sessions (6 during enrollment and 3 by 2 during recall) for a total of 2,160 audio clips per genuine user participant. Imposter participants engaged in a reduced data collection process. They visited only once and completed two donning sessions. This resulted in 3 microphones by 20 examples by 3 audio cues by 2 sessions or 360 samples per imposter. In all sessions we asked participants to put on and take off the glasses by themselves to capture natural donning behaviors and variations. Genuine users were compensated with the equivalent of 26 USD in local currency and imposters with 13 USD.

5.3 Demographics

We recruited 30 participants in total (see Appendix A.2 Table 3). 23 were male and seven were female. Half were considered genuine users and half imposters. The genuine user evaluation set consisted of 11 male and four female participants—their average age was 23.3 years (SD=2.3). We did not observe any statistically significant difference between the genuine users and imposters with respect to their physiological characteristics such as age, height, weight, and head circumference (two-sample t-tests $p = 0.571$, 0.873, 0.696, and 0.569, respectively).

5.4 Evaluation Setup

We trained per-user binary classifiers based on the concatenated MFCC/PSD feature set by selecting genuine samples from the five enrollment donnings. The samples from the remaining sessions were used to measure FRRs. A fixed imposter train set was selected from the 15 imposters: we selected three samples from each of the two sessions to generate a set of 90 samples. FRRs were measured based on the immediate recall (sixth donning session), followed by three prolonged recalls performed under normal conditions, and three additional recalls performed in the presence of audio noise. To create an imposter test set for a given genuine user, we selected all samples from the other 14 users in the genuine user set, creating a test set consisting of 3,360 samples per cue signal.

5.5 Authentication Accuracy

Figure 8 shows the ROC curves by microphone locations and cue sounds for all genuine users. The brow and mastoid combination reported EERs of between 2.72–4.22% across all recall sessions. It is worth highlighting several aspects of these results. First, compared to the first study, immediate recall rates are notably improved to between 0.03% and 2.61% from a single microphone. We attribute this to the increased stability of microphone placement enabled by our glasses format prototype. Second, recall rates from subsequent sessions are somewhat elevated compared to immediate recall, with data from a single microphone achieving between 3.44% and 9.82% EERs. However, we note that concatenating EERs from both microphones results in EERs of between 2.69–4.56%. This validates our prior claim about the importance of capturing signals through multiple channels to stabilize performance.

5.6 Robustness to Audio Noise

To investigate the effect of a noisy environment, we compared the performance between the regular recall sessions and those featuring background noise. The error rates are presented in Figure 8. Background noise led to minor changes in EERs when both microphones were used: 4.56% vs. 4.42%, 2.69% vs. 3.24%, and 3.17% vs. 2.81% for the chirp, melody, and speech cue types, respectively. Bonferroni corrected pairwise t-tests did not find any statistically significant differences in the mean FRR and FAR values between the regular and noisy recall sessions (all p values range between 0.21 and 0.87). This result is expected: the surface transducers in our prototype are minimally responsive to air-transmitted sounds.

5.7 Cue Type Usability

During the post-study survey, we asked the participants to rank the three cue types in the order of preference from one to three, one being the "most preferred" and three being the "least preferred" option. Out of the 30 participants, 16 participants chose the wake-up melody as their top preference; 13 participants chose the speech sample, and just one participant preferred the chirp sample. The average preference ranks for the chirp, wake-up melody, and speech sample were 2.63, 1.70, 1.67, respectively. The chirp, a highly artificial sound, was clearly disliked.

5.8 Performance of In-Air Microphones

In this section, we investigate the effects of collecting response signals through a standard in-air microphone on recall performance. To find an optimal feature set for classifying in-air microphone samples, we first experimented with MFCC delta and MFCC delta-delta features used in prior work [55]. However, our MFCC/PSD concatenation features, alone, outperformed all other feature combinations, and demonstrated top performance on the last enrollment session samples. Hence, we performed all subsequent comparative analyses based on our MFCC/PSD concatenation feature set. To validate the correctness of our SkullConduct implementation, we applied the one-class based evaluation method employed in [55] on the 15 genuine users, achieving 4.40%, 9.24%, and 4.34% EERs on the immediate recall data for the chirp, wake-up melody, and speech cues, respectively. These results are comparable to the 6.9% EER reported in the original article [55]. Figure 15 (see Appendix A.4) shows the recall performance for the in-air microphone: all three graphs demonstrate significant elevations in the EERs compared to those reported for SkullID in the normal sessions (9.54–14.20%). Additionally, we observed substantially higher EERs (58–66%) when background noise was played. Figure 9 explains this. It depicts the high variances that appear in PSDs computed on the air-transmitted signals when background noise is present. We conclude standard in-air microphones are impractical for reliable bio-acoustic authentication: they are simply too susceptible to background audio noise.

5.9 System Optimization

This study used a substantial amount of data for enrollment: 100 repetitions of each of three samples over five smartglass donnings.

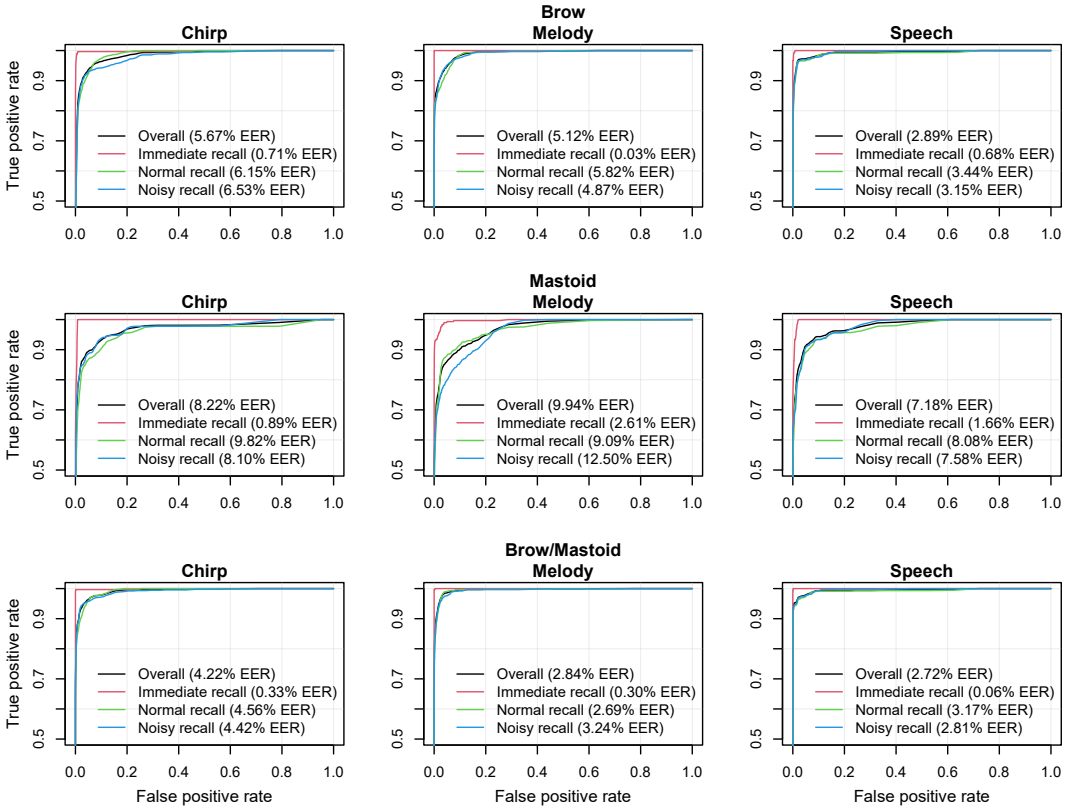

Figure 8: ROC curves by microphone location and cue type. "Immediate recall" represents the performance at the sixth donning in enrollment session, "Normal recall" the performance over three normal sessions, and "Noisy recall" the performance over three noisy recall sessions. "Overall" represents the performance over all seven recall sessions.

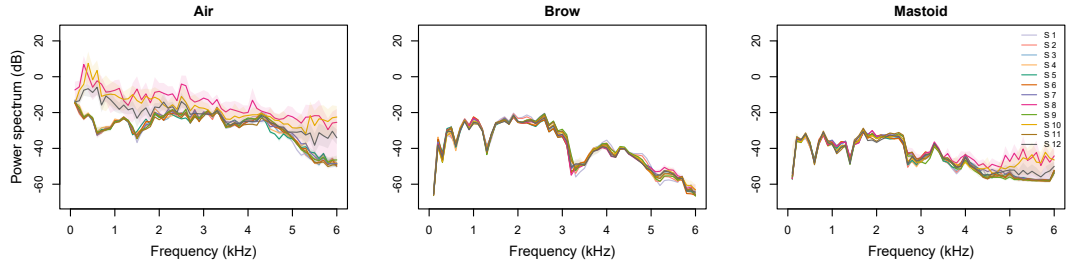

Figure 9: The average power spectrum for one subject across all twelve (both enrollment and recall) donning sessions based on the speech cue. Sessions 7, 9, 11 represent normal recall sessions, and sessions 8, 10, 12 represent noisy recall sessions.

Such prolonged effort represents a major burden to users. Accordingly, we examined our data to make recommendations about how to optimize enrollment procedures. We first investigated the effects of reducing the number of training donning sessions and samples per donning session. Data showing the resultant performance variations are in Figure 10. This figure indicates that recall performance tends to stabilize after between three and five donnings; erring on the side of caution, we recommend maintaining five donning sessions to capture signals representing natural variations in device fit during enrollment. Additionally, we note melody and speech cues are relatively unaffected by the number of samples captured per donning: performance with five samples is close to that attained with 20. Accordingly, we recommend that capturing five samples per donning is appropriate during enrollment. Finally, training on three audio samples is unnecessary. Based on its good performance with low enrollment repetitions, its consistently low error rates throughout and its high popularity with users, we recommend use of our speech cue alone.

5.10 Overheads

To analyze the model training and authentication time overheads, we measured the MFCC and PSD feature extraction time and the SVM training and prediction time using the Python `sklearn` libraries on both a Raspberry Pi 3 Model B (1.2GHz CPU, 1GB RAM) [52] and a Linux desktop equipped with an Intel i7-9700 Processor (3.00GHz CPU, 16GB RAM). We logged this data for all

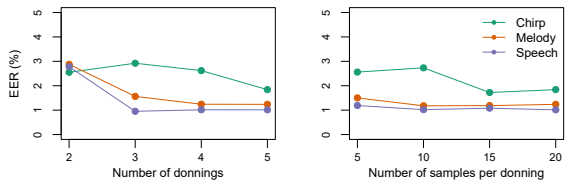

Figure 10: EERs with respect to varying number of enrollment donnings (left) and number of samples per donning (right).

15 genuine users assuming 25 genuine samples and 30 imposter samples are used to train each classifier. On average, it took 620.40 (SD=202.57) and 19.10 (SD=4.33) milliseconds to extract both MFCC and PSD feature sets from a single two-second mono audio clip, and 32.26 (SD=5.32) and 3.15 (SD=0.36) milliseconds to train a classifier with the speech cue on the Raspberry Pi and the PC, respectively. Assuming imposter feature vectors will be pre-loaded on the target device, we therefore calculate the full SVM binary classifier training time as the time required to extract features from 50 two-second audio clips (25 genuine user recordings from each microphone) plus the SVM classifier training time. Real-time execution of this full model training pipeline will therefore take approximately 30 seconds on the Raspberry Pi and 0.5 seconds on the PC. The size of the extracted features from a single two-second audio mono sample was 1.00 KB, and final model size was 12.87 KB (SD=2.75). The authentication time (including extracting features from two audio clips and classification time) was 1.2 (SD=0.2) seconds and 38.3 (SD=4.3) milliseconds for the Raspberry Pi and PC, respectively. Our evaluation demonstrates that training a lightweight SVM classifier is feasible even on a low-resource device like the Raspberry Pi. Modern smartglasses are substantially more powerful than this Pi—Google Glass Enterprise Edition 2, for example, features a Qualcomm Snapdragon XR1, 3GB RAM and 32GB storage [20], and Apple Vision Pro will be equipped an M2 chip (3.49GHz CPU and 24GB RAM) [6]. Thus, we believe training and running our classifiers directly on smartglasses is feasible; the performance we report on the Raspberry Pi is both already sufficient to achieve a good user experience in terms of supporting rapid authentication and, additionally, greatly slower than that which would be achieved on current smartglass platforms.

6 STUDY 3: MULTI-DAY PERFORMANCE AND USABILITY

To measure the multi-day recall performance and perceived usability of the enrollment and authentication process, we conducted a one-week study based on a more robust single-unit frame. We used the speech cue signal and optimized enrollment settings described in Section 5.9 to finalize the system configuration.

6.1 Single-Unit Frame Prototype

To support this study, we iterated on the smartglass prototype. Specifically, we used the same speaker and microphone modules and maintained the brow and mastoid locations, but developed a single frame, 3D printed in Nylon 12, to house these components. In addition, we miniaturized the audio PCBs and integrated a battery and charger for the speaker and integrated these into the frame.

Figure 11: Final prototype showing mastoid and brow speaker and microphone positions (left and right). Center image shows frame—the battery and charger for the speaker integrated into the right arm and the microprocessor and audio boards integrated into the left.

The resulting prototype is substantially more robust than prior iterations and can be connected to a PC by a single USB cable, greatly simplifying study execution. These changes make the prototype suitable for a longer, multi-day study. This final prototype can be seen in Figure 11.

6.2 User Study Methodology

We again recruited both genuine user and imposter participants. Genuine users first followed the enrollment procedures recommended in Section 5.9: five donning sessions each involving five repetitions of the speech cue. After enrollment, genuine users participated in an immediate recall session involving two cue repetitions. They then attended two additional multi-day recall sessions separated by a minimum of 24 hours, spread over up to one week. Each session involved two separate donnings: participants experienced a single cue playback in the initial donning; they then experienced two cue repetitions in the subsequent donning, the first playback taking place in normal conditions, and the second taking place in the presence of a vibration stimulus (generated by an off-the-shelf massage device attached to their chair). This was intended to assess the impact of vibratory noise, such as might occur while riding public transport. After system enrollment and the initial donning in the second and third recall sessions, participants completed subjective measures: the system usability scale (SUS) [8], the unweighted NASA Task Load Index (TLX) [27], and the the Borg-10 perceived exertion scale [66]. This was intended to capture their experiences after both enrollment and also exposure to just a single recall cue presentation. Imposter participants completed a single enrollment session. Genuine users were compensated with the equivalent of 23 USD in local currency and imposters with 8 USD.

6.3 Demographics

We recruited 27 new participants—17 genuine users and 10 imposters—aged between 18 and 28. Fourteen were female and thirteen were male (see Appendix A.2 Table 3 for full details). The genuine user group consisted of eight male and nine female participants. There was no statistically significant difference between the genuine user group and the imposter group with respect to their age, height, weight, and head circumference (all two-sample t-tests $p \geq 0.104$).

6.4 Authentication Accuracy

For each genuine user, we trained a classifier on all enrollment samples (25 in total) and five samples from each imposter (50 in

total). We assessed FRRs using the two samples from the first immediate recall, and all three samples from each of the two multi-day recalls (eight samples in total). To measures FARs, we created an imposter test set for each genuine user by selecting all samples from all other genuine users (528 samples in total). Mean EER measured using samples from all three recall tests (including both normal and vibration conditions) was 2.94%—a figure that is on par with those recorded in our second study. However, because finding an optimal EER threshold in a real-world deployment setting is, in practice, infeasible [30], we primarily report the third study results based on the HTER (or mean of the FAR and FRR) at a fixed 0.5 probability threshold. This was 3.10% (5.15% FRR and 1.04% FAR). Indeed, we recorded 7 authentication failures from 136 total attempts across all 17 users, a positive result that strongly suggests that SkullID deployed with the optimized enrollment settings we propose is capable of reliably authenticating users over the course of several days. In addition, to measure the effects of being exposed to vibration interference during recall, we compared the FRR between the normal and vibration conditions, and found no change between the two conditions: SkullID appears to be robust against vibration interference.

6.5 Usability Results

The mean Borg CR10 score during enrollment was 2.7 (SD=2.1)—indicating low to moderate exertion was experienced over the five donnings. The mean Borg CR10 score for the shorter authentication sessions dropped to 1.9 (SD=2.2). This indicates participants felt low exertion during recall, and suggests that their perceptions of the authentication experience may be improving as they gain experience with it. Raw data for TLX and SUS questionnaires are shown in Appendix A.5 Table 6. To summarize, mean SUS scores for enrollment and authentication were 76.2 (SD=12.7) and 81.1 (SD=11.5), respectively. These figures are associated with "good" levels of usability [8]. Drilling down into these aggregates, we note the scores for the "easy to use" and "quick to learn" questions were particularly high (both 4.35 from 5 for enrollment, and 4.79 and 4.85 for recall). This suggests SkullID was particularly effective in terms of these qualities. Similarly, the unweighted overall workload scores from the NASA TLX were 3.7 (SD=2.1) and 2.3 (SD=2.6) for enrollment and authentication sessions, respectively. To provide a general context for these results, we refer to Grier [24]'s analysis of 200 studies deploying TLX in a wide range of tasks; scores in the ranges reported here (2.3 to 3.7) were recorded in the 10% of tasks with the lowest workload levels encountered in Grier's survey. To more specifically contextualize these results in authentication tasks, we refer to overall unweighted TLX scores for entering PINs on smartwatches: approximately 6/20 [47]. The fact our enrollment processes yield lower workload scores than a familiar PIN authentication task is compelling evidence that participants did not experience undue workload while operating SkullID. That said, one aspect of the TLX data is worth pulling out: while still relatively low, ratings for effort expended show a modest spike in authentication (5.73) and, particularly, enrollment (7.59) tasks. We suggest this is due to the work involved in donning (and in the case of enrollment, repeatedly donning) our tethered and snugly fitting glasses frame. Improving this aspect of user experience may require further iteration on our prototypes (e.g., by creating a wireless version or developing models with different sizes to support improved fit) or further reducing the data collection requirements during enrollment. Regardless, taken together, these results indicate that the usability of SkullID is good: authenticating daily will require only modest effort and represent a small burden to users.

7 SECURITY ANALYSIS

To measure the robustness of SkullID against replay attacks described in Section 3.6, we conducted a separate attack study by recruiting ten new attackers, and using the third study genuine user samples as compromised (replay attack) response signals. Further, to explore the impact of imitation attacks, we applied a clustering technique to identify six separate groups with similar physiological characteristics, and performed an attack using all other users within the same cluster as attack samples.

7.1 Robustness to Replay Attacks

We recruited 10 new participants as attackers: 5 males and 5 females (see Table 3 in Appendix A.2 for full details). For each attacker, we provided a set of 68 genuine samples: one sample from each of the two microphone locations and two recall sessions (the second and third regular recall sessions) for all 17 genuine users. These signals were played on a standalone speaker unit (one of speaker modules extracted from the glass prototype) firmly pressed by the attacker participant against their skull in three different locations: adjacent to the mastoid speaker, the brow microphone and the mastoid microphone. Thus, each attacker recorded a pair of brow and mastoid microphone responses to 204 signals (68 samples by 3 speaker locations) while each genuine user (victim) was exposed to 120 replay attack responses (each composed of a pair of brow and mastoid recordings). The results led to 0.44% FAR across all attack attempts (at an unbiased 0.5 threshold) and a total of four compromised users: participants 1, 8, and 10 were each spoofed once whereas participant 5 was spoofed six times.

A closer investigation of the attack signals of the four spoofed victims revealed substantially weaker energy transfer rates on the microphone that was located further from the attacker's speaker unit. Figure 12 shows those attenuated attack signals, and the resultant PSD patterns. We note these are highly distinct from the original genuine and imposter samples. As our binary classifiers were trained using a small number of standard genuine and imposter samples (recorded normally through the frame), such novel PSD patterns (unknown to the classifiers) could be classified in a random manner, and result in authentication failures. To mitigate this sampling limitation, we experimented with a training method that entails adding attack samples to the train set. We first recruited two additional attackers and two new genuine/victim users following the same data collection protocol—this procedure ensured there is no overlap between the new imposter train samples added (those collected through the two new attackers) and the original attack samples used for testing. We then selected 10 random samples from each of the new attackers, included those samples in the final imposter train set, and measured attack success rates again. The new results were promising: no participants were spoofed (0% FAR). This improvement demonstrates the effectiveness of using

attack-specific training samples. It leads to more robust classifiers with an improved awareness of replayed signals.

7.2 Robustness to Imitation Attacks

A second practical attack scenario would involve recruiting attackers who are physiologically similar to a victim (e.g., with respect to head circumferences), and simply asking them to put on the stolen glasses and authenticate as normal. To study the robustness of SkullID to such imitation attempts, we used the genuine user data collected from the third study, and applied a clustering technique to group participants based on similarities in physiological characteristics. After excluding two participants (13 and 16) who opted not provide their physiological information, we applied the k-means clustering algorithm to the remaining 15 participants based on weight, height, and head circumference information. We found six separate clusters, one of which featured a single individual (participant 8) and was culled. For each participant in the remaining five clusters, we measured FARs (representing the attack success rates) using all other users within the same cluster. The attack results are shown in Table 7 and indicate the attack was largely unsuccessful: the majority of participants recorded zero FARs, while participant 7 recorded a FAR of 1.52%. These results suggest SkullID is robust against imitation efforts that involve recruiting subjects with similar physiological characteristics.

8 DISCUSSION

This paper investigated the feasibility of measuring through-skull bone conduction to authenticate smartglass users. With contact microphones mounted on the brow and mastoid, we achieve EERs of 5.68% and 7.95% in a lab study conducted on 25 participants. Concatenating the two response signals led to substantial improvement in the EERs (2.35%), a figure that improves considerably over closely related prior work (e.g., 5.48% [18], or see Table 2 for a full comparison). In addition, through a multi-recall session study (N=30), we demonstrated that the proposed two-channel system is robust to microphone contact variability and background audio noise (2.72% EER), issues that have seen scant attention in prior work on bio-acoustic authentication on smartglasses [31, 32]. We also demonstrate that a prior solution [55] that uses air-transmitted audio signals reported significant elevations in the error rates in the presence of background noise (65.61% EERs). A more extensive study conducted over a week (N=27) reported consistent multi-day recall performance (2.94% EER), and robustness to vibration interference. Our system is also feasible for real world deployment: an SVM-based implementation took only 30 seconds to train a full classifier on a low-power Raspberry Pi. In addition, we designed and validated a lightweight SVM solution that requires just 25 enrollment training samples from a user. Furthermore, participants experienced low exertion and workload levels during enrollment and recall, and reported high SUS scores. Finally, our security analysis demonstrated low success rates for various forms of imitation and signal replay attacks. We highlight key takeaways, future directions and limitations of our work in the sections below.

8.1 In-air versus Bone Conduction Microphones

Our second study revealed that in-air microphones suffer from two major performance limitations (Section 5.8). First, they showed a substantial drop in accuracy drop during longitudinal recall. Second, they were highly susceptible to background noise. Taken together, these issues make in-air microphones impractical for real world authentication. Contact microphones, in comparison, were more reliable over time and largely immune to background noise: this makes them more suitable for supporting bio-acoustic authentication. Integration of such devices into next generation smartglass products is also likely—reflecting the fact that skull conducted sound, in general, includes user speech but excludes externally originating sounds, several bone conduction sensors (a.k.a. voice accelerometers or voice pickup bone sensors) have recently been introduced in the market (e.g., Sonion VPU [59] and Vesper VA1200 [63]) and integrated into the current crop of high-end earbuds (e.g., Apple AirPods Pro [5], Samsung Galaxy Buds Pro [54], and Huawei FreeBuds Pro [29]). While the general use of such sensors is to work in tandem with in-air microphones to improve voice recognition, other applications have also been pursued. In particular, the Huawei FreeBuds Pro uses their integrated bone conduction sensor to authenticate users' speech [28]. As voice assistants are a major use-case for smartglasses, we expect such sensors will also appear in next generation smartglass products. Given these trends, a good direction for future work is exploring how to improve bio-acoustic authentication by combining signals from in-air and bone conduction microphones.

8.2 Microphone Location Recommendations

While our results indicate bone conducted audio supports reliable authentication, they also suggest that not all sensing locations are created equal. In the single-session study (see Table 4) the brow and mastoid were the best two performing microphone locations. This is, at least in part, due to the relatively large size of the underlying bone structures and the lack of obstructions (e.g., hair) on their surfaces. However, our analysis also cautions against relying on any single microphone location. In our studies, variability in microphone fit clearly impacted performance. Having multiple microphones helped to mitigate this problem, as variability in fit was relatively independent—when one microphone exhibited unreliable or uncharacteristic data due to misalignment, another could still be well-seated and return typical signals. We expect to see similar effects in real smartglass systems, as users are unlikely to don their glasses in exactly the same way each time (e.g., due to variations in hair style or use of accessories such as hair bands, ear jewelry or face masks). Accordingly, we recommend that smartglass systems that seek to incorporate bio-acoustic authentication target use of at least two well separated sensors. We specifically recommend the mastoid and brow as candidate locations as this pair achieved good performance (2.35–2.84% EERs) in our single session study. Our subsequent multi-session studies provide a strong validation of this multi-channel recommendation: demonstrating 2.72% and 2.94% EERs, respectively. We conclude that multiple sensor points can help reduce the impact of device placement variability on authentication system performance, and strongly recommend that any real world smartglass implementations pursue this strategy.

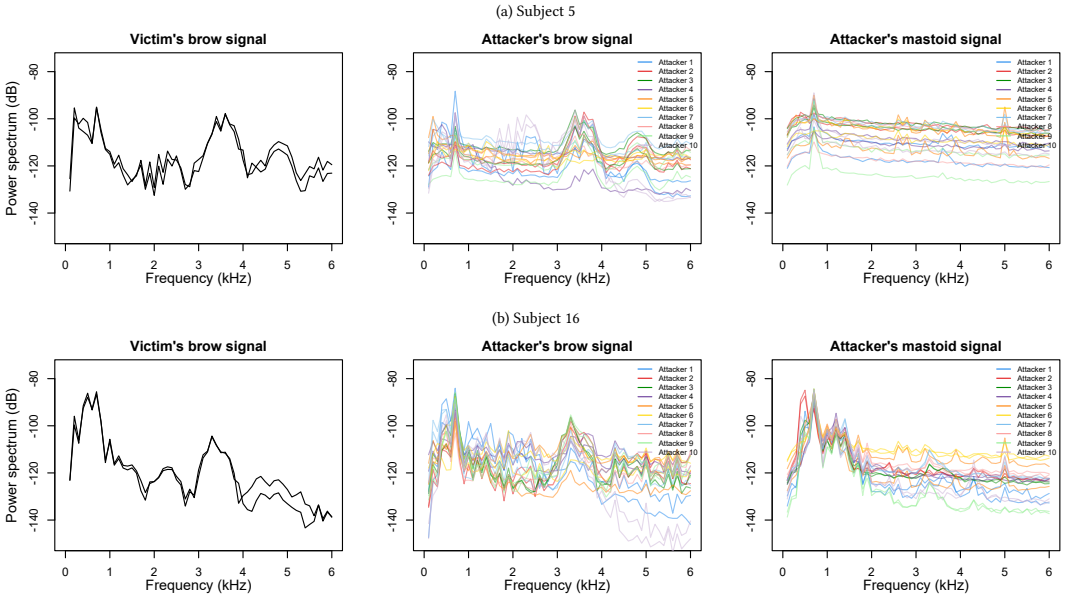

Figure 12: The two graphs in the first column represent power spectrum of the two samples collected on the brow microphone of subjects 5 and 16. The graphs in the second and third columns represent the power spectrum of the replayed signals collected on the attackers' brow and mastoid process, respectively (the original victim signals were injected through the surface transducer mounted on the attackers' brow). Subject 16 was never spoofed whereas subject 5 was spoofed six times.

8.3 Limitations

There are a number of limitations to our work. While our sample sizes (N=25 for study 1, N=30 for study 2, and N=27 for study 3) exceed those in much prior work in this area [32, 55], more substantial studies would increase confidence in the results we report. In addition to recruiting larger participant groups, it would also be beneficial to engage them for prolonged periods. While we are the first group to study bio-acoustic authentication for smartglasses is a sustained multi-recall session, longer term studies, spanning months, would also us to examine gradual changes due to factors like hair growth [48] or fluctuations in weight and body composition. Such extended studies are a clear next step for this work. In addition, we note that our participant sample is Asian (covering multiple countries). The performance of our system on other ethnic groups thus remains unknown and increasing the diversity of our participant pool is an imperative for future studies.

Beyond these relatively generic concerns, there are also more specific issues. While we showed the robustness of SkullID to surrounding in-air audio noise (e.g., music, news broadcast, and public cafe noise) and vibration noise, we did not explore other physical disturbances or more realistic use-contexts. To enable us to do so, future work should develop more portable and readily deployable versions of our system. This would first require engineering efforts to miniaturize components by, for example, integrating recently developed VPUs [59, 63] in place of contact microphones, or by using existing bone conductance headphone platforms (e.g., [56]). In addition, it would be valuable to integrate SkullID with existing HMDs, such as Meta's Quest devices [45] to verify its feasibility on consumer products. While our work has demonstrated the effectiveness of SkullID when mounted on the head with fabric bands, a common approach in consumer HMDs, actually integrating it with such products to formally test its performance would be a key next step for this work. With such improvements in place, it would then be useful to evaluate SkullID's performance in contexts such as walking or riding both private vehicles (representing relatively stable travel conditions) and also in more noisy settings such as on public transit (e.g., buses and subways). Such studies will provide important data to enhance our understanding of the susceptibility of bone-conduction authentication systems to noise. Finally, our replay attack analysis revealed that our binary classifiers can fail when presented with unknown signals, such as those replayed through a separate speaker unit. Nevertheless, we were able to effectively mitigate replay attacks by adding a few attack samples in the final train set, and improving the classifiers' awareness of the replayed signal characteristics. Future research should investigate how such adaptive training methods could be applied to mitigate other forms of signal replay or manipulation attacks, and optimize overall system accuracy.

9 CONCLUSION

We presented SkullID, a system that authenticates users by the unique characteristics of audio transmitted through their skulls. We demonstrate it is deployable, effective, immune to noise, reliable over time and resistant to attack. We highlight how signal selection and sensor placement impact performance. Taken together the results in this paper represent a robust demonstration of the feasibility of using bone conducted audio to authenticate smartglass users and provide precise, relevant and actionable recommendations for how such systems should be designed in the future.

ACKNOWLEDGMENTS

This work was supported by Samsung Research and the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (2023R1A2C1004046).

REFERENCES

[1] Adafruit 2014. *Small Surface Bone Conductor Transducer Exciter with Wires 8 Ohm 1 Watt.* Retrieved July 28, 2023 from https://www.adafruit.com/product/1674

[2] Amazon 2020. *Echo Frames (2nd Gen).* Retrieved July 28, 2023 from https://www.amazon.com/All-new-Echo-Frames/dp/B083C58VDP

[3] Bradley W. Anderson, Michael W. Kortz, Asa C. Black, and Khalid A. Al Kharazi. 2023. *Anatomy, Head and Neck, Skull.* StatPearls Publishing.

[4] Marios Antonakakis, Sophie Schrader, Ümit Aydin, Asad Khan, Joachim Gross, Michalis Zervakis, Stefan Rampp, and Carsten H Wolters. 2020. Inter-subject variability of skull conductivity and thickness in calibrated realistic head models. *Neuroimage* 223 (2020).

[5] Apple 2019. *AirPods Pro–Technical Specifications.* Retrieved July 28, 2023 from https://support.apple.com/kb/SP811?viewlocale=en_US&locale=en_US

[6] Apple 2023. *Apple Vision Pro.* Retrieved July 28, 2023 from https://www.apple.com/apple-vision-pro/

[7] Avid 2022. *Avid Pro Tools.* Retrieved July 28, 2023 from https://www.avid.com/pro-tools

[8] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies (JUS)* 4, 3 (2009), 114–123.

[9] Fadi Boutros, Naser Damer, Kiran Raja, Raghavendra Ramachandra, Florian Kirchbuchner, and Arjan Kuijper. 2020. Iris and periocular biometrics for head mounted displays: Segmentation, recognition, and synthetic data generation. *Image and Vision Computing* 104 (2020), 104007.

[10] MW Boyce, RH Thomson, JK Cartwright, DT Feltner, CR Stainrod, J Flynn, C Ackermann, J Emezie, CR Amburn, and E Rovira. 2022. Enhancing Military Training Using Extended Reality: A Study of Military Tactics Comprehension. *Front. Virtual Real.* 3 (2022), 9 pages. https://doi.org/10.3389/frvir.2022.754627

[11] Peter J Brockwell and Richard A Davis. 1991. *Time Series: Theory and Methods.* Springer-Verlag.

[12] Zhi Cai, Douglas G Richards, Martin L Lenhardt, and Alan G Madsen. 2002. Response of human skull to bone-conducted sound in the audiometric-ultrasonic range. *The International Tinnitus Journal* 8, 1 (2002), 3–8.

[13] Geumhwan Cho, Jun Ho Huh, Soolin Kim, Junsung Cho, Heesung Park, Yenah Lee, Konstantin Beznosov, and Hyoungshick Kim. 2020. On the Security and Usability Implications of Providing Multiple Authentication Choices on Smartphones: The More, the Better? *ACM Transactions on Privacy and Security* 23, 4 (2020).

[14] Ivo Dobrev, Jae Hoon Sim, Stefan Stenfelt, Sebastian Ihrle, Rahel Gerig, Flurin Pfiffner, Albrecht Eiber, Alexander M Huber, and Christof Röösli. 2017. Sound wave propagation on the human skull surface with bone conduction stimulation. *Hearing Research* 355 (2017), 1–13.

[15] 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 (MobiCom '17).* Association for Computing Machinery, New York, NY, USA, 343–355. https://doi.org/10.1145/3117811.3117823

[16] Focusrite 2019. *Focusrite Scarlett 18i8.* Retrieved July 28, 2023 from https://focusrite.com/en/usb-audio-interface/scarlett/scarlett-18i8

[17] 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 (IMWUT)* 5, 1 (2021), 1–25.

[18] Yang Gao, Wei Wang, Vir V Phoha, Wei Sun, and Zhanpeng Jin. 2019. EarEcho: Using ear canal echo for wearable authentication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 3, 3 (2019), 1–24.

[19] Google 2013. *Google Glass.* Retrieved July 28, 2023 from https://www.google.com/glass/start/

[20] Google 2019. *Glass Enterprise Edition 2 Tech Specs.* Retrieved July 28, 2023 from https://support.google.com/glass-enterprise/customer/answer/9220200?hl=en

[21] GoPro 2019. *GoPro Head Strap.* Retrieved July 28, 2023 from https://www.amazon.com/GoPro-Strap-QuickClip-Official-Mount/dp/B00F19PYR4?th=1

[22] Claire C Gordon, Cynthia L Blackwell, Bruce Bradtmiller, Joseph L Parham, Patricia Barrientos, Stephen P Paquette, Brian D Corner, Jeremy M Carson, Joseph C Venezia, Belva M Rockwell, et al. 2014. *2012 anthropometric survey of us army personnel: Methods and summary statistics.* Technical Report. Army Natick Soldier Research Development and Engineering Center MA.

[23] Michael S Gordon, Michael D Hall, Jeremy Gaston, Ashley Foots, and Jitwipar Suwangbutra. 2019. Individual differences in the acoustic properties of human skulls. *The Journal of the Acoustical Society of America* 146, 3 (2019), 191–197.

[24] Rebecca A. Grier. 2015. How High is High? A Meta-Analysis of NASA-TLX Global Workload Scores. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 59, 1 (2015), 1727–1731. https://doi.org/10.1177/1541931215591373 arXiv:https://doi.org/10.1177/1541931215591373

[25] Matti D. Groll, Jennifer M. Vojtech, Surbhi Hablani, Daryush D. Mehta, Daniel P. Buckley, J. Pieter Noordzij, and Cara E. Stepp. 2020. Automated Relative Fundamental Frequency Algorithms for Use With Neck-Surface Accelerometer Signals. *Journal of Voice* 36, 2 (2020), 156–169.

[26] Bo Håkansson, Anders Brandt, Peder Carlsson, and Anders Tjellström. 1994. Resonance frequencies of the human skull in vivo. *The Journal of the Acoustical Society of America* 95, 3 (1994), 1474–1481.

[27] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, Oxford, England, 139–183.

[28] Huawei 2020. *Bone Voiceprint Recognition update rolling out for Huawei FreeBuds 3.* Retrieved July 28, 2023 from https://consumer.huawei.com/en/community/details/Bone-Voiceprint-Recognition-update-rolling-out-for-Huawei-FreeBuds-3/topicId_88745/

[29] Huawei 2020. *Huawei FreeBuds Pro Specifications.* Retrieved July 28, 2023 from https://consumer.huawei.com/sa-en/headphones/freebuds-pro/specs/

[30] Jun Ho Huh, Hyejin Shin, HongMin Kim, Eunyong Cheon, Youngeun Song, Choong-Hoon Lee, and Ian Oakley. 2023. WristAcoustic: Through-Wrist Acoustic Response Based Authentication for Smartwatches. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 6, 4 (2023), 1–34.

[31] Irwansyah, Sho Otsuka, and Seiji Nakagawa. 2021. Evaluation of Bone-Conducted Cross-Talk Sound in the Head for Biometric Identification. In *The 2021 International Conference on Computer, Control, Informatics and Its Applications (IC3INA).* 76–80.

[32] Kaito Isobe and Kazuya Murao. 2021. Person-identification Method using Active Acoustic Sensing Applied to Nose. In *Proceedings of 2021 International Symposium on Wearable Computers (ISWC '21).* Association for Computing Machinery, New York, NY, USA, 138–140. https://doi.org/10.1145/3460421.3480425

[33] Shiqi Jiang, Zhenjiang Li, Pengfei Zhou, and Mo Li. 2019. Memento: An Emotion-Driven Lifelogging System with Wearables. *ACM Trans. Sen. Netw.* 15, 1, Article 8 (jan 2019), 23 pages. https://doi.org/10.1145/3281630

[34] Moritz Kassner, William Patera, and Andreas Bulling. 2014. Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-Based Interaction. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '14) Adjunct Publication.* Association for Computing Machinery, New York, NY, USA, 1151–1160. https://doi.org/10.1145/2638728.2641695

[35] Jee-Hyun Kwon, Jong S Kim, Dong-Wha Kang, Kyun-Seop Bae, and Sun U Kwon. 2006. The thickness and texture of temporal bone in brain CT predict acoustic window failure of transcranial Doppler. *Journal of Neuroimaging* 16, 4 (2006), 347–352.

[36] Lavalier 2019. *Lavalier GO Professional Microphone.* Retrieved July 28, 2023 from https://rode.com/microphones/lavalier-wearable/lavalier-go

[37] Jin-hee Lee, Su-Jeong Hwang Shin, and Cynthia L Istook. 2006. Analysis of human head shapes in the united states. *International Journal of Human Ecology* 7, 1 (2006), 77–83.

[38] Haiyan Li, Jesse Ruan, Zhonghua Xie, Hao Wang, and Wengling Liu. 2007. Investigation of the critical geometric characteristics of living human skulls utilising medical image analysis techniques. *International Journal of Vehicle Safety* 2, 4 (2007), 345–367.

[39] 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 (SenSys).* 312–325.

[40] Sugang Li, Ashwin Ashok, Yanyong Zhang, Chenren Xu, Janne Lindqvist, and Macro Gruteser. 2016. Whose move is it anyway? Authenticating smart wearable devices using unique head movement patterns. In *2016 IEEE International Conference on Pervasive Computing and Communications (PerCom).* IEEE Computer Society, USA. https://doi.org/10.1109/PERCOM.2016.7456514

[41] 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 (ISWC '23').* 6–10.

[42] Rui Liu, Cory Cornelius, Reza Rawassizadeh, Ronald Peterson, and David Kotz. 2018. Vocal resonance: Using internal body voice for wearable authentication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 2, 1 (2018), 1–23.

[43] Ivan Martinovic, Marc Roeschlin, and Ivo Sluganovic. 2017. *Mobile Biometrics in Financial Services: A Five Factor Framework.* Technical Report. University of Oxford, Oxford, UK.

[44] Maranda McBride, Tomasz R Letowski, and Phuong K Tran. 2005. *Bone conduction head sensitivity mapping: Bone vibrator.* Technical Report. ARMY RESEARCH LAB ABERDEEN PROVING GROUND MD.
[45] Meta 2023. *Meta Quest VR Headsets.* Retrieved December 10, 2023 from https://www.meta.com/us/en/quest/
[46] Microsoft 2021. *U.S. Army to use HoloLens technology in high-tech headsets for soldiers.* Retrieved July 28, 2023 from https://news.microsoft.com/transform/u-s-army-to-use-hololens-technology-in-high-tech-headsets-for-soldiers/
[47] Ian Oakley, Jun Ho Huh, Junsung Cho, Geumhwan Cho, Rasel Islam, and Hyoungshick Kim. 2018. The Personal Identification Chord: A Four ButtonAuthentication System for Smartwatches. In *Proceedings of the 2018 ACM on Asia Conference on Computer and Communications Security (ASIACCS '18).* Association for Computing Machinery, New York, NY, USA, 75–87. https://doi.org/10.1145/3196494.3196555
[48] Satoki Ogiso, Koichi Mizutani, Keiichi Zempo, Naoto Wakatsuki, and Yuka Maeda. 2017. Estimation of contact force and amount of hair between skin and bone-conducted sound transducer using electrical impedance. *Japanese Journal of Applied Physics* 56 (2017).
[49] Alexi Orchard, Marcel O'Gorman, Chelsea La Vecchia, and Jason Lajoie. 2022. Augmented Reality Smart Glasses in Focus: A User Group Report. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22).* Association for Computing Machinery, New York, NY, USA, Article 20, 7 pages. https://doi.org/10.1145/3491101.3503565
[50] Ge Peng, David T Nguyen, Gang Zhou, and Shuangquan Wang. 2015. A continuous and noninvasive user authentication system for google glass. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '15)').* 487–487.
[51] Pauline Pfeifer, Tim Hilken, Jonas Heller, Saifeddin Alimamy, and Roberta Di Palma. 2023. More than meets the eye: In-store retail experiences with augmented reality smart glasses. *Computers in Human Behavior* 146 (2023), 107816. https://doi.org/10.1016/j.chb.2023.107816
[52] Raspberry Pi 2016. *Raspberry Pi 3 Model B.* Retrieved July 28, 2023 from https://www.raspberrypi.com/products/raspberry-pi-3-model-b/
[53] Giovanni Saggio, Angela Scioscia Santoro, Vito Errico, Maurizio Caon, Alfiero Leoni, Giuseppe Ferri, and Vincenzo Stornelli. 2021. A Novel Actuating–Sensing Bone Conduction-Based System for Active Hand Pose Sensing and Material Densities Evaluation Through Hand Touch. *IEEE Transactions on Instrumentation and Measurement* 70 (2021), 1–7.
[54] Samsung 2021. *Galaxy Buds Pro.* Retrieved July 28, 2023 from https://www.samsung.com/global/galaxy/galaxy-buds-pro/
[55] 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 (CHI '16).* Association for Computing Machinery, New York, NY, USA, 1379–1384. https://doi.org/10.1145/2858036.2858152
[56] Shokz 2023. *Shokz Bone Conduction Headphone.* Retrieved December 10, 2023 from https://shokz.com/
[57] Shutterhealth 2020. *Remote Scribes for Clinicians – Google Glass and Sutter Health.* Retrieved July 28, 2023 from https://vitals.sutterhealth.org/remote-scribes-for-clinicians-google-glass-and-sutter-health/
[58] Sabrina Sobieraj, Sabrina Eimler, and Gerhard Rinkenauer. 2023. Can smart glasses change how people evaluate healthcare professionals? A mixed-method approach to using smart glasses in hospitals. *International Journal of Human-Computer Studies* 178 (2023), 103081. https://doi.org/10.1016/j.ijhcs.2023.103081
[59] Sonion 2018. *Sonion Voice Pick Up Bone Sensor (VPU).* Retrieved July 28, 2023 from https://www.sonion.com/hearing/bone-conduction-sensors-and-actuators/vpu-voice-pick-up-sensor/
[60] Sophie Stephenson, Bijeeta Pal, Stephen Fan, Earlence Fernandes, Yuhang Zhao, and Rahul Chatterjee. 2022. SoK: Authentication in Augmented and Virtual Reality. In *Proceedings of IEEE Symposium on Security and Privacy (S&P '22).* IEEE Computer Society, USA, 267–284. https://doi.org/10.1109/SP46214.2022.9833742
[61] Hemant Bhaskar Surale, Yu Jiang Tham, Brian A. Smith, and Rajan Vaish. 2022. ARcall: Real-Time AR Communication Using Smartphones and Smartglasses. In *Proceedings of the Augmented Humans International Conference 2022* (Kashiwa, Chiba, Japan) *(AHs '22).* Association for Computing Machinery, New York, NY, USA, 46–57. https://doi.org/10.1145/3519391.3519398
[62] Jitwipar Suwangbutra, Rachelle Tobias, and Michael S Gordon. 2013. Music of the body: An investigation of skull resonance and its influence on musical preferences. In *Proceedings of Meetings on Acoustics.* AIP Publishing. https://doi.org/10.1121/1.4805756
[63] Vesper 2021. *Vesper VA1200 Bone Conductor Sensor – Analog Piezoelectric MEMS Voice Accelerometer.* Retrieved July 28, 2023 from https://vespermems.com/products/va1200/
[64] Vue 2016. *Vue: Your Everyday Smart Glasses.* Retrieved July 28, 2023 from https://www.kickstarter.com/projects/vue/vue-your-everyday-smart-glasses
[65] Zi Wang, Sheng Tan, Linghan Zhang, Yili Ren, Zhi Wang, and Jie Yang. 2021. EarDynamic: An Ear Canal Deformation Based Continuous User Authentication Using In-Ear Wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 5, 1 (2021), 1–27.
[66] Nerys Williams. 2017. The Borg Rating of Perceived Exertion (RPE) scale. *Occupational Medicine* 67, 5 (2017), 404–405.
[67] Dhruv Kumar Yadav, Beatrice Ionascu, Sai Vamsi Krishna Ongole, Aditi Roy, and Nasir Memon. 2015. Design and analysis of shoulder surfing resistant pin based authentication mechanisms on google glass. In *Proceedings of International Conference on Financial Cryptography and Data Security (FC '15).* Springer, Berlin, Heidelberg, 281–297. https://doi.org/10.1007/978-3-662-48051-9_21
[68] Shanhe Yi, Zhengrui Qin, Ed Novak, Yafeng Yin, and Qun Li. 2016. Glassgesture: Exploring head gesture interface of smart glasses. In *Proceedings of the 35th Annual IEEE International Conference on Computer Communications (INFOCOM '16).* IEEE Computer Society, USA, 1–9. https://doi.org/10.1109/INFOCOM.2016.7524542
[69] Cheng Zhang, Anandghan Waghmare, Pranav Kundra, Yiming Pu, Scott Gilliland, Thomas Ploetz, Thad E Starner, Omer T Inan, and Gregory D Abowd. 2017. FingerSound: Recognizing unistroke thumb gestures using a ring. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 1, 3 (2017), 1–19.
[70] Cheng Zhang, Qiuyue Xue, Anandghan Waghmare, Ruichen Meng, Sumeet Jain, Yizeng Han, Xinyu Li, Kenneth Cunefare, Thomas Ploetz, Thad Starner, Omer Inan, and Gregory D. Abowd. 2018. FingerPing: Recognizing Fine-Grained Hand Poses Using Active Acoustic On-Body Sensing. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18).* Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3173574.3174011
[71] Ziqing Zhuang, Chang Shu, Pengcheng Xi, Michael Bergman, and Michael Joseph. 2013. Head-and-face shape variations of US civilian workers. *Applied Ergonomics* 44, 5 (2013), 775–784.

A APPENDIX

A.1 Comparing SkullID to Prior Work on Smartglass Authentication

Table 2 summarizes the key differences between the proposed system and prior work on bio-acoustic authentication with respect to the device form factors, sensors used, required authentication actions, studied (experimental) conditions, and reported accuracy.

A.2 User Study Demographics

Table 3 shows the demographics of the 25 participants recruited in the first single-session study with a headband prototype, 30 participants in the second single-day multi-session study with a glasses frame prototype, 27 participants in the third multi-day study with a single frame prototype, and ten replay attackers in study 3.

A.3 First Study Results

A.3.1 User Identification Results. As the first step towards investigating the feasibility of using through-skull bone conduction information to identify individuals, we trained multi-class SVM classifiers using the samples collected from the first four donnings, and evaluated their performance using samples from the last donning. We used the MFCC/PSD concatenation feature set for this analysis. The user identification accuracy ranged from between 71–89% (across all four locations and three cue signals) when a single microphone was used. We observed significant elevations in the accuracy when two or more microphones were used together, e.g., demonstrating 93–98% accuracy after concatenating the brow and mastoid features. Figure 13 summarizes these results.

A.3.2 Performance by Microphone Location. To see whether there are statistical differences in authentication accuracy between the four microphone locations, we performed pairwise paired t-tests between locations. We present all location pairs whose authentication accuracy is statistically significantly different (Bonferroni corrected $p < 0.05$) in Table 4.

Table 2: A comparison of bio-acoustic authentication methods. Note that ✓denotes "studied" condition, and × denotes conditions "not considered" in the user study.

	Biometric	Form factor	Required sensors	Required actions	Accuracy		Experimental conditions				
					Single-session	Multi-session	Multiple sensor locations	Sensor location variability	Multiple cue sounds	Ambient noise	Security analysis
SkullID	Skull structure	Smartglasses	Bone conduction transducer and two bone conduction microphones	None	2.35% EER	2.94% EER	✓	✓	✓	✓	✓
Schneegass et al. [55] (SkullConduct)	Skull structure	Smartglasses	Bone conduction transducer and in-air microphone	None	6.94% EER	×	×	×	×	×	×
Isobe and Murao [32]	Nose structure	Smartglasses	Piezoelectric speaker and microphone	None	9% EER	×	×	×	×	×	×
Irwansyah et al. [31]	Skull structure	Headset	Bone conduction transducer and inward-facing microphone	None	2.6% EER	×	×	×	×	×	×
Feng et al. [15]	Voice and vocal resonance	Smartglasses, earphones, and necklace	Accelerometer and in-air microphone	Voice commands	3% FRR, 0.1% FAR	×	✓	×	N/A	×	✓
Liu et al. [42]	Vocal resonance	Smartglasses, earphones, and necklace	Piezo contact microphone	Speaking	×	3.9% HTER	✓	×	N/A	×	✓
Gao et al. [18]	Ear canal structure	Earphones	Speaker and inward-facing microphone	None	5.48% HTER	×	×	✓	✓	✓	✓
Wang et al. [65]	Ear canal deformation	Earphones	Speaker and inward-facing microphone	Voice commands	97.38% recall and 95.02% precision	95% accuracy	×	×	N/A	✓	✓
Gao et al. [17]	Throat-to-ear body structure and body asymmetry	Earphones	Two outward-facing and inward-facing microphones	Speaking	3.64% EER	×	×	×	N/A	×	✓

Table 3: Demographics of study participants.

Variable	Value	Study 1 (N=25)	Study 2 (N=30)		Study 3 (N=27)		Study 4 (N=10)
		Single-session	Single-day multi-recall session		Multi-day recall session		Replay attack
			Genuine users	Imposters	Genuine users	Imposters	Attackers
Gender	Female	4 (16%)	4 (27%)	3 (20%)	9 (53%)	5 (50%)	5 (50%)
	Male	21 (84%)	11 (73%)	12 (80%)	8 (47%)	5 (50%)	5 (50%)
Age	Not respond	0 (0%)	1 (7%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	18-20	1 (4%)	2 (13%)	5 (33%)	5 (29%)	2 (20%)	5 (50%)
	21-30	23 (92%)	12 (80%)	9 (60%)	12 (71%)	8 (80%)	5 (50%)
	31-40	1 (4%)	0 (0%)	1 (7%)	0 (0%)	0 (0%)	0 (0%)
Weight (kg)	Not respond	0 (0%)	1 (7%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	[45, 50)	0 (0%)	0 (0%)	0 (0%)	3 (18%)	0 (0%)	1 (10%)
	[50, 55)	0 (0%)	2 (13%)	1 (7%)	4 (23%)	3 (30%)	2 (20%)
	[55, 60)	3 (12%)	1 (7%)	2 (13%)	3 (18%)	1 (10%)	3 (30%)
	[60, 65)	4 (16%)	2 (13%)	2 (13%)	4 (23%)	1 (10%)	1 (10%)
	[65, 70)	5 (20%)	4 (26%)	4 (26%)	2 (12%)	0 (0%)	1 (10%)
	[70, 75)	7 (28%)	3 (20%)	3 (20%)	0 (0%)	4 (40%)	1 (10%)
	[75, 80)	4 (16%)	1 (7%)	1 (7%)	1 (6%)	0 (0%)	0 (0%)
	[80, 90)	2 (8%)	1 (7%)	2 (13%)	0 (0%)	1 (10%)	1 (10%)
Height (cm)	Not respond	0 (0%)	1 (7%)	0 (0%)	2 (12%)	0 (0%)	0 (0%)
	[150, 160)	2 (8%)	0 (0%)	0 (0%)	3 (18%)	2 (20%)	2 (20%)
	[160, 170)	6 (24%)	3 (20%)	4 (27%)	6 (35%)	3 (30%)	4 (40%)
	[170, 180)	15 (60%)	9 (60%)	8 (53%)	5 (29%)	5 (50%)	3 (30%)
	[180, 190)	2 (8%)	2 (13%)	3 (20%)	1 (6%)	0 (0%)	1 (10%)
Head circumference (cm)	[54, 55)	0 (0%)	0 (0%)	1 (7%)	1 (6%)	0 (0%)	1 (10%)
	[55, 56)	0 (0%)	2 (13%)	0 (0%)	2 (12%)	0 (0%)	0 (0%)
	[56, 57)	3 (12%)	3 (20%)	2 (13%)	2 (12%)	2 (20%)	3 (30%)
	[57, 58)	4 (16%)	1 (7%)	2 (13%)	4 (23%)	1 (10%)	1 (10%)
	[58, 59)	11 (44%)	5 (33%)	3 (20%)	3 (18%)	0 (0%)	1 (10%)
	[59, 60)	4 (16%)	3 (20%)	6 (40%)	3 (18%)	4 (40%)	1 (10%)
	[60, 61)	2 (8%)	1 (7%)	1 (7%)	2 (12%)	2 (20%)	1 (10%)
	[61, 62)	1 (4%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	2 (20%)

A.3.3 Performance by Cue Type. To explore whether there are differences in authentication accuracy between the three cue types (chirp, wake-up melody, and speech), we performed pairwise paired t-tests between cue signals at different location combinations. Table 5 shows the cue signal pairs and the location combinations whose authentication accuracy is statistically significantly different (Bonferroni corrected $p < 0.05$).

A.3.4 Signal Consistency. Figure 14 shows the Wilks' Lambda statistics for each microphone with the chirp cue. Small Wilks' Lambda implies small within-subject variation, i.e., high signal pattern consistency. This figures indicates the brow and mastoid locations show high consistency, with a notable drop in performance on the temple and, in particular, on the squamous.

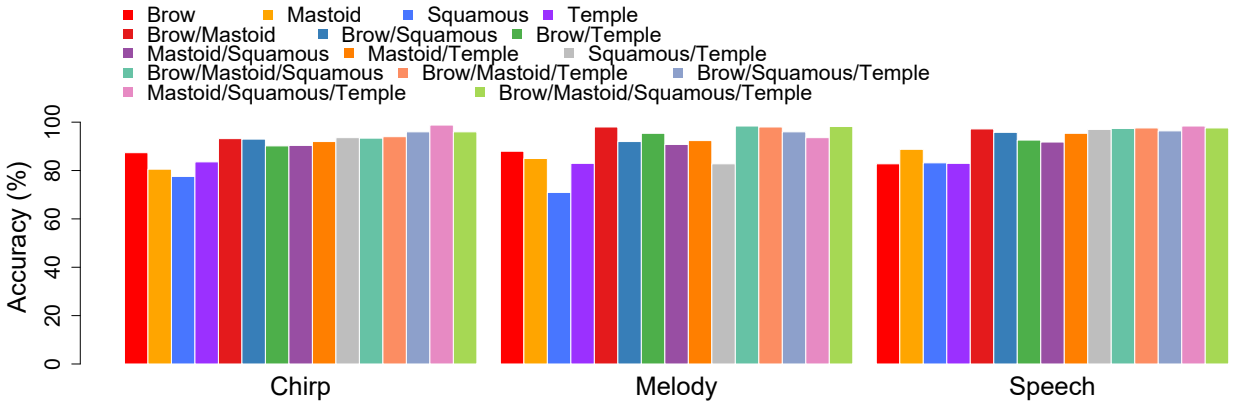

Figure 13: Mean user identification accuracy measured across all 25 participants, separately presented for each cue signal.

Table 4: Statistical significance of pairwise difference in authentication accuracy between microphone locations. Included location pairs were all those where mean EERs were statistically significantly different. Microphone locations indicated by initial letter: B(row), M(astoid), S(quamous),

Cue	Location Pair	Mean EER Difference (%)	Cohen's d	Bonferroni Corrected p
Chirp	B vs S	-2.22	-4.89	0.0001
	B vs T	-3.44	-7.86	0.0000
	M vs S	-1.67	-3.66	0.0073
	M vs T	-2.89	-5.41	0.0000
Melody	B vs M	-2.27	-5.68	0.0000
	B vs S	-4.25	-8.79	0.0000
	B vs T	-4.01	-9.29	0.0000
	M vs S	-1.99	-4.53	0.0003
	M vs T	-1.74	-4.82	0.0001
Speech	B vs S	-2.02	-4.96	0.0001
	B vs T	-2.07	-4.63	0.0002

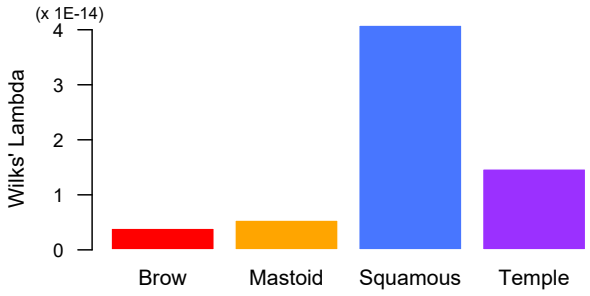

Figure 14: Wilks' Lambda statistics for the four microphone locations when a chirp from 50 Hz to 6 kHz was used.

A.4 In-air Microphone Performance

Figure 15 shows the recall performance for the in-air microphones: all three graphs demonstrate significant elevations in the EERs compared to those reported for SkullID both in the normal sessions (9.54–14.20%) and sessions recorded with audio noise (58–66%).

Table 5: Statistical significance of pairwise difference in authentication accuracy between cue signals. Selected cue pairs were those where average EERs were statistically significantly different. Microphone locations indicated by initial letter: B(row), M(astoid), S(quamous), T(emple).

Cue pair	Location(s)	Mean EER Difference (%)	Cohen's d	Bonferroni Corrected p
Chirp vs Melody	B	1.35	4.31	0.0017
	B/T	2.27	8.83	0.0000
	M/S	-0.90	-4.29	0.0018
	S/T	-3.11	-13.95	0.0000
	B/M/S	-0.64	-5.12	0.0001
	M/S/T	-2.36	-9.88	0.0000
	B/M/S/T	-0.61	-9.17	0.0000
Chirp vs Speech	M	-0.62	-3.47	0.0345
	T	1.53	4.26	0.0021
	M/T	1.59	6.23	0.0000
	S/T	-0.98	-4.84	0.0002
	M/S/T	0.48	3.97	0.0062
	B/M/S/T	0.18	3.39	0.0450
Melody vs Speech	B	-1.19	-4.26	0.0021
	B/S	0.77	8.45	0.0000
	B/T	-2.12	-8.38	0.0000
	M/S	1.09	5.44	0.0000
	M/T	1.80	10.42	0.0000
	S/T	2.13	13.09	0.0000
	B/M/S	0.68	7.76	0.0000
	B/M/T	0.42	5.21	0.0000
	M/S/T	2.84	10.05	0.0000
	B/M/S/T	0.80	11.71	0.0000

A.5 SUS and NASA TLX Results

Mean scores for all individual questions in the SUS and NASA TLX questionnaires are summarized in Table 6.

A.6 Imitation Attack Results

Table 7 shows the imitation attack results. Participants were grouped based on weight, height, and head size using the k-mean clustering algorithm ($k = 6$).

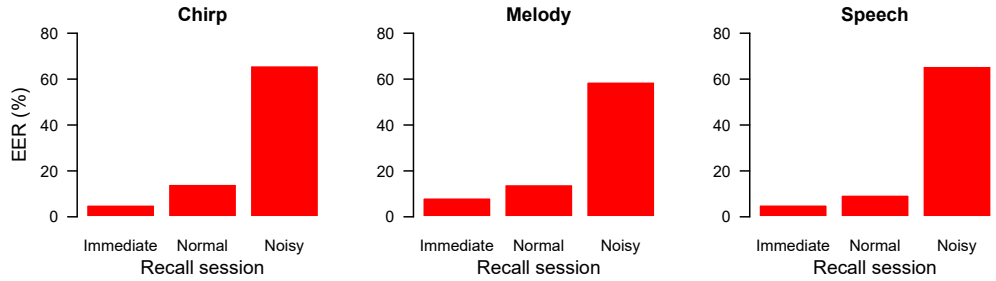

Figure 15: EERs measured on the in-air microphone signals.

Table 6: Scores for individual SUS (5-point Likert scale from "Strongly Disagree" (1) to "Strongly Agree" (5), starred items are inverted to calculate overall score) and NASA TLX (0-20 scale) questions. Mean and SD (in brackets) are presented.

Questionnaire	Question	Response	
		Enrollment	Authentication
SUS	I think that I would like to use this system frequently.	3.47 (0.80)	4.00 (0.71)
	* I found the system unnecessarily complex.	1.65 (0.79)	1.36 (0.60)
	I thought the system was easy to use.	4.35 (0.61)	4.58 (0.56)
	* I think that I would need the support of a technical person to be able to use this system.	1.59 (0.71)	1.61 (0.66)
	I found the various functions in this system were well integrated.	3.82 (0.64)	3.82 (0.73)
	* I thought there was too much inconsistency in this system.	2.06 (1.03)	2.03 (0.88)
	I would imagine that most people would learn to use this system very quickly.	4.35 (0.70)	4.79 (0.42)
	* I found the system very cumbersome to use.	2.29 (1.21)	2.09 (0.95)
	I felt very confident using the system.	3.29 (1.05)	3.73 (0.84)
	* I needed to learn a lot of things before I could get going with this system.	1.24 (0.44)	1.39 (0.66)
TLX	Mental Demand	2.35 (4.06)	1.88 (3.77)
	Physical Demand	3.06 (3.85)	1.00 (1.50)
	Temporal Demand	1.24 (1.71)	0.36 (0.74)
	Performance Achieved	4.94 (6.74)	2.94 (5.86)
	Effort Expended	7.59 (7.63)	5.73 (8.01)
	Frustration Experienced	3.06 (4.04)	2.06 (4.37)

Table 7: "Clustering attack" FARs (threshold 0.5) show the imitation attack success rates based on the brow/mastoid classifiers. Subjects that belong to the same "C" were used as the attack set.

Cluster	Participant	Physical characteristic				FAR (%)
		Gender	Height (cm)	Weight (kg)	Head front/side length (cm)	
C1	Subject1	Female	161	50–55	15.9/16.9	0
	Subject3	Female	163	50–55	15.5/17.9	0
	Subject11	Female	158	45–50	15.8/18.4	0
	Subject15	Female	159	45–50	15.9/17.7	0
C2	Subject2	Male	184	65–70	16.9/19.2	0
	Subject10	Male	178	75–80	16.8/19.6	0
C3	Subject4	Male	168	60–65	16.1/19.5	0
	Subject7	Male	173	55–60	16.3/19.5	1.52
	Subject17	Male	165	55–60	16.1/19.1	0
C4	Subject5	Female	168	60–65	16.3/18.4	0
	Subject9	Male	170	65–70	16.5/18.6	0
C5	Subject6	Female	168	50–55	15.4/18.9	0
	Subject12	Male	174	60–65	15.6/18.8	0
	Subject14	Female	170	55–60	15.7/18.8	0