WristAcoustic: Through-Wrist Acoustic Response Based Authentication for Smartwatches

JUN HO HUH, Samsung Research, Republic of Korea
HYEJIN SHIN, Samsung Research, Republic of Korea
HONGMIN KIM, Ulsan National Institute of Science and Technology, Republic of Korea
EUNYONG CHEON, Ulsan National Institute of Science and Technology, Republic of Korea
YOUNGEUN SONG, Ulsan National Institute of Science and Technology, Republic of Korea
CHOONG-HOON LEE, Samsung Research, Republic of Korea
IAN OAKLEY, Ulsan National Institute of Science and Technology, Republic of Korea

PIN and pattern lock are difficult to accurately enter on small watch screens, and are vulnerable against guessing attacks. To address these problems, this paper proposes a novel implicit biometric scheme based on through-wrist acoustic responses. A cue signal is played on a surface transducer mounted on the dorsal wrist and the acoustic response recorded by a contact microphone on the volar wrist. We build classifiers using these recordings for each of three simple hand poses (*relax*, *fist* and *open*), and use an ensemble approach to make final authentication decisions. In an initial single session study (N=25), we achieve an Equal Error Rate (EER) of 0.01%, substantially outperforming prior on-wrist biometric solutions. A subsequent five recall-session study (N=20) shows reduced performance with 5.06% EER. We attribute this to increased variability in how participants perform hand poses over time. However, after *retraining* classifiers performance improved substantially, ultimately achieving 0.79% EER. We observed most variability with the relax pose. Consequently, we achieve the most reliable multi-session performance by combining the fist and open poses: 0.51% EER. Further studies elaborate on these basic results. A usability evaluation reveals users experience low workload as well as reporting high SUS scores and fluctuating levels of perceived exertion: moderate during initial enrollment dropping to slight during authentication. A final study examining performance in various poses and in the presence of noise demonstrates the system is robust to such disturbances and likely to work well in wide range of real-world contexts.

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

Additional Key Words and Phrases: smartwatch authentication, bone conduction, acoustic response

ACM Reference Format:
Jun Ho Huh, Hyejin Shin, HongMin Kim, Eunyong Cheon, Youngeun Song, Choong-Hoon Lee, and Ian Oakley. 2022. WristAcoustic: Through-Wrist Acoustic Response Based Authentication for Smartwatches. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4, Article 167 (December 2022), 34 pages. https://doi.org/10.1145/3569473

Authors' addresses: Jun Ho Huh, Samsung Research, Seoul, Republic of Korea, junho.huh@samsung.com; Hyejin Shin, Samsung Research, Seoul, Republic of Korea, hyejin1.shin@samsung.com; HongMin Kim, Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea, khm489@unist.ac.kr; Eunyong Cheon, Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea, beer@unist.ac.kr; Youngeun Song, Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea, soyo61@unist.ac.kr; Choong-Hoon Lee, Samsung Research, Seoul, Republic of Korea, choonghoon.lee@samsung.com; Ian Oakley, Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea, ian.r.oakley@gmail.com.

2474-9567/2022/12-ART167 $15.00
https://doi.org/10.1145/3569473

1 INTRODUCTION

Smartwatches support a very wide range of applications including many that are privacy or security critical, such as those that monitor or record health data, process or mediate financial transactions, or serve as a trusted device that enables access to other devices such as smartphones or laptops [21, 38]. As these applications proliferate, it is becoming increasingly important to secure access to smartwatches. However, explicit smartwatch authentication schemes, such as PIN or pattern, suffer from critical issues in terms of both usability and security. They suffer from the fat-finger problem [39]: displayed targets are small and obscured by a user's own finger leading to prolonged entry times and high error rates. This is particularly problematic for the precise input required during explicit authentication [33]. In addition, and partly to mitigate these usability issues, people tend to choose easy-to-remember and easy-to-enter PINs and patterns that are vulnerable to guessing attacks [3, 6, 28, 43].

To overcome these security and usability issues, recent research [7, 24, 46] has focused on developing implicit biometric authentication schemes for smartwatches. Cornelius et al. [7] proposed a solution that captured wrist bio-impedance (tissue response to electrical current), data that should vary with the arrangement of relatively stable internal structures within the wrist. A single day field study on eight participants achieved 13.10% average equal error rate (EER). Watanabe et al. [46] measured ultrasonic sound signal reflection on the wrist and used this data to authenticate users. They recorded signals during four different hand poses, and combined these to attain an average EER of 2.94% with nine participants. Recently, Lee et al. [24] study the authentication performance that can be achieved with wrist vibrations generated by the haptic feedback motors built into a commercial smartwatch. In a two session lab study of 20 participants, they achieve an EER of 1.37% in their first session, and apply this decision threshold to achieve a false rejection rate (FRR) of 4.99% in their second session.

While this work showcases the potential of wrist based biometrics, we note a number of limitations in the evaluation methods it deploys. Perhaps most importantly, while this body of work has proposed combining multiple hand poses to improve authentication accuracy, it offers very limited insight into the details of such a scheme. We know little about how sensitive systems are to hand pose variations, how accurately and reliably users can produce various hand poses, and how users respond to the idea of performing one or more specific hand poses to unlock their watch. This paper demonstrates that through-wrist biometrics can be highly sensitive to hand pose changes and explores the implications, in terms of both authentication accuracy and usability, of this issue. We argue that this type of in-depth exploration of the impact of hand pose variations, analyses that are currently lacking in the literature, are essential to understand the practicality of wrist-based biometric systems. Further, prior work suggests that there may be variability in data recorded over several temporally separated sessions [7, 24] but provides little in the way of data directly addressing this issue. We identify a need for studies that examine how performance may change over time due to factors such as inevitable variations in the hand poses users adopt during authentication. Only by assessing and understanding longitudinal performance, and the factors that can introduce variability over multiple sessions, can we be confident that wrist-based biometrics are practical, effective and usable. In addition, prior work relies on presenting EER results based on optimal thresholds for studied data sets. Determining such thresholds, however, is likely highly impractical for a real world deployment [42]—in such settings, fixed threshold values are more realistic. We identify a lack of literature contrasting and examining system performance in terms of both optimal ERRs and metrics (e.g., FRRs) based on practical and readily deployable fixed decision thresholds.

Building on prior audio/vibration approaches, we propose *WristAcoustic*, a novel biometric authentication system for smartwatches based on through-wrist sound conduction. In our system, a white noise signal is played through a surface transducer (or bone conduction speaker) on the dorsal wrist, and acoustic signal responses are measured by a contact (or bone conduction) microphone on the volar wrist. We measure three separate responses for simple *relax*, *fist*, and *open* hand poses. Differences in the internal wrist structure between individuals leads to characteristic response patterns, which we use as features to train three pose-specific authentication classifiers.

We also use data from these poses together in a set of ensemble blenders. To evaluate the performance of WristAcoustic we conducted four user studies. Through the first single-session study (N=25) we demonstrate the effectiveness of combining multiple poses. Our best performing blender uses all three poses to achieve an average EER of 0.01% on this data set. To measure multi-recall performance, we conducted a 5-recall session study (with an interim period of at least four hours), demonstrating reduced performance of 5.06% EER with the same three pose blender. The performance degradation is due to increasing variability in the way people perform their poses over time. We then conducted a usability study (N=15) to measure users' experience (e.g., perceived exertion and workload) in performing multiple hand poses during system enrollment and authentication. Finally, we conducted a study (N=10) to document performance of the system while users adopt various arm poses and in the presence of noise (both audio and vibration). We summarize our key contributions and findings:

- **Development of an authentication system based on audio transmission through the wrist.** To the best of our knowledge we are the first group to build a through-wrist prototype combining a surface transducer and contact microphone on opposite (dorsal and volar) sides of the wrist, and also to study the feasibility of using the signal responses they generate and capture to authenticate smartwatch users.
- **Evaluation of multi-pose based authentication performance.** Our first study shows that training separate classifiers for three different hand poses, and using them in an ensemble blender—as opposed to just using a single pose classifier—substantially improves authentication accuracy: we achieved an average 0.01% EER with this blender, substantially outperforming prior approaches to wrist-based biometrics.
- **Characterization of multi-session performance.** Results from our multi-session study indicated that people do not perform hand poses consistently over time. These variations in user behavior present major challenges to achieving low error rates over prolonged periods. Retraining approaches, in which data from recall sessions is used to update classifiers, led to improved performance, ultimately achieving a mean EER of 0.79% with all three poses. Participants struggled more with the relaxed pose—often loosely forming a fist or modestly splaying their fingers instead. A blender using data from only the fist and open poses, which represent extreme, and thus distinct, points on the scale of finger flexion/extension showed more consistent multi-recall performance: 2.76% EER without retraining and 0.51% EER after retraining. Our final recommendation is to make use of those two poses.
- **Usability evaluation.** Requiring users to consistently perform multiple hand poses during authentication may impose new usability challenges. To that end, we propose an optimized system configuration that achieves relatively rapid multi-pose setup and recall times while maintaining high authentication accuracy. Our usability study shows that the perceived level of exertion while performing the fist and open poses during enrollment is moderate (average Borg CR10 scores of 3.1 and 3.3, respectively) and slight during recall (scores were 2.1 and 2.3). The average system usability scale (SUS) scores for enrollment and recall were 76.9 and 76.4, ratings associated with "good" usability [1]. In addition, NASA TLX scores indicate participants experience low levels of workload (relative to, for example, the broad spectrum of tasks analysed by Grier [11]): 3.76 and 3.42 for enrollment and recall. These results indicate that WristAcoustic is easy to learn and use. We note that the usability of performing one or more specific hand poses during watch authentication has not been studied in prior work; we contribute the first data on this topic.
- **Robustness to variations in pose and background noise.** Real world watch authentication will take place in a wide variety of contexts and in the presence of various forms of interference, such as background noise, motion or vibration. To explore the robustness of WristAcoustic to such variations, we conducted a final study (N=10) combining the optimized enrollment protocol defined and tested in the usability evaluation with 22 recall sessions conducted with users adopting various arm poses and experiencing various forms of noise (e.g., music, ambient noise) and motion/vibration (e.g., from a smart phone). The results

demonstrate WristAcoustic performs well in these settings, achieving a mean FRR of 0.5% (corresponding to a single failure by a single participant) and a mean FAR of 1.67%.

- **Threshold evaluations.** We present all study results in terms of both optimal thresholds (e.g., EERs) and an unbiased 0.5 threshold. This analysis reveals that FRRs can be high during the initial recall sessions—due to more cluttered score distributions forming below the 0.5 level. In order to combine the practical benefits of fixed thresholds with an acceptable level of initial performance, we recommend adopting relaxed (e.g., 0.2) thresholds immediately after enrollment, and adjusting this to an unbiased threshold (e.g., 0.5) after several rounds of retraining. This approach sidesteps the need to collect representative samples, and based on those samples, infer and validate theoretically optimal per user threshold values. We argue it is therefore highly practical and suitable for real world system deployments.

2 RELATED WORK

2.1 Explicit Authentication on Smartwatches

Unlocking smartwatches with PINs or graphical patterns can be challenging, as watch screens and graphical targets are small and input needs to be precise [33]. These usability problems are exacerbated by fundamental security problems: people tend to choose easy-to-remember and quick-to-enter PINs or patterns that are vulnerable to dictionary attack. An attacker with perfect knowledge [2] who uses an entropy estimation approach may successfully crack 8% to 19% of 4-digit PINs within 10 online guesses [28]. Many users also select memorable dates as PINs [3]—such PINs would be additionally vulnerable to both brute-force attacks and those that leverage personal data. Pattern exhibit similar trends: crack rates range from 13.33% in a real-world mobile application [5] to 32.55% in an MTurk study [6]; various well-documented biases [6, 30] also influence pattern selection. Given the fact that smartwatches increasingly offer security-critical features (e.g., unlocking paired laptops or monitoring health), we argue it is necessary to develop secure and usable authentication schemes that do not require users to enter passwords on small watch screens and are not susceptible to guessing attacks.

2.2 Smartwatch Biometric Authentication

Inspired by these motivations, researchers have begun to explore various biometric based authentication schemes for smartwatches. Cornelius et al. [7] study the feasibility of using on-wrist bioimpedance (tissue responses to electrical current) to authenticate users, achieving 13.1% EER on eight participants. Zhao et al. [53] explore the use of wrist photoplethysmography (PPG) signals—PPG monitors blood volume changes from light reflected on the skin—to authenticate users. In a continuous authentication scenario, their solution achieves approximately 10% FRR and 10% FAR (false acceptance rate). Watanabe et al. [46] explore the feasibility of using ultrasonic signal reflection on wrists to authenticate users. Their approach requires users to perform four different hand poses. Based on a nine participant lab study they report a 2.94% EER. However, their work suffers from the limitation that the same set of imposters (or attackers) was used to both train and test the system, likely resulting in artificially elevated performance. For a more realistic/meaningful evaluation, these train and test imposter sets need to be distinct. Finally, Lee et al. [24] use low-frequency vibration motors available on smartwatches to generate vibrations, and measure vibration responses using accelerometer and gyroscope sensors to authenticate users. Through a two-session recall study conducted on twenty participants, they reported 1.37% EER on the first day, and 4.99% FRR after seven days. While this level of performance is promising, doubts remain about the information used to achieve it: signal frequencies (170–240 Hz) exceed sample recording frequencies (100 Hz). In addition, we argue that the reduction in performance in their follow up session indicates some uncertainty regarding the long-term reliability of their scheme. This highlights a need for more substantial multi-session studies in this area. In this paper, WristAcoustic also demonstrates a similar upward trend in FRRs over multiple sessions but is able to stabilize this effect by introducing a model retraining process.

To summarize: we note two inadequacies in existing literature. First, most studies have assumed that people would authenticate with a single (typically relaxed or undefined) hand pose. However, there may be considerable variability in how hand poses are enacted over prolonged periods. How such variations will affect overall performance is currently unknown. Second, prior work primarily focuses on reporting EER results based on one or two recall sessions. We argue this focus on short-term performance under optimal threshold values likely overlooks key aspects of performance. For example, it may obscure issues related to deteriorating error rates over time (e.g., due to changing behaviors), and miss the need to update or retrain classifiers. In this paper, we report EERs in tandem with results using fixed decision thresholds to elucidate the ways in which EERs can provide incomplete or misleading information. We believe these issues are of particular importance when considering practical real-world deployments [42]: realistically, optimal thresholds will not be available, and performance with respect to fixed thresholds will need to be studied.

2.3 Bioacoustics and Bone Conduction

The hand and arm are effective mediums for transmitting audible bioacoustic signals. Prior work exploring these signals has focused on using sensors on the hand, wrist or arm to achieve goals as diverse as localizing finger taps [14, 49], identifying held objects [23], detecting finger gestures [50], and recognizing hand poses [51]. We know of no prior work exploring acoustic response signals on the wrist for authentication. Acoustic sensing has been studied extensively in the context of head-mounted wearables too—authenticating voice commands through speaking-induced body sounds [9, 27], authenticating users on smartglasses through bone conduction acoustic responses [37], and authenticating users on wireless earphones through in-ear sound reflection [10, 45]. In closely related work, Schneegass et al. [37] explored the feasibility of transmitting white noise acoustic signals on the skull through a bone conduction speaker, and measuring transferred signal responses to authenticate users. Their approach, however, uses an air-gapped, in-air microphone to record responses—hence, it is unclear as to what types of energy transfer paths are being analyzed; the majority of energy transfers may occur through the glasses frame rather than the skull. They report a 6.9% mean EER with ten participants. In this paper, we build on this concept and measure the acoustic signal responses of wrists to authenticate smartwatch users. We use a surface transducer and contact microphone in place of Schneegass et al.'s in-air microphone, focus on recording through-wrist energy transfer and conduct considerably extended empirical evaluations.

3 DESIGN AND IMPLEMENTATION

In this section, we describe the hardware implementation and optimization efforts. We also explain the overall data processing pipeline, covering classifier features and training.

3.1 Theory of Operation

The human wrist is made up of many different anatomical parts such as skin, bones, tendons, ligaments, nerves, and blood vessels. These anatomical parts differ in layout, size, and density from person to person. In 2009, Kumar et al. [22] noted that despite the promising stability of internal hand and wrist structures in diverse environmental conditions (such as temperature and humidity), challenges in reliable imaging has meant that relatively little literature has explored their potential as a biometric. The advent of smartwatches is changing this and a range of more recent work has documented the strong potential of wrist imaging for biometrics by assessing properties such as the visual spectral response [19], the impedance [7] and the response to vibration [24]. This work strongly suggests that the internal and physiological structures of the wrist can provide useful information for biometrics. We also note that these internal structures, and the external form of the wrist, also vary in response to different hand articulations (such as making a fist or open palm) with sufficient regularity that monitoring them is reported to support accurate recognition of hand pose [18]. Furthermore, the wrist and indeed, the human body in general,

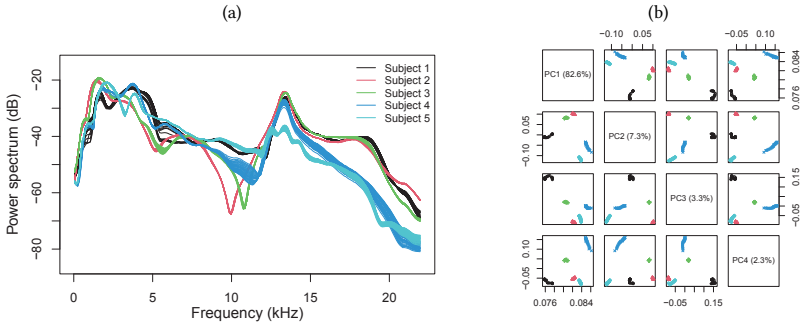

Fig. 1. (a) The power spectrum of 30 response signals recorded under the "sit" and "fist" conditions for five subjects; (b) the first four principal component scores from 150 power spectrum curves.

is also by-and-large non-compressible and, therefore, an excellent carrier of vibration [23]. Indeed, bio-acoustic techniques, such as ultrasound [29], are commonly used for high resolution medical imaging of internal body structures. These three properties combine to suggest that when a sound is emitted from a surface transducer on the dorsal wrist, and the resultant vibrations transmitted to a contact microphone on the volar wrist, they will be distinctively transformed by their passage through the particular arrangement of skin [14], bones [41], soft tissues and fluids [40] they encounter. Figure 1(a) shows the power spectrum of 30 responses to an identical signal recorded from the wrists of five different individuals, each holding a specific hand pose: a fist. This figure shows clear differences in the data captured from different individuals and also highly consistent patterns among the samples recorded from each individual. This combination of high variability between individuals and a low variability for a given individual (and pose) suggests that the range of anatomical differences in human wrists may transform audio signals sufficiently uniquely that they can serve as an effective biometric. Figure 1(b) reinforces this point. It shows the principal component (PC) scores on the first 4 PC directions (which explain more than 95% of the variation in the power spectrum curves). It clearly indicates that the response signals between different individuals are readily separable.

3.2 Hardware Design and Implementation

We constructed two sensor enhanced wrist-bands for the work reported in this paper. We note that watch wristbands have been frequently proposed as a site for embedded electronics that seek to enhance smartwatch functionality, such as extending the touch input space [36], providing an larger display [20] or supporting various forms and modalities of hand gesture recognition [52]. Both our wristbands were built around simple 22 millimeter fabric wristwatch straps and involved separate modules, each of which could be slid along the strap in order to accommodate different wrist sizes, for the vibration transducer and contact microphone. Hardware components in both versions of the wristband modules were identical. For sensing, we used a Knowles BU-21771 contact microphone, previously deployed in a range of closely related prior work studying through body audio transmission [12, 35] together with a modified version of the amplifier design proposed for this device by Zhang et al. [50]. 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: during piloting, we ultimately selected a gain of two. For actuation, we used a commercially available surface transducer[1] also deployed in prior work on active in-body audio transmission [51]. We drove the speaker using a breakout for the TPA2012 class D audio

[1]https://www.adafruit.com/product/1674

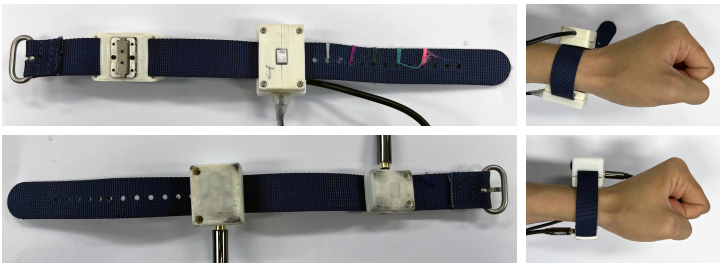

Fig. 2. Hardware prototypes used in first (top) and second (bottom) studies. Units on the left of the watch straps are the surface transducers and units on the right are the contact microphones. Image panels on the right show worn prototypes.

amplifier[2] configured for 24 dB gain. Both amplifiers were located immediately adjacent to the microphone and transducer (i.e., mounted on the wrist) in order to minimize the impact of RF noise.

All modules themselves were 3D printed in skin-safe Thermoplastic polyurethane (TPU), specifically NinjaTek SemiFlex[3]. We note TPU is often indicated for use as a vibration damper. The primary differences between the two prototypes were in the design of these enclosures: in the first study, the modules exposed transducer and microphone surfaces directly to the skin and used jumper cables for connections. In addition, the 350mAh 3.7V li-ion battery used to power the transducer was located off the wrist. We note this battery is typical for the current generation of smartwatches. For example, the Apple Watch 7[4] and Samsung Galaxy Watch 4[5] product families feature 3.8V li-ion batteries rated at, respectively, up to 309mAh and 350mAh. In the second study, a thin film of TPU (0.35 millimeters) covered actuator/sensor surfaces, we used 3.5 millimeter audio jack connectors and shielded cables and also integrated the 3.7 li-ion battery into the wrist unit. These changes were made to increase the reliability of the prototype (to better support repeated study sessions over a protracted period) and to more closely follow commercial products in this space—systems in which actuators and sensors are invariably enclosed in plastic to enhance robustness and reliability. Both prototypes are shown in Figure 2.

To ensure close synchronization between signal generation and recording we opted to use an embedded audio system based around the Teensy 4.1[6] platform, and its audio expansion board and library[7]. Using this system, we developed software to play 2 seconds and record 2.025 seconds of single channel 44.1 kHz 16 bit audio simultaneously. Recording commenced immediately (between 7-8 milliseconds) before playback to ensure all transmitted signals were captured. All samples (for both playback and recording) were buffered in RAM to minimize the impact of latency during data loading or saving. As RAM on such embedded systems is highly limited, the system flushed recorded data to an SD card in-between each recording. Additionally, the system was capable of loading arbitrary new samples for playback. The system was connected to a host PC to coordinate, and execute experimental procedures via either Bluetooth or a wired RS232 connection. This system enabled accurate synchronization of playback and recording activities, sufficient to support our empirical objectives.

3.3 System Overview

Use of WristAcoustic involves three distinct phases of activity: enrollment, authentication, and retraining.

[2]https://www.adafruit.com/product/1552
[3]https://ninjatek.com/
[4]https://en.wikipedia.org/wiki/Apple_Watch
[5]https://en.wikipedia.org/wiki/Samsung_Galaxy_Watch_4
[6]https://www.pjrc.com/store/teensy41.html
[7]https://www.pjrc.com/store/teensy3_audio.html

3.3.1 Enrollment. During enrollment, the system collects reference response signals, pre-processes them, extracts classification features, and trains user-specific binary classifiers. During this stage, users are required to put on their watches several times and perform one or more poses (relax, open, or fist). While performing each pose, a cue signal (e.g., white noise) is played through a surface transducer resting on the dorsal wrist. Transferred response signals are recorded through a contact microphone situated on the volar wrist, and processed into one or more of the signal processing features presented in Section 3.4. Multiple samples per pose are collected each time the watch is worn (subsequently referred to as a *donning*). In Section 5.10 we introduce and justify recommendations for the number of times (four) the device should be donned and the number of samples to be collected each time (five) it is put on. When data collection is complete, user-specific binary classifiers are trained for each pose (and their blended combinations) using the features extracted from reference signals and a pre-deployed imposter train set, which comprises 100 samples collected from a separate group of 10 individuals during the system development stage. This is a prerequisite for training binary classifiers on smartwatches: a small-scale data collection effort would be needed to acquire approximately 100 samples per pose, process them into features, and deploy them on the watches before shipping. Since only 20 genuine user samples and 100 imposter train samples are used for training a lightweight (pose-specific) classifier, all model training can be performed quickly on smartwatches within a few seconds (see Section 5.11).

3.3.2 Authentication. WristAcoustic is a smartwatch unlock scheme—implying that, during authentication, users don their watch, and perform the required (one or more) poses synchronized with audio cues (e.g., "beep" sounds played in ascending frequencies) to unlock their watch. Authentication response signals are collected and processed into features. This data is then submitted separately to the appropriate classifiers in order to generate probability scores. The probability scores are then compared against a pre-defined threshold value. Users are successfully authenticated if the probability scores are higher than the threshold value.

3.3.3 Retraining. Finally, the *implicit* retraining phase involves accumulating the response signal samples from successful authentication attempts, and using those new samples to periodically retrain authentication classifiers—this ensures that the classifiers adapt to users' changing donning and posing behaviors, and physiological and biological characteristics. Samples from successful WristAcoustic authentication sessions are selected, and added to the training set. Samples from failed sessions are also added if users eventually manage to unlock their watch through WristAcoustic or by using another authentication method (e.g., PIN or pattern). New samples are collected automatically, and users are never asked to perform additional sessions solely for the purposes of extending data collection. To maintain a balanced train set, our sampling algorithm ensures that those new samples (representing users' latest posing behaviors) make up approximately half of the training set; the remainder is always selected from the original enrollment set. Considering battery constraints, we envisage that retraining could be performed once a day—ideally at night while users are asleep and their watches are being recharged.

3.4 Pre-processing and Feature Selection

Our hardware prototype records the responses to each two-second stimulus cue signal bracketed by a short (25 millisecond) buffer to accommodate any system latencies and ensure the capture of full audio signals. To precisely determine the start and end of recorded samples we computed the ratio of signal variances between an empty signal (where cue stimuli are absent) and the time frames of a given response signal, and checked this ratio against a threshold value. We also applied a bandpass filter between 300 Hz and 19 kHz to reflect the frequency response range of our surface transducer. The details of these pre-processing steps are explained in Appendix A. Using those pre-processed samples, we compute each of the following features over the full two seconds of recorded audio.

3.4.1 Power Spectral Density (PSD). PSD is a frequency domain feature that describes how the power of a signal is distributed over a given frequency range. There are multiple methods to estimate PSD. As we expect the response to a stationary white noise stimulus to also be stationary, we estimated PSD by calculating the periodogram of the whole signal using FFT, and smoothing the periodogram with modified Daniell smoothers [4]. For smoothing and frequency sampling, we used 50 Hz windows without overlapping in the range from 300 Hz to 19 kHz: 375 windows in total. The resulting 375 item log-PSD vector was used as a feature. Note that one can use a time-averaged periodogram for PSD calculation (e.g., Welch's method [47]). The time-averaging approach involves dividing the signal into overlapping segments, computing a periodogram for each segment, and averaging the periodograms. For stationary signals, the time-averaging approach results in a coarser frequency resolution; calculating a smoothed periodogram, the method used here, is therefore preferred.

3.4.2 Transfer function. A transfer function is a frequency domain feature that measures how a cue stimulus signal s is converted to a response signal x by passage, in this case, through the wrist. 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 of s. We used the same frequency range and windowing parameters as for PSD (leading to 375 windows) and used log magnitude of this transfer as a feature.

3.4.3 Mel-frequency Cepstral Coefficients (MFCCs). MFCCs are time-frequency domain features widely used in speech recognition systems. They have also been used in prior authentication systems [37]. We computed 39 cepstral coefficients based on a Hamming sliding window. We used a window length of 25 milliseconds and overlap length of 10 milliseconds. We computed means across the time-series data, and used those 39 mean values as features.

In addition to the feature sets described above, we also experimented with simple concatenation of two feature sets (e.g., concatenating PSD and MFCC features to create an one-dimensional vector consisting of 414 features).

3.5 Classification Algorithms

We build our authentication classifiers using a binary support vector machine (SVM) with radial basis function kernel. We chose a binary SVM classifier as it showed peak performance among the following classification algorithms: SVM, Random Forest (RF), XGBoost, and Neural Network for binary classification, and one-class SVM and kNN. Throughout these tests, we used the python hyperopt package to optimize hyperparameters for the classification algorithms.

4 STUDY 1: SINGLE-SESSION STUDY FOR PARAMETER OPTIMIZATION

We conducted a single-session study to characterize basic system performance, and explore the impact of different hand poses (*relax*, *fist* and *open*), and common body postures (*sitting* and *standing*). This section presents the study methods and authentication accuracy results. We note that rather than test a realistic authentication system, the goal of this study was to identify the best performing combinations of poses, postures, features, and classifiers. As such, the study collected extensive data: multiple cue repetitions from multiple sessions spanning a prolonged study session of approximately one hour. The study protocols were approved by the university's IRB. We explicitly informed the participants that the purpose of the data collection was to develop and evaluate a smartwatch authentication solution based on through-wrist acoustic response information.

4.1 Methods

This study involved participants wearing our wristband prototype while we played audio cues and recorded the responses. Based on closely related prior work [37], we selected a white noise signal with 44.1 kHz sampling rate (frequency range between 0 to 22.05 kHz) as the audio cue. In addition to transmitting sound through the wrist,

this cue produced an audible noise that resembled that of the vibration feedback on a smartphone. We chose a cue duration of two seconds to keep authentication times close to those reported for explicit PIN based smartwatch authentication [31, 33]. Participants were asked to put on our wristband prototype (see Section 3.2) on their left wrist five times in total. Each time they wore it, they performed three hand poses—*relax*, *fist* and *open*—in two body postures—*sitting* and *standing*. For each combination of these variables, we played and recorded responses to 30 samples. As such, we logged 900 samples (3 gestures × 2 body poses × 5 donnings × 30 samples) from each participant. To reduce wear and tear on the prototype and ensure similarity (e.g., band tightness, approximate watchband location) between donnings, study moderators supported participants in putting on and removing the wristband. In the sitting context, the participants were asked to rest their elbows comfortably on the study table with their hand held in free space above the table. In the standing context, the participants were asked to have their arms resting downward, adjacent to their thighs, in a relaxed pose. After the data collection phase, the participants were asked to complete a short demographics questionnaire. The study took approximately one hour to complete, and the participants were compensated with the equivalent of 13 USD in local currency.

4.2 Demographics

Out of the 25 participants, 14 were female (see Table 6 in Appendix B). The average age was 23.3 years (SD=5.5), and 88% were right-handed. We measured wrist sizes as these could impact authentication performance. The average wrist circumference for female and male participants was 15.2 centimeters (SD=1.2) and 16.2 centimeters (SD=0.8), respectively. Unsurprisingly, we observed a statistically significant difference in the wrist sizes between female and male participants (two-sample t-test $p = 0.024$). We also recorded height and weight: the average height and weight for female participants were 161.5 centimeters (SD=4.2) and 59.3 kilograms (SD=8.0). For the male participants, these data were 174.4 centimeters (SD=4.4) and 67.0 kilograms (SD=7.2). We observed a strong correlation between the wrist size and weight (Pearson's correlation coefficient 0.85, $p < 0.0001$), and a moderate one between wrist size and height (Pearson's correlation coefficient 0.56, $p = 0.003$).

4.3 Evaluation Setup

We evaluated how the authentication performance differs between the three hand poses and sitting/standing body postures, and the effectiveness of combining multiple hand poses as ensemble blenders. To explore these variations, we used a binary SVM classifier and the features described in detail in Section 3.4. For each classifier we trained, we divided the 25 participants into two groups: 10 participants were used for imposter training and the remaining 15 participants were set as genuine users. For each genuine user, the imposter set was used to train a binary classifier and the other 14 genuine users served for unknown user (imposter) testing. In addition, due to the fact that a single evaluation of an authentication system with a specific set of genuine users and imposters does not represent authentication performance in general, we repeated the classifier training process with a random selection of 10 imposters and 15 genuine users 100 times. We report all results as means over all hundred permutations. To examine the effects of hand poses on authentication, we evaluated two authentication models: (i) per-pose authentication, and (ii) multi-pose authentication.

In per-pose authentication, a single hand pose is required for authentication. So, for each pose, we used the first k donning sessions of each genuine user in both sit and stand positions for training. In our evaluation, we set $k = 3$ and 4, which resulted in 180 and 240 samples for genuine user training. The last donning session for each pose was always used for genuine user testing (corresponding to 60 samples). To create a balanced training set, we used 300 samples for imposter training: 30 samples from each of the ten imposters. As attackers can try any hand pose under any body posture to compromise the target device, the samples for each imposter were composed of 5 randomly selected samples for each combination of the three hand and two body poses. For unknown user

(imposter) testing, we used 60 samples (3 hand poses × 2 body poses × 5 donnings × 2 samples) from each of the other 14 subjects in the genuine user set: 840 samples in total.

In multi-pose authentication, we assumed users were requested to perform a given set of hand poses. We examined all possible combinations—three hand pose pairs and one triple. We used ensemble models of multiple per-pose classifiers each with equal weight for authentication. We used the same training sets as in the per-pose authentication for training the ensemble model, while 60 pairs or tuples of samples of the appropriate hand poses were used for genuine user testing. For unknown user (imposter) testing, we chose 4 pairs or tuples of samples under each of all possible hand pose sequences (i.e., nine combinations of two hand poses and 27 combinations of three hand poses) and ensured that both sit and stand positions were included, as attackers can try any sequence of hand poses for authentication. This resulted in 504 pairs of samples covering any two hand poses, and 1,512 tuples of samples for all three hand poses.

We use both (1) equal error rate (EER), and (2) half total error rate (HTER) as evaluation metrics. EER is defined as the rate at which false acceptance rate (FAR) and false rejection rate (FRR) are equal. We use individual user thresholds to compute EERs. FRR measures error rates for incorrectly classifying users' samples as attackers' samples (affects usability); FAR measures attack success rates (affects security). EER values are readily obtained from the ROC curves, but are dependent on test data—indicating that for a given user, inferring a threshold value would depend on a combination of both the user's own data and data from other users. For this reason, a fixed threshold (e.g., probability = 0.5) can also be used across users. In that case, HTER, defined as the average of FAR and FRR, can be used as the preferred performance measure.

4.4 Results

To simplify data presentation, we select the top performing feature combination, which was the MFCC and PSD concatenation feature, and summarize the key accuracy results in Table 1. The full results containing all individual feature sets can be found in Appendix C (see Tables 7 and 8). These results show the mean and standard deviation (over 100 random permutations) of the average genuine user EER and HTER for each of the three hand poses. The results indicate that hand pose affects authentication performance. Specifically, the fist pose showed stronger performance compared to the other two poses. Overall, all three individual feature sets demonstrated strong performance: e.g., achieving between 0.7–2.41% average EER, and 3.23–3.39% average HTER on the fist pose with $k = 4$. As for concatenated features, MFCC and PSD concatenation showed superiority across all three hand poses compared to all other feature sets. It led to average EERs for fist, open, and relax poses of 0.72%, 1.59%, and 1.75%, respectively, when $k = 4$. We found that combining multiple hand poses significantly reduces both EER and HTER, demonstrating that each hand pose offers unique information that enables these ensemble approaches to be effective. Combining all three gestures performed the best, showing an average EER of 0.01% and an HTER of 0.39% with both PSD and MFCC features when $k = 4$. Among the two-pose ensemble blenders, the combination of fist and open demonstrated the lowest error rates of 0.11% EER and 0.67% HTER, when $k = 4$.

4.5 Context Analysis

In our user study, we collected each participant's bio-acoustics responses from five donning sessions, each featuring three different hand poses and two different body postures. These different contexts can affect the response signal. For example, extension and flexion of the fingers during the open and fist gestures involves the tendons, ligaments and muscles in the wrist, causing numerous internal changes. Blood pressure in the sitting pose (with the hand roughly level with the heart) would be expected to be higher than in the standing pose (with the wrist well below the heart). Subjectively, this change of pose corresponds to a sense of tightening of the band during standing. To better understand how authentication accuracy might vary when users adopt different hand or body poses, we opted to examine results from a single balanced set of imposters and genuine users: we

Table 1. Authentication accuracy for per-pose classifiers and multi-pose ensemble blenders. Mean FAR, FRR and EER measured over 100 permutations (randomly selecting 15 genuine users and 10 imposters each time). k indicates the number of donnings used for training.

Feature	Hand pose	EER (%)		Probability threshold = 0.5					
				$k = 3$			$k = 4$		
		$k = 3$	$k = 4$	FAR (%)	FRR (%)	HTER (%)	FAR (%)	FRR (%)	HTER (%)
		Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
MFCCs + PSD	Fist	1.48 (0.64)	0.72 (0.43)	1.99 (1.05)	8.02 (3.02)	5.01 (1.48)	2.22 (1.07)	3.99 (2.44)	3.10 (1.27)
	Open	2.23 (0.82)	1.59 (0.65)	2.76 (1.28)	8.84 (3.68)	5.80 (1.70)	3.09 (1.39)	4.52 (2.26)	3.81 (1.18)
	Relax	2.01 (1.06)	1.75 (1.02)	2.51 (1.43)	7.88 (3.68)	5.19 (1.92)	3.06 (1.66)	3.69 (2.57)	3.38 (1.56)
	Fist/Open	0.44 (0.35)	0.11 (0.13)	0.62 (0.52)	3.54 (2.16)	2.08 (1.12)	0.79 (0.59)	0.54 (0.63)	0.67 (0.45)
	Fist/Relax	0.28 (0.19)	0.19 (0.16)	0.48 (0.41)	5.08 (2.61)	2.78 (1.29)	0.74 (0.54)	1.09 (1.42)	0.92 (0.72)
	Open/Relax	0.54 (0.40)	0.34 (0.28)	0.39 (0.30)	6.78 (2.06)	3.58 (1.03)	0.63 (0.40)	4.16 (1.61)	2.39 (0.80)
	Fist/Open/Relax	0.17 (0.14)	0.01 (0.02)	0.24 (0.25)	3.41 (2.47)	1.82 (1.24)	0.39 (0.34)	0.39 (0.52)	0.39 (0.31)

manually selected this set from the one hundred previously evaluated permutations. For the ten imposters, six were female and the mean age, wrist size, weight, and height were, respectively, 23.6 years, 15.39 centimeters, 63 kilograms, and 167.76 centimeters. For the 15 genuine users, eight were female and these figures were 23.1 years, 15.75 centimeters, 62.5 kilograms, and 166.8 centimeters, a close match to the imposters. HTERs with this set for per-pose classifiers (trained with PSD) and $k = 4$ were 1.33%, 5.59%, and 2.66%, respectively, for fist, open, and relax.

Table 2 shows how authentication accuracy changes when we test with hand poses that are different to the trained poses. Average FRRs (measured across all 15 genuine users) increase steeply when a genuine user seeks to authenticate with the wrong hand pose—using the open pose to authenticate with a fist-trained classifier, for example, led to a 36 fold increase in average FRRs. We also tried training a single classifier with all three hand poses (randomly selecting 10 samples per pose from each of the k donning sessions): these classifiers (the "all" condition) demonstrated modestly worse HTERs for the individual poses of fist (3.74%) and relax (3.74%), suggesting that (1) hand gestures are critical for stable on-wrist authentication using our technique, and (2) training separate per-pose classifiers is more effective than training a single classifier that accommodates all three poses.

Using a similar process, and the same set of users, we also evaluated how sitting and standing postures affect authentication accuracy for the top performing fist hand pose. Table 3 shows the resulting cross-body posture authentication accuracy. If the fist-specific classifier was trained under either sit or stand body posture and then tested on the other posture, FRRs increased by 3–9 times. On the other hand, when samples from both sit and stand postures were used for training (we randomly selected 15 samples per posture from each of the k donning sessions), peak authentication accuracy was achieved. This suggests that training a diversity of body postures in enrollment phase can be effective, albeit at the cost of requiring more complex procedures.

4.6 Physical Characteristics

To analyze how authentication accuracy changes by physical characteristics of a genuine user, we picked the fist-based classifier trained with PSD features and $k = 4$. In each of the 100 permutation (genuine/imposter) user sets, we evaluated whether there is a statistically significant difference in authentication accuracy by each physical characteristic (see Figure 9 in Appendix D). Note that a two-sample t-test was used for gender, and correlation test with Pearson's correlation coefficient was used for age, wrist size, weight, and height. In these 100 sets, only wrist size (once) and weight (three times) led to statistically significant variations in authentication performance. We conclude that physical characteristics had little impact on authentication accuracy.

Table 2. Cross-pose authentication accuracy. Results in %

Training pose	Testing pose	$k=3$ FAR	$k=3$ FRR	$k=3$ HTER	$k=4$ FAR	$k=4$ FRR	$k=4$ HTER
Fist	Fist	0.97	8.39	4.68	1.21	1.44	1.33
	Open	0.97	51.56	26.26	1.21	52.78	26.99
	Relax	0.97	28.39	14.68	1.21	19.11	10.16
Open	Fist	1.21	53.00	27.10	1.84	40.89	21.37
	Open	1.21	19.33	10.27	1.84	9.33	5.59
	Relax	1.21	58.28	29.74	1.84	51.33	26.59
Relax	Fist	1.44	30.33	15.89	1.87	21.56	11.72
	Open	1.44	59.50	30.47	1.87	53.44	27.66
	Relax	1.44	9.11	5.28	1.87	3.44	2.66
All	Fist	3.41	6.67	5.04	4.14	3.33	3.74
	Open	3.41	10.39	6.90	4.14	7.00	5.57
	Relax	3.41	5.00	4.21	4.14	3.33	3.74

Table 3. Cross body posture authentication accuracy. Hand pose is fixed to the fist pose. Results in %.

Training posture	Testing posture	$k=3$ FAR	$k=3$ FRR	$k=3$ HTER	$k=4$ FAR	$k=4$ FRR	$k=4$ HTER
Sit	Sit	0.24	12.78	6.51	0.41	15.33	7.87
	Stand	0.24	62.78	31.51	0.41	59.56	29.99
Stand	Sit	0.44	41.11	20.78	0.71	46.22	23.47
	Stand	0.44	15.56	8.00	0.71	4.44	2.58
All	Sit	0.98	0.56	0.77	1.27	1.56	1.41
	Stand	0.98	15.78	8.38	1.27	1.33	1.30

Table 4. FARs of individual users from clustering attacks where unknown attackers are the other users in the same cluster of physical characteristics.

Cluster	Participant	Physical condition Gender	Height (cm)	Weight (kg)	FAR (%) F	O	R	F/O	F/R	O/R	F/O/R
C1	Subject1	female	164	50–55	33.33	5.00	1.67	5.56	0	0	0
	Subject16	female	162	45–50	0	0	0	0	0	0	0
C2	Subject3	female	155	50–55	0.83	0.83	0.83	0	0	0	0
	Subject5	female	157	50–55	5.00	35.42	5.00	8.33	2.78	6.94	3.24
	Subject9	female	159	55–60	0	0	7.08	0	0	0	0
	Subject17	female	157	55–60	8.33	4.58	8.75	0	0	0	0
	Subject21	female	158	50–55	0	0	2.08	0	0	0	0
C3	Subject6	female	165	60–65	0	2.22	5.56	0	0	5.56	1.85
	Subject13	female	163	60–65	3.33	4.44	0	0	0	0	0
	Subject18	female	163	60–65	0	0	0	0	0	0	0
	Subject22	female	163	55–60	0	7.22	0	0	0	0	0
C4	Subject15	female	171	65–70	3.33	0	0	0	0	0	0
	Subject23	female	165	65–70	0	0	0	0	0	0	0
C6	Subject4	male	170	50–55	0	0	0	0	0	0	0
	Subject14	male	170	55–60	0	0	0	0	0	0	0
C7	Subject7	male	170	65–70	1.67	0	0.56	0	0	0	0
	Subject10	male	173	65–70	0	1.11	3.33	0	0	0	0
	Subject12	male	174	65–70	0.56	0	6.67	0	0	0	0
	Subject19	male	175	65–70	0	0	0	0	0	0	0
C8	Subject20	male	181	75–80	1.67	1.67	0	0	0	0	0
	Subject24	male	180	70–75	0	0	0	0	0	0	0
	Subject25	male	180	70–75	0	0	0	0	0	0	0
Average					2.64	2.84	1.89	0.63	0.13	0.57	0.23

4.7 Robustness against Imitation Attacks

Biometric authentication systems, such as based on face or voice recognition, are often targeted with imitation or replay attacks that involve presenting media (e.g., photographs, audio recordings) to spoof the system. While such methods could be applied to WristAcosutic, there are major challenges: it would be inherently difficult to both acquire and accurately inject the response of a user's wrist to an acoustic signal. A more practical approach may be to recruit attackers who are physically and physiologically similar (e.g., in terms of height, weight, wrist size) to a target victim and simply ask them to don the watch and attempt to authenticate as normal.

To explore the robustness of WristAcoustic to such efforts, we conducted two analyses. We first explored the impact of gender, height and weight: we split participants into gender sub-groups, then applied k-means clustering to the weight and height characteristics of each sub-group (with $k = 5$) to identify seven clusters (encompassing 22 participants) composed of two or more participants with identical genders and similar builds. For each cluster participant, we then trained a binary classifier using an imposter train set of ten randomly

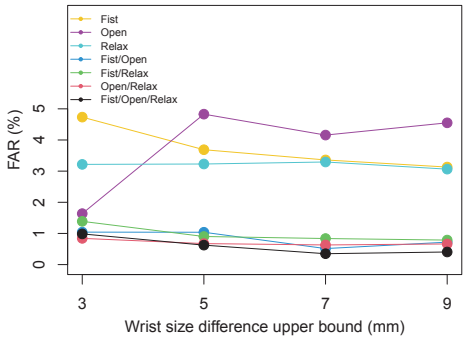

Fig. 3. Mean FARs from imitation attacks where unknown attackers are those whose wrist circumferences are within 3, 5, 7, or 9 mm of each genuine user.

selected participants from outwith their clusters and calculated FARs, representing attack success rate, using all participants from within their cluster. The results are shown in Table 4 and indicate that single pose systems can be vulnerable to this attack—although generally low, FARs peak at 35.42%. Two and three pose blenders showed improved performance with peak FARs of 8.33% and the vast majority of participants recording 0% FRRs. This result emphasizes the importance of combining multiple poses.

Our second analysis examined wrist size. For each participant we created four different groups of attackers based on wrist circumference similarity. These featured attackers with wrist circumferences less than 3mm, 5mm, 7mm and 9mm different from the participant. These groups featured a mean of, respectively, 3.4, 5.8, 7.8, and 9.3 individuals. For each participant, imposter sets (N=10) were again drawn randomly from the remaining, non-attacker, participants. We then trained classifiers and calculated FARs for each participant. Figure 3 shows the mean FARs achieved. These data again highlight the importance of using multiple poses: FARs with single pose configurations are three to four times greater. In addition, performance is modestly worse (approximately 1% FAR) in multi-pose systems when attackers have similar wrist sizes (less than 3mm different): the average FARs for the two-pose fist/open blender and the three-pose blender were 1.04% (SD=3.3) and 0.99% (SD=2.1), respectively. This indicates that, even if adversaries are able to accurately determine victim wrist circumference and recruit attackers based on this physical characteristic, the performance of multi-pose systems remains reasonably robust.

4.8 Authentication Model Generalizability Validation

To further validate the performance of WristAcoustic, we employed a widely used user identification scheme [8, 16, 17, 37]. The technique involves first assessing multi-class accuracy on a known set of users, then measuring authentication performance (FRR/FAR) with one-class classifiers using a "leave-one-subject-out" cross validation procedure. We trained the multi-class classifiers using samples from the first four donning sessions and used samples from the final donning session to assess accuracy. The multi-class user identification results, shown in Table 5, show single pose classifiers achieve accuracies of 95.47% or higher and that performance with multi-pose classifiers (configured as an ensemble blender) is near perfect. We then measured authentication error over 25 cross-validation rounds. In each round a single participant is left out as an unknown attacker: all data from this participant serves as the imposter test set (used to determine FARs). We then use the remaining 24 participants to train a multi-class classifier and, additionally, a one-class classifier unique to each participant. In this process, we again use the first four donning sessions for training and the final donning session as the genuine user test set (for determining FRRs). We then calculate performance by first submitting each test sample to the multi-class

Table 5. User identification and authentication accuracy.

Feature	Hand pose	User Identification	User Authentication		
		Accuracy (%)	FAR (%)	FRR (%)	HTER (%)
			Mean (SD)	Mean (SD)	Mean (SD)
MFCCs + PSD	Fist	95.47	4.07 (5.80)	4.06 (0.65)	4.06 (2.82)
	Open	95.73	8.70 (12.45)	8.45 (0.74)	8.57 (6.27)
	Relax	97.93	5.32 (6.70)	4.65 (0.53)	4.98 (3.28)
	Fist/Open	100.00	1.89 (6.04)	1.20 (0.17)	1.54 (3.03)
	Fist/Relax	100.00	1.33 (4.61)	1.27 (0.26)	1.30 (2.32)
	Open/Relax	99.27	3.22 (7.68)	3.47 (0.50)	3.34 (3.79)
	Fist/Open/Relax	100.00	0.59 (2.07)	0.40 (0.08)	0.50 (1.04)

classifier; the result of this process determines which one class classifier should be selected. The sample is then submitted to the selected one-class classifier. If the resulting score exceeds a given threshold—for this analysis, we computed an optimal one-class classification threshold value based on all 25 users—the sample is considered genuine. Table 5 shows the user authentication results, in terms of mean FARs, FRRs and HTERs, over all 25 cross-validation rounds. For single pose systems, these figures are somewhat elevated compared to the results reported from the binary classifiers described in Table 1: 4.06% to 8.57% here compared to 3.1%–3.81% with binary classifiers. However, WristAcoustic continues to achieve low error rates with the best-performing three pose blender (0.40% FRR and 0.59% FAR here), once again demonstrating the importance of using multiple poses. In addition, these results indicate that our feature sets are robust and perform generally well regardless of the classification and validation methods applied.

5 STUDY 2: MULTI-RECALL SESSION PERFORMANCE

To build on the results from our first study, we conducted a multi-session study to explore performance over time. We expected to record increased diversity in recorded data due to small variations that may accumulate over longer periods. In addition to characterizing the impact of changes, we also sought to establish the need for and extent to which retraining can accommodate the resultant signal variations and maintain high levels of authentication performance. The study protocols were reviewed and approved by the university's IRB.

5.1 User Study Methods

We recruited two groups of ten new participants: imposters and genuine users. For both groups, we collected data during an enrollment session that closely followed procedures in the single session study. Each participant completed five donning sessions and trials with each of the three hand poses (relaxed, fist, and open), as these showed a strong impact on authentication performance. However, we opted to study only a single body posture (seated), as this variable led to more minor variations (although training with both postures did lead to reliable overall performance), and dropping it enabled a substantial reduction in data collection time. In addition, the use of a single pose allowed us to stably record videos of all trials. To achieve this we set up a camera to capture video of the participant's arm in profile, providing a clear side view of the watch, forearm and hand gestures being performed. In this study, we again used the two-second white noise cue, captured 30 cue repetitions in each donning, and had a moderator help the participants during donnings to ensure the device was always buckled with the same tightness, and in approximately the same location on the wrist. Imposters completed their participation in the study after this initial session. Genuine users continued the study by attending a series of five separate recall sessions. We selected a session schedule based on data gathered from the first study. Of the participants who wore a watch regularly (62%), half indicated that they took it off mid-day at least once for exercise, personal hygiene, charging, or to sleep. Based on this pattern of activity we targeted semi-regular watch donning through the course of several days in our study design. Specifically, recall sessions were each

separated by a minimum of four hours, and a maximum of 72 hours (e.g., if study sessions fell over a weekend). Each recall session involved a single donning session including 30 cue repetitions for each hand pose. In total, we collected 900 samples for each genuine user and 450 for each imposter. Genuine users were compensated with the equivalent of 26 USD in local currency and imposters with 13 USD.

5.2 Demographics

Appendix B includes the demographics of our multi-recall session study participants. During recruitment, our primary goal was to assemble genuine user and imposter sets that were well matched rather than broadly representative of the wider population. This is because our first study robustly demonstrated our technique is effective for a relatively diverse group of participants. Furthermore, demographic variations between the relatively small genuine user and imposter groups in this study would weaken confidence in our results, as such differences might artificially inflate performance. Accordingly, to simplify the matching process, we opted to recruit only male participants. The average age of participants in the genuine user set was 25.9 years (SD=3.0), and nine were right-handed. Their average height, weight, and wrist circumference were 172.9 centimeters (SD=3.2), 70.5 kilograms (SD=6.3), and 15.7 centimeters (SD=0.9), respectively. The average age of the imposters was 25.6 (SD=2.0) and their average height, weight, and wrist circumference were 173.6 centimeters (SD=4.8), 74 kilograms (SD=7.5), and 16.2 centimeters (SD=0.9), respectively. There were no statistically significant differences in the heights, weights, and wrist sizes of the genuine and imposter user sets (two-sample t-test $p = 0.705$, 0.273, and 0.262, respectively).

5.3 Evaluation Setup

To evaluate multi-recall session performance, we trained three binary classifiers using the PSD and MFCC concatenated features that performed best in the first study for each genuine user: one for each hand pose. For each, we used all genuine user samples in the first four enrollment donnings (120 samples in total) and an *imposter train set* composed of 10 samples from each imposter (100 in total). All imposter samples were randomly selected from all poses in the first donning session and fixed for each user. One participant in the genuine user set showed low compliance with study instructions, executing hand poses forcefully despite instructions to the contrary. Accordingly, we excluded this participant's data in the accuracy evaluations.

For each genuine user, we measure per session recall FRRs using all samples from the final enrollment donning and all five recall sessions. We include data from the final enrollment donning to compare performance in this study with that attained in the first study. To measure FARs, we created a fixed *imposter test set* for each genuine user. This was composed of 30 samples from each other genuine user, randomly selecting these from all sessions and poses (240 samples in total). In this analysis, we create classifiers for each individual pose, and ensemble blenders for all possible pose combinations (three pairs and one triple).

We extend these results to evaluate model retraining effects by including data from recall sessions in the genuine user train set: we used between 1 and 3 recall sessions for retraining. To create retraining sets, we extended the initial 120 sample train sets with all 30 samples from the retraining recall sessions. For example, when using three recall sessions for retraining, we add 90 retraining samples to the original train set. During this process, each genuine user's imposter train sets and imposter test sets (for calculating FARs) are unchanged. We calculate FRR with the remaining recall sessions.

5.4 Single-session Results

To compare and validate our results in this study against those achieved in the first study, we measured the single-visit FRRs based on the fifth donning (enrollment) samples. FARs were measured with the imposter test set described above. The first row in Table 9 in Appendix E shows these results. All single-pose classifiers performed

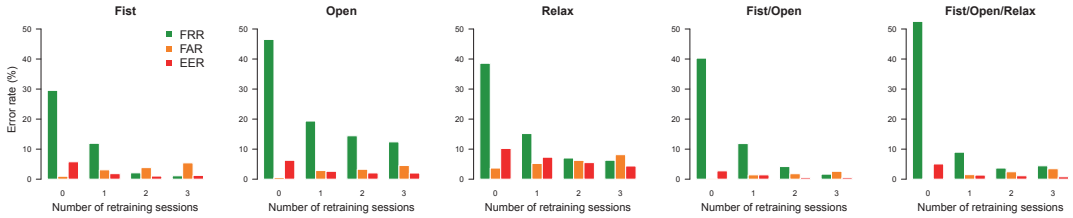

Fig. 4. Multi-session average FRR, FAR (threshold 0.5), and EER with respect to varying number of recall sessions included in the retrain set ("number of retraining sessions"). "0" indicates that classifiers trained only on enrollment data were used. Concatenated PSD and MFCC features were used for all evaluations.

as expected, achieving between 0.28–1.85% average HTER when both MFCC and PSD were used as a concatenated feature set. An ensemble blender that uses all three poses achieved peak performance with a mean HTER of 0%. These observations are consistent with the HTER results presented in Table 1.

5.5 Multi-session Results

We present multi-recall session results with respect to average FRRs, FARs, and EERs in Figure 4. Full details are in Table 9 in Appendix E. For all three poses and ensemble blenders, average FARs remain low while FRRs are markedly elevated (ranging between 29.56–52.52% across different hand poses and blenders). The resultant EERs 2.76%–10.23% are broadly in line with those in related work reporting on performance with a vibration-based wrist biometric and a 7th-day recall session [24]. We surmise these variations may be due to participants failure to perform the three hand poses consistently over time—more variability in hand poses may creep in after some hours or days. We explore the veracity of this somewhat surprising observation in more detail in Section 5.8.

5.6 Retraining Results

Figure 4 shows performance after retraining. Full details are again in Table 9 in Appendix E. The headline observation is that there is a steep downward trend in FRRs as the number of recall sessions contributing data to the retraining set increases. For example, the peak performing single pose fist classifiers exhibit a drop in FRRs from 29.56% without any retraining to 1.11% when three sessions are included. These improvements are also associated with a more modest, but still undesirable, increase in FARs—0.97% to 5.82% for fist. In addition, we note that, in general, ensemble blenders of multiple hand poses outperform individual poses. Peak performance, in terms of HTER, is 2.14%, attained with the fist/open blender after including data from three retraining sessions. These results confirm that enrollment data alone are insufficient to train robust classifiers, most likely due to variations in the way people perform hand poses over time. We argue that either a prolonged multi-session enrollment, or a longitudinal retraining protocol, will be a requirement for any practical bio-acoustic wrist authentication system.

One limitation of our evaluation method is that the size of the genuine test set shrinks as we increase the number of retraining sessions. In order to ensure this change is not impacting our FRR results, we conducted an additional analysis in which we fixed the recall sessions to the two sessions following (re-)training. For example, when not retraining, we used the first two recall sessions—those directly after enrollment. Similarly, when retraining with three recall sessions, we used the last two recall sessions—those directly afterwards. Figure 5 contrasts the retraining performance between those two evaluation methods for two classifier configurations (fist and the fist/open ensemble). Both methods show broadly similar trends, suggesting our retraining methods are not unduly impacting the results.

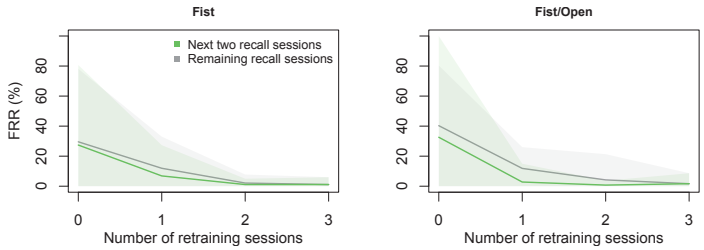

Fig. 5. Average FRRs measured based with two different methods: "grey" line represents FRRs measured with all remaining recall sessions; "green" line represents FRRs measured with the following two recall sessions after retraining. Left chart shows data using the fist classifier while right chart shows data from the fist/open blender.

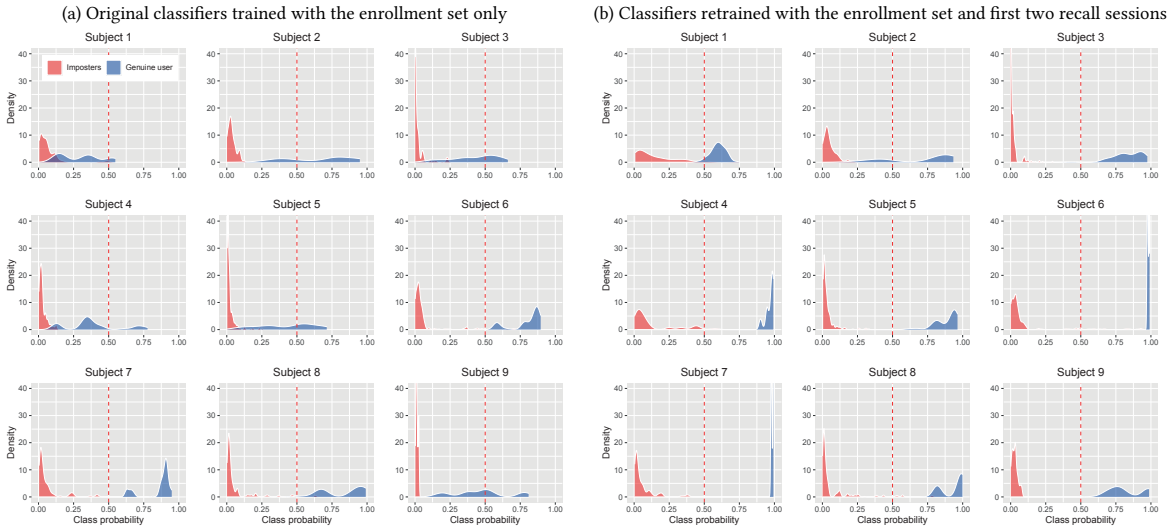

Fig. 6. Probability score distributions [42] of genuine and imposter test samples across all nine subjects using the fist/open blender. (a) shows scores using the enrollment classifiers, and (b) shows scores after including two retraining sessions.

To shed further light on the impact of retraining, we plot probability score distributions of each genuine user and their imposter test sets in Figure 6. We include a set of plots from training with the enrollment set alone (left) and a second set in which the first two recall sessions have been used for retraining (right). These charts clearly demonstrate the effectiveness of retraining: without retraining, user and imposter distributions are relatively proximate to each other and, at times, overlapped; there is also a wide spread of genuine user scores, suggesting a default 0.5 threshold will lead to poor performance. After retraining, the two distributions are well separated with genuine scores accumulating above 0.5.

5.7 Selecting Hand Poses

In both the first study and second study, single-pose classifiers built on data from the fist trials showed the best performance in terms of average HTERs, followed by the classifiers built on relax data, with those built on open data leading to the poorest results. Among the two-pose classifiers, blenders built from the combination of fist and open poses performed the best in both studies, achieving average HTERs of 0.67% ($k = 4$) and 2.14% (after

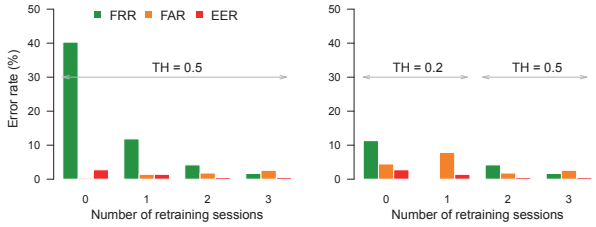

Fig. 7. The first graph shows average FRR/FAR (with respect to varying number of recall sessions in the retrain set) based on a fixed 0.5 threshold. The second graph shows the effects of starting with a 0.2 threshold, and increasing it to 0.5 after two rounds of retraining. Both charts use the fist/open blender.

retraining with data from three recall sessions), respectively. There were mixed results with respect to the three pose classifier: in the single session study, it led to peak performance (0.31% HTER) while in the multi-session it achieved 3.96% HTER, and was outperformed by the fist and open blender. This could be due to the increased diversity in the poses in the multi-session study manifesting as a greater overlap between the poses adopted for relax and fist trials—both poses involve flexing the fingers to a greater or lesser degree and over time, the boundaries between the two may have blurred, reducing the value of the data that can be captured from the combination of these poses. Evidence to support this assertion comes from the fact that while the relax classifiers generally outperformed the open classifiers in both studies, the combination of fist and relax was also inferior to fist and open in both studies. Compared to the highly distinct poses of fist and open, the data that can be captured from fist and relax is, at least partially, redundant.

5.8 Hand Pose Variability

Examination of the classifier score distributions in Figure 6 reveals participants one through five show a relatively broad spread of results with un-retrained classifiers (compared to participants six through nine) and well-separated scores after retraining. To examine the potential causes for these changes, we viewed video recordings for participants one through five. This revealed all five participants showed inconsistencies in performing the hand poses. This was particularly prominent for the relax pose, which all five participants performed with highly varying degrees of finger flexion (fingers more or less flexed), ultimately forming relax poses that were frequently similar to the fist and, less often, to the open poses. There was also variability in how forcefully the poses were enacted. Participants one through four, for example, produced the open pose with fingers both naturally and comfortably held and also, at other times, very widely splayed. These variations may help explain the superiority of the fist based classifiers over the other two poses—in general there was lower variability in how fist was enacted. In addition, the relatively high frequency with which relax resembled fist helps explain why blenders which combined these poses were less effective than expected.

5.9 Adapting Thresholds

A real world biometric authentication system needs to operate with a minimum of training. However, Figure 4 shows that FRRs with our system are high for the initial few donnings—before there is sufficient data to retrain models. However, EER results tell a different story: peak EERs without retraining were competitively low for the fist and open blender (2.76%) and even the fist-only classifier (5.78%). These results suggest that classifiers trained on the enrollment sets alone can effectively distinguish between users and imposters. The sample probability score distributions from the fist classifiers trained on the enrollment set shown in Figure 6(a) sheds light on how these distinctions are achieved. This figure suggests that although many genuine user authentication attempts result in scores beneath 0.5, they remain separable from imposter scores, which are generally clustered below 0.2.

Based on these observations, we suggest that lowering thresholds for the sessions immediately after enrollment would enable considerably improved FRRs at the cost of a small rise in FARs. Figure 6(b) shows probability score distributions after two retraining sessions: user and imposter scores are more widely separated, with imposter scores broadly unchanged in their distributions and user scores accumulating above 0.5. This demonstrates that the classifiers are becoming more effective at distinguishing between the two classes based on the addition of a more diversified set of genuine samples. This suggests that threshold values after several rounds of retraining could be incrementally raised to 0.5, a typical and well-balanced arrangement.

Figure 7 provides an example to illustrate the impact of this type of threshold adjustment—it assumes an initial post-enrollment threshold of 0.2 that is adjusted to 0.5 after two rounds of model retraining. With this setup, the FRRs for the fist/open blender immediately after enrollment are substantially improved over those reported in Figure 4. We also note that the adjusted FRR for the first recall session, at around 10%, is broadly comparable to the 11.5% real-world error rate reported for Android pattern lock [13]. This suggests that these elevated initial FRRs will not result in major disruptions to user experience: they are on par with many other forms of authentication technique.

5.10 Reducing Authentication Time and Enrollment Time

Reflecting our prior discussion of the hand pose variability and error rates over multiple recall sessions, we recommend a final system design based on the fist and open blender. However, setup and authentication times with such a system, if used in the configuration we deployed in our prior studies (see Section 5.3), remain prolonged. Enrollment will involve 30 repetitions of playback of two seconds of audio in each of two poses and through four donnings—480 seconds of audio playback in total. Authentication will involve two seconds of audio played in each pose, thus taking a minimum of four seconds. Such extended authentication and, particularly, enrollment times may represent a substantial burden to users.

Accordingly, in this section, we investigate the impact of reducing the overall enrollment and authentication times, via the mechanisms of reducing sample duration and repetition count, on authentication accuracy. We first examined this in terms of reducing the sample duration by clipping all recordings from the original two seconds to both one second and half a second in length. This implies reducing the authentication time with two poses to, respectively, two seconds and one second. Figure 8(a) shows how the error rates of the last two recall sessions vary with this change. The FRRs increase (from 1.67% to 4.63%) as the audio clips are shortened while the FARs fall slightly (from 2.62% to 2.39%). While we note that the increase in FRRs represents an additional burden to users in the case of a single failed authentication attempt, this is offset by the reduction in overall time spent authenticating over many attempts. For example, a user requires 400 seconds to complete 100 four second authentication attempts and just 200 seconds to complete 100 attempts that each take two seconds. The resulting uptick in FRRs (1.67% to 3.33%) is insufficient to impact the substantial aggregate usability improvement imparted by the shorter cues: repeating a pair of four second authentications would take 8 seconds, as would repeating four shorter two second authentications. As FARs remain steady with shorter cues, we argue there is considerable usability advantages to using briefer audio clips.

Building on this result, we sought to explore the impact of the number of repetitions used during enrollment on final recall session error rates. To do this we used clipped one second audio cues and explored reducing enrollment repetitions from 30 to ten and five. These configurations lead to enrollment times of 240 seconds, 80 seconds and 40 seconds (plus any time required to follow instructions). Figure 8(b) shows that FRRs modestly increase (from 3.33% to 4.81%) and FARs fall slightly (from 2.4% to 2.08%) as the number of repetitions decreases. Considering the sixfold reduction in enrollment times achieved for a modest increase in FRRs, and the previously demonstrated effectiveness of continuous retraining (suggesting performance will rapidly reach original levels

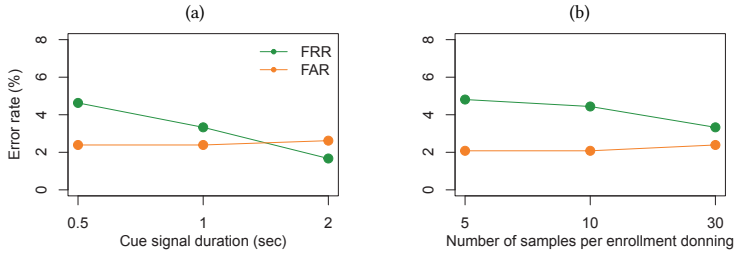

Fig. 8. Average FRRs and FARs (after three rounds of retraining) with respect to varying (a) cue signal duration and (b) number of samples collected during enrollment. Both charts use the fist/open blender.

as re-training data accumulates), we suggest the collection of five samples per pose in each of four enrollment donnings, taking just 40 seconds, will support reliable user authentication.

5.11 Model Training and Authentication Time Overheads

To determine the model training overheads, we measured the MFCC and PSD feature extraction time and SVM training time on a Linux machine equipped with Intel i7-7700 CPU (3.60G Hz) using the `sklearn` libraries in Python. Concatenated MFCC and PSD consists of 414 features. Based on our final recommendation (1 second cue signal duration per pose, and 4 enrol donnings with 5 samples per pose), we measured the feature extraction time and training time across the 10 genuine users. We used 20 genuine samples and 100 imposter samples to train fist and open ensemble classifiers. On average, it took 882 milliseconds (SD=50) to extract the full feature set, and took 60 milliseconds (SD=4.6) to train the two classifiers. The total size of the extracted feature sets was 448 Kilobytes, and the total model size was 193 Kilobytes (SD=16.4). The prediction (authentication) time was also fast, taking only 44 milliseconds (SD=1.5) on average. Although we evaluated these overheads on a PC, considering the small training set size and model size, we expect training and loading classifiers directly on smartwatches should be feasible.

6 STUDY 3: ENROLLMENT AND AUTHENTICATION USABILITY

To assess the usability of our authentication system, we conducted a study using the optimal system configurations identified in Section 5.10. Specifically, we used the combination of fist and open hand poses, one second audio cues and enrollment sessions involving four donnings, each featuring five cue repetitions. The goal of this study was to assess the subjective experience of using WristAcoustic.

6.1 Methods

The study took place in a quiet office environment. Participants were provided with detailed instructions and were encouraged to ask questions to the moderator. They also had the opportunity to practice both enrollment and authentication sessions for a few minutes. Participants then experienced the full enrollment process, and then immediately performed two authentications, removing and re-donning the watch between each session. Enrollment was moderated by a researcher who aided participants donning the prototype, controlled playback of the audio cues and instructed participants to perform the fist and open poses at the appropriate times. After enrollment the participants completed a short survey consisting of the system usability scale (SUS) [1] and the unweighted NASA Task Load Index (TLX) [15] questions. In addition, they were asked to rate each hand pose on the the Borg-10 perceived exertion scale [48].

After completing this survey, the participants completed two authentication sessions, taking the watch off and re-donning it prior to each one. To deliver a fully automatic authentication process that did not require

moderator invention, short audio cues (150 millisecond beeps) were used to mark pose changes: an initial beep signified the start of the authentication session. Participants then adopted the fist pose and held it while one second of white noise was played through the wrist. After a second beep, they moved to the open pose and held that while another second of white noise was played. Data from both poses was collected and the moderator observed participant performance live. If a participant was judged to have failed to produce the correct pose at the designated time, they were requested to continue to complete authentication sessions until a total of two correct sessions were recorded. We logged the number of pose production failures as measure of task difficulty. After the two correctly completed sessions, participants were provided with a final survey assessing levels of SUS, TLX and Borg-10 experienced during authentication sessions.

6.2 Demographics

We recruited a new group of 15 participants. Seven were female and eight male. The average age was 26.2 (SD=5). The average height and wrist circumference were 168.1 centimeters (SD=7.5) and 15.5 centimeters (SD=1.6), respectively. Full details can be found in Appendix B.

6.3 Authentication Accuracy

In order to validate the recommendations made in Section 5.10, we first evaluated authentication performance. To train the fist and open binary classifiers for each genuine user, we used all genuine user samples collected from the four enrollment donnings (20 per pose), and selected ten samples from each of the ten imposters (100 in total) that we recruited during the Study 2. For each genuine user, we measured FRRs using the four samples collected through the two authentication tests. To measure FARs, we created a fixed imposter test set for a given genuine user, selecting all four samples from all other genuine users. We trained a separate classifier for each pose, and used them together as an ensemble blender to predict the final probability (authentication) scores. Using this setup, we recorded a single authentication failure across all 15 genuine users (30 tests), corresponding to an FRR of 3.33% and in line with our expectations for this system configuration. The average FAR was 1.90% (SD=3.7).

6.4 Usability Results

There were two failures to perform the correct hand poses in the correct sequence. Two participants each failed during the their first authentication attempt. As such, the failure rate for this task of was 6.25%. This is likely due to the novelty of the task. The fact there were no failures in the second authentication sessions suggests that learning times for this task may be short. In terms of the subjective measures, mean Borg CR10 scores during enrollment were 3.1 (SD=1.9) and 3.3 (SD=2.4) for fist and open poses respectively. This indicates a moderate level of exertion and likely reflects the somewhat prolonged nature of the task, involving wearing and re-wearing the watch four times. Supporting this explanation, mean Borg CR10 scores for the shorter authentication sessions were 2.1 (SD=2.0) and 2.3 (SD=2.4) for fist and open, levels corresponding with slight exertion. Mean SUS scores for enrollment and authentication were 76.9 (SD=11.2) and 76.4 (SD=13.7), indicating that the participants generally felt that WristAcoustic—configured with the fist and open blender and based on a total of two seconds of cue signal playback duration—achieves "good" usability [1]. In addition, the unweighted overall workload scores from the NASA TLX for enrollment and authentication sessions were 3.76 (SD=2.6) and 3.42 (SD=3.7), respectively, indicating that the participants experienced low levels of mental and physical workload, interpreted according to Grier [11]'s analysis of 200 studies deploying TLX in a wide range of tasks. Indeed, these workload scores are lower than the mean values of approximately 6 reported for entering 4-digit PIN items on smartwatches [33]. Considering that these data prior are related to PIN item entry usability—i.e., the workload costs of PIN setup and memorability are not included—the low scores we record in this work indicate there may be usability advantages to our approach over traditional PINs. Taken together, these results suggest that participants' experience with

WristAcoustic was positive: although enrollment involved a moderate level of exertion, actual authentication processes were perceived as requiring only slight effort, and measures of usability and workload were good throughout. This study suggests that users will face few usability barriers to setting up and authenticating with WristAcoustic.

7 STUDY 4: POSTURE VARIATIONS AND NOISE

While the previous studies in this paper have demonstrated the performance of WristAcoustic with a variety of analytic approaches, over a sustained period of time and in a realistic, more usable, configuration, they all took place in a relatively controlled setting. In the real world, users can be expected to don their watches and authenticate in somewhat more diverse environments and situations—in various body poses and in the presence of various forms of distraction and interference. To address the impact of these issues on the performance of WristAcoustic we conducted a study in which participants adopted different typical body poses and were exposed to different forms of acoustical and motion/vibration interference during recall sessions.

7.1 Methods

We applied the procedures from the usability study to collect enrolment data. After completing enrolment, participants put the watch back on and completed twenty-two recall sessions, also following the procedures in the usability study. These sessions were organized into two examples of each of 11 different pose or noise conditions. Six of these explored arm poses - this was embodied by asking participants to hold their arms near horizontal (approximately 20 degrees), raised (approximately 45 degrees) and near vertical (approximately 70 degrees) with their elbows resting both on and off a desk in front of them. These variations represent a supported or at-rest forearm versus one suspended in free space at a quite wide spectrum of comfortable orientations. They embody typical poses a user might hold their arm in during authentication. Three further conditions explored audio noise: participants adopted the default pose (arm raised and resting on the desk) while samples of music, speech (in the form of a news clip) and ambient noise (recorded from a busy coffee shop) were played through a standard speaker at between 60 and 70 dB. The final two conditions also used the default pose and involved physical disturbance: in one, participants sat on a massage chair (a surrogate for various situations, e.g., transportation, in which a person may be regularly jostled) as it operated while in the second they held a vibrating mobile phone in their right hand (the one not wearing the watch prototype) in comfortable proximity to the watch (approximately 20 cm away). Due to the requirement to precisely perform hand poses during authentication, it is not possible for participants to simultaneously hold a phone in their watch hand. This set of situations was selected to represent common everyday scenarios in which people might authenticate with WristAcoustic.

7.2 Demographics

We recruited a new group of ten participants. Four were female and six male. The average age was 24.7 (SD=3.8). The average height and wrist circumference were 168.9 centimeters (SD=7.9) and 15.4 centimeters (SD=1.1), respectively. Full details can be found in Appendix B.

7.3 Results

We applied the usability study evaluation methodology (see Section 6.3) to determine performance across the twenty-two authentication trials. Arm pose variations led to minimal changes in FRR: we observed just a single failure in the study. This suggests that WristAcoustic is robust to a wide range of arm poses and postures and will continue to offer high authentication performance regardless of how users choose to hold their upper bodies. In addition, we recorded no authentication failures in the presence of audio noise; this result is expected, as the microphones used in the system are surface transducers that pick up minimal air transmitted sound. We

showed similar strong performance in the motion and vibration conditions: no authentication failures. This strong performance is likely because motions delivered by the massage chair are very low frequency, that vibrations from the buzzing phone attenuate strongly during their passage through the body, and that relatively low amplitude in-air signals generated by phone's vibration are, again, simply not picked up by the surface transducers. In addition to this strong performance in terms of FRRs, we recorded a mean FAR of 1.67% (SD=1.9) across the whole study, a figure comparable to that attained in the usability study. Taken together, these results suggest WristAcoustic will perform well in the kinds of diverse situations that users may encounter during real world authentication attempts.

8 DISCUSSION

8.1 Performing Multiple Poses over Time

Single-session study results (see Table 8 and Table 9 in Appendix E) show promising levels of performance for both classifiers built on both individual and combined hand poses: mean HTERs approach zero in our first two studies. However, our multi-session study presents a starkly different profile of highly elevated FRRs when users seek to authenticate after a break. Analysis of study videos suggests these results are likely due to inconsistencies in the way participants perform hand poses between multiple recall sessions. Participants may use more or less finger flexion, more or less force, or present other variations such as in wrist extension/flexion. These results suggest that individuals are poorly equipped to perform functionally identical hand poses over time—although they exhibit a clear conceptual understanding of the basic poses, they interpret and express these instructions variously.

This finding leads to a number of design implications and suggestions for future avenues of investigation. First and foremost, the intuition that requiring more hand poses will increase performance by sampling more unique data points [46] may not hold up to scrutiny. In our multi-session study, a blender on all three poses led to reduced performance (3.96% HTER with 4.44% FRR) compared to that achieved with the classifier based on the fist alone (3.26% HTER with 1.11% FRR). This may be because sampling more poses involves increased variability compared to that achievable with a single (or small number of) well-defined poses. As such, we suggest poses should be carefully selected to be as simple and unambiguous as possible—in our studies, fist is likely the best example here. Future systems in this area should minimize the number of poses from which they get sample data: based on the multi-session recall results, our recommendation is to use the fist and open poses. In addition, our usability evaluation showed that the enrollment and authentication UX involving those two poses is generally easy to learn and use, and entails low levels of workload and exertion. Finally, we note clear guidance and instructions will be necessary to ensure the consistent performance of hand poses.

8.2 Retraining Recommendations

We demonstrated that error rates can be stabilized over time by retraining classifiers. For example, fist classifiers led to an average HTER of 15.26% after enrollment but improved substantially with retraining, ultimately reaching 3.26% HTER with three retraining sessions. Blenders showed similar trends, with the best performing combination of fist and open showing an improvement from 20.32% HTER to 2.14% after retraining. We believe that retraining will be a necessary part of any future system in this space in order to accommodate variations in pose enactment and other contextual changes that go beyond the current scope of our studies.

That said, retraining was not unreservedly beneficial. In addition to achieving greatly improved FRRs, it was also associated with a more modest—and undesirable—increase in FARs. While we can observe this trend, our current data set offers few insights into how it might play out over longer periods. However, we can speculate about the potential causes: during retraining, we continuously expand the genuine user test set while we maintain a fixed imposter train set; this likely results in an increase in genuine user classification boundaries while imposter

boundaries shrink. Accordingly, we recommend that future work update the imposter train sets in tandem with the genuine user train set by selecting new and unique imposters to diversify the set's overall coverage.

8.3 Evaluation Metrics: EER vs FAR/FRR

Existing work on biometric authentication [7, 24, 25, 37, 44, 46] typically presents EERs as the sole means to measure authentication accuracy. Results in this paper, however, suggest that this approach may not be sufficient and may, in fact, lead to misleading characterizations of performance [34, 42]. For example, average EERs for our multi-recall study are low without retraining and improve markedly after retraining: with the fist and open blender, they reduce from 2.76% to 0.51%. However, the corresponding FRR results, show a dramatically different trend with a probability threshold of 0.5. For the fist and open blender, a high figure of 40.3% without retraining reduces to a reasonable 1.67% after retraining. This suggests that, despite low EERs, users may encounter usability issues in the form of high rejection rates during initial use.

There is a simple explanation for the differences between EERs and FRRs. EER figures represent optimal probability thresholds for each user, while FRRs are calculated using a fixed threshold (e.g., 0.5) for all users. Placing the decision point at an optimal per-user location almost inevitably leads to improved performance, as it can accommodate a wide range of genuine user probability score distributions. Figure 6 illustrates this point well: subjects 1,3,4,5, and 9, for instance, show score distributions that spread well below the 0.5 threshold while still remaining distinct from the scores returned by imposters. These results beg a question: which type of threshold is valid? The answer may be dependent on context. We argue that it is infeasible to determine optimal thresholds in real-world systems and deployments—it is unclear what sampling strategies could be used to derive and validate them. As such, a more realistic approach may be to deploy the methods outlined in this paper: using preset thresholds that adapt according to a fixed schedule as more genuine user data is accumulated. For example, in the multi-session study reported in this paper, using a threshold that changes from 0.2 to 0.5 as more retraining sessions become available achieves a competitive level of performance over all sessions. In general, we recommend that future work that seeks to understand the security and usability implications of biometric classifiers focuses on (1) how probability score distributions change over time, (2) emphasizes the use of fixed thresholds, and (3) explores how these impact unbiased FARs and FRRs. In many cases, over reliance on EERs, and the optimal thresholds they rely on, will lead to reporting results that tell incomplete, impractical and possibly misleading stories.

8.4 Limitations

There are a number of limitations to our work. We summarize these in the sections below.

8.4.1 Sample Limitations. While sample sizes in our first two studies (N=25 and N=20) exceed those in much prior work in this area [24, 37, 46], more substantial studies would increase confidence in the results we report. Our second study participants were intentionally screened to be male, a choice that facilitated creating matched groups of imposters and genuine users. While we believe that future studies should target broader demographics, we also note that analysis in our first study (100 permutation evaluations) showed no statistical differences in performance between male and female participants. Thus, we believe our screening procedures in the second study do not threaten the validity of the results. Another limitation of our sample is that almost all participants were Asian. As such, the performance of WristAcoustic on other ethnic groups remains unknown, and needs to be investigated as a part of future work.

8.4.2 Multi-session Limitations. The size of our current studies also presents limits to our classifier evaluation procedures and the extent to which we can make practical, actionable recommendations about likely real world performance. Perhaps most critically, in our current study when we evaluate retraining, the size of the test sets

we use changes—as we retrain with more sessions, we have fewer sessions that can be used for testing. Smaller test sets may be associated with improvements in classifier performance on key metrics such as FRR. The best solution to this problem would be collection of substantially extended data sets (e.g., with a large number of sessions captured over several weeks or months) that can support a robust analysis of this issue. We present an initial exploration of how this could be achieved by fixing the size of the test sets (two sessions) used to evaluate all trained and retrained classifiers in Figure 5. These results show similar trends of decreasing FRRs to those in the main analysis, suggesting that changes in test set size may not be unduly impacting the results we report. Beyond this issue, our current data is insufficient to support reliable recommendations about a range of practical issues, such as the number of retraining sessions that will be required to achieve robust long-term model performance or the kinds of sampling strategy (for both genuine user and imposter data) that are optimal. Only by collecting extended data sets, featuring large numbers of recall sessions, will we be able to explore these issues in the detail they deserve.

8.4.3 Procedural Limitations. There are also limitations to our final two studies. In our usability study, a moderator assisted participants putting on and taking off the watch and, during enrolment, also cued production of the appropriate hand pose at the appropriate time. The usability results for enrollment may have been impacted by this in-person support and future studies of this system should re-examine usability after creating complete and robust prototypes that do not require live moderator intervention. In addition, while our study assessed a common set of usability measures, it cannot be considered a complete and thorough evaluation of this multi-faceted construct. For example, we did not assess the learnability or memorability [32] of using WristAcoustic. Longer term studies on more mature prototypes may be able to present a more rounded and complete evaluation in the future. In addition, in our final study, we explore body poses and noise conditions. This work could also be extended. While our current (tethered) prototype did not allow us to explore more extreme body posture variations, such as lying down, they should be examined in the future. In addition, while we examined some forms of vibration, future studies on standalone prototypes should go further: it would be interesting to examine system performance while, for example, participants are walking or riding public transport. Extended studies of real world usability and robustness to various environmental conditions would do much to complement data and results reported in the current paper.

8.4.4 Prototype Limitations. Finally, it is also worth discussing the limitations inherent to our prototype device. We developed a bespoke system capable of generating and recording through-wrist audio signals over a wide frequency range. To achieve this, we selected relatively high-end actuators and sensors that do not currently appear in smart watches. Integrating such components into practical, real world devices would require that they add substantial value to use cases including, but also beyond, secure authentication. This may be realistic: the proliferation and rapid development of ear wearables is resulting in major advances in various forms of novel bio-acoustic sensor, such as voice pickup units [54]. This is increasing the performance and driving down the cost, size and power consumption of such devices. In addition, advanced vibration actuators are already a standard and ever evolving part of smartwatches and the list of applications to which wrist bio-acoustics have been employed for is growing: hand gesture recognition [18], the recognition of grasped objects [23], the localization of on-body touches [49]. However, beyond this speculative future, we also note that our use of relatively high end components is also a useful starting point for developing more realistic versions of our system. If the data and results we report represent something approximating peak performance, then it will be interesting and valuable to determine whether performance degrades or is maintained on more realistic, and currently deployed, hardware. A specific point of interest here would be to compare performance directly with a system configured akin to that described in Lee et al. [24]. In this paper, built-in watch vibration motors and accelerometers are used to generate and record vibrations used for authenticating users. While same session performance is modestly reduced (to 1.37% EER) from that reported in this paper (as low as 0.01% EER in our first study), there is limited

data on multiple sessions and how the system would respond to varying hand poses. Given the high variability we observe in response to these factors, we identify a need for future studies to replicate the extended (multi-session) study protocols proposed in this paper. An important topic for future work on WristAcoustic would be to test its performance on existing consumer devices, and assess whether the findings reported in this paper can be maintained with lower-end built-in actuators and sensors.

9 CONCLUSION

We studied the feasibility of authenticating smartwatch users based on acoustic responses measured through the wrist. By training separate classifiers for three hand poses (relax, fist, and open), and using them as an ensemble blender, we achieved an average EER of 0.01% in a single-session study (N=25), significantly outperforming prior on-wrist biometric authentication systems. Our multi-recall session study (N=20), however, demonstrated variability in the way people perform hand poses over time, leading to significant elevations in FRRs. The relax pose, in particular, was highly problematic: many participants interpreted it as either a loose fist or relatively open palm. Hence, the blenders that used relax classifiers (which offered diverse and/or redundant information) were less effective. Consequently, the fist/open blender (two distinctive poses) demonstrated the most consistent performance over multiple sessions: an average EER of 2.76%. In addition, we show that it is necessary to periodically retrain the classifiers to stabilize error rates for long-term use. For instance, the average FRR for the fist/open blender dropped rapidly from 40.3% (without retraining) to 1.67% after including samples from the first few recall sessions and updating the classifiers. Furthermore, our usability evaluation demonstrates that people generally find the overall enrollment and authentication process requiring those two poses easy to learn and use, and that enacting them results in low levels of mental and physical workload. In addition, our final study demonstrates that WristAcoustic performs robustly in a wide variety of typical situations and environments, demonstrating strong performance in the kinds of situations users might experience during everyday real-world use.

ACKNOWLEDGMENTS

This work was supported by Samsung Research and a Korea Institute for Advancement of Technology (KIAT) grant funded by the Korean Government (MOTIE) (P0012725, The Competency Development Program for Industry Specialist).

REFERENCES

[1] 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.

[2] Joseph Bonneau. 2012. The Science of Guessing: Analyzing an Anonymized Corpus of 70 Million Passwords. In *Proceedings of IEEE Symposium on Security and Privacy (S&P)*. IEEE Computer Society, USA, 538–552.

[3] Joseph Bonneau, Sören Preibusch, and Ross J. Anderson. 2012. A Birthday Present Every Eleven Wallets? The Security of Customer-Chosen Banking PINs. In *Proceedings of International Conference on Financial Cryptography and Data Security (FC)*. Springer Berlin Heidelberg, Berlin, Heidelberg, 538–552.

[4] Peter J Brockwell and Richard A Davis. 1991. *Time Series: Theory and Methods*. Springer-Verlag, New York, NY, USA.

[5] Seunghun Cha, Sungsu Kwag, Hyoungshick Kim, and Jun Ho Huh. 2017. Boosting the Guessing Attack Performance on Android Lock Patterns with Smudge Attacks. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security* (Abu Dhabi, United Arab Emirates) *(ASIA CCS '17)*. Association for Computing Machinery, New York, NY, USA, 313–326. https://doi.org/10.1145/3052973.3052989

[6] Geumhwan Cho, Jun Ho Huh, Junsung Cho, Seongyeol Oh, Youngbae Song, and Hyoungshick Kim. 2017. SysPal: System-Guided Pattern Locks for Android. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, USA, 338–356. https://doi.org/10.1109/SP.2017.61

[7] Cory Cornelius, Ronald Peterson, Joseph Skinner, Ryan Halter, and David Kotz. 2014. A Wearable System That Knows Who Wears It. In *Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services* (Bretton Woods, New Hampshire,

USA) *(MobiSys '14)*. Association for Computing Machinery, New York, NY, USA, 55–67. https://doi.org/10.1145/2594368.2594369

[8] Rig Das, Emanuela Piciucco, Emanuele Maiorana, and Patrizio Campisi. 2018. Convolutional neural network for finger-vein-based biometric identification. *IEEE Transactions on Information Forensics and Security* 14, 2 (2018), 360–373.

[9] 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.

[10] 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.

[11] 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

[12] 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 (J Voice)* 36, 2 (2020), 156–169.

[13] Marian Harbach, Alexander De Luca, and Serge Egelman. 2016. The Anatomy of Smartphone Unlocking: A Field Study of Android Lock Screens. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4806–4817. https://doi.org/10.1145/2858036.2858267

[14] Chris Harrison, Desney Tan, and Dan Morris. 2010. Skinput: Appropriating the Body as an Input Surface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 453–462. https://doi.org/10.1145/1753326.1753394

[15] 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.

[16] Christian Holz, Senaka Buthpitiya, and Marius Knaust. 2015. Bodyprint: Biometric User Identification on Mobile Devices Using the Capacitive Touchscreen to Scan Body Parts. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3011–3014. https://doi.org/10.1145/2702123.2702518

[17] Anna Huang, Dong Wang, Run Zhao, and Qian Zhang. 2019. Au-id: Automatic user identification and authentication through the motions captured from sequential human activities using rfid. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 3, 2 (2019), 1–26.

[18] Yasha Iravantchi, Yang Zhang, Evi Bernitsas, Mayank Goel, and Chris Harrison. 2019. Interferi: Gesture Sensing Using On-Body Acoustic Interferometry. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300506

[19] Hanvit Kim, Haena Kim, Se Young Chun, Jae-Hwan Kang, Ian Oakley, Youryang Lee, Jun Oh Ryu, Min Joon Kim, In Kyu Park, Hyuck Ki Hong, et al. 2018. A wearable wrist band-type system for multimodal biometrics integrated with multispectral skin photomatrix and electrocardiogram sensors. *Sensors* 18, 8 (2018), 2738.

[20] Konstantin Klamka, Tom Horak, and Raimund Dachselt. 2020. Watch+Strap: Extending Smartwatches with Interactive StrapDisplays. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376199

[21] Masoud Mehrabi Koushki, Borke Obada-Obieh, Jun Ho Huh, and Konstantin Beznosov. 2021. On Smartphone Users' Difficulty with Understanding Implicit Authentication. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 690, 14 pages. https://doi.org/10.1145/3411764.3445386

[22] Amioy Kumar, Tanvir Singh Mundra, and Ajay Kumar. 2009. *Anatomy of Hand*. Springer, Boston, MA, 28–35. https://doi.org/10.1007/978-0-387-73003-5_267

[23] Gierad Laput, Robert Xiao, and Chris Harrison. 2016. ViBand: High-Fidelity Bio-Acoustic Sensing Using Commodity Smartwatch Accelerometers. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 321–333. https://doi.org/10.1145/2984511.2984582

[24] Sunwoo Lee, Wonsuk Choi, and Dong Hoon Lee. 2021. Usable User Authentication on a Smartwatch Using Vibration. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security* (Virtual Event, Republic of Korea) *(CCS '21)*. Association for Computing Machinery, New York, NY, USA, 304–319. https://doi.org/10.1145/3460120.3484553

[25] Jingjie Li, Kassem Fawaz, and Younghyun Kim. 2019. Velody: Nonlinear Vibration Challenge-Response for Resilient User Authentication. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security* (London, United Kingdom) *(CCS '19)*. Association for Computing Machinery, New York, NY, USA, 1201–1213. https://doi.org/10.1145/3319535.3354242

[26] Jian Liu, Chen Wang, Yingying Chen, and Nitesh Saxena. 2017. VibWrite: Towards Finger-Input Authentication on Ubiquitous Surfaces via Physical Vibration. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (Dallas, Texas, USA) *(CCS '17)*. Association for Computing Machinery, New York, NY, USA, 73–87. https://doi.org/10.1145/3133956.3133964

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

[28] Philipp Markert, Daniel V. Bailey, Maximilian Golla, Markus Dürmuth, and Adam J. Aviv. 2020. This PIN Can Be Easily Guessed: Analyzing the Security of Smartphone Unlock PINs. In *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, USA, 286–303. https://doi.org/10.1109/SP40000.2020.00100

[29] Jess McIntosh, Asier Marzo, Mike Fraser, and Carol Phillips. 2017. EchoFlex: Hand Gesture Recognition Using Ultrasound Imaging. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 1923–1934. https://doi.org/10.1145/3025453.3025807

[30] Collins W. Munyendo, Miles Grant, Philipp Markert, Timothy J. Forman, and Adam J. Aviv. 2021. Using a Blocklist to Improve the Security of User Selection of Android Patterns. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*. USENIX Association, Santa Clara, CA, 37–56. https://www.usenix.org/conference/soups2021/presentation/munyendo

[31] Toan Nguyen and Nasir Memon. 2017. Smartwatches Locking Methods: A Comparative Study. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*. USENIX Association, Santa Clara, CA, 5 pages. https://www.usenix.org/conference/soups2017/workshop-program/way2017/nguyen

[32] Jakob Nielsen. 1993. *Usability Engineering*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[33] 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 on Asia Conference on Computer and Communications Security* (Incheon, Republic of Korea) *(ASIACCS '18)*. Association for Computing Machinery, New York, NY, USA, 75–87. https://doi.org/10.1145/3196494.3196555

[34] John Oglesby. 1995. What's in a number? Moving beyond the equal error rate. *Speech Communication* 17, 1–2 (1995), 193–208.

[35] 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.

[36] Léa Saviot, Frederik Brudy, and Steven Houben. 2017. WRISTBAND.IO: Expanding Input and Output Spaces of a Smartwatch. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 2025–2033. https://doi.org/10.1145/3027063.3053132

[37] 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* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 1379–1384. https://doi.org/10.1145/2858036.2858152

[38] Muhammad Shahzad and Munindar P. Singh. 2017. Continuous Authentication and Authorization for the Internet of Things. *IEEE Internet Computing* 21, 2 (2017), 86–90.

[39] Katie A. Siek, Yvonne Rogers, and Kay H. Connelly. 2005. Fat Finger Worries: How Older and Younger Users Physically Interact with PDAs. In *Proceedings of the 2005 IFIP TC13 International Conference on Human-Computer Interaction* (Rome, Italy) *(INTERACT'05)*. Springer-Verlag, Berlin, Heidelberg, 267–280. https://doi.org/10.1007/11555261_24

[40] Haim Sohmer, Sharon Freeman, Miriam Geal-Dor, Cahtia Adelman, and Igal Savion. 2000. Bone conduction experiments in humans–a fluid pathway from bone to ear. *Hearing Research* 146, 1–2 (2000), 81–88.

[41] Stefan Stenfelt. 2013. Skull vibration during bone conduction hearing. *20th International Congress on Sound and Vibration 2013, ICSV 2013* 2 (01 2013), 1394–1401.

[42] Shridatt Sugrim, Can Liu, Meghan McLean, and Janne Lindqvist. 2019. Robust Performance Metrics for Authentication Systems. In *Network and Distributed Systems Security (NDSS)*. The Internet Society, USA, 15 pages. https://doi.org/10.14722/ndss.2019.23351

[43] Sebastian Uellenbeck, Markus Dürmuth, Christopher Wolf, and Thorsten Holz. 2013. Quantifying the Security of Graphical Passwords: The Case of Android Unlock Patterns. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer &; Communications Security* (Berlin, Germany) *(CCS '13)*. Association for Computing Machinery, New York, NY, USA, 161–172. https://doi.org/10.1145/2508859.2516700

[44] Wei WANG, Lin Yang, and Qian Zhang. 2018. Resonance-Based Secure Pairing for Wearables. *IEEE Transactions on Mobile Computing* 17, 11 (2018), 2607–2618.

[45] 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.

[46] Hiroki Watanabe, Hiroaki Kakizawa, and Masanori Sugimoto. 2021. User Authentication Method Using Active Acoustic Sensing. *Journal of Information Processing* 29 (2021), 370–379.

[47] Peter Welch. 1967. The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics* 15, 2 (1967), 70–73.
[48] Nerys Williams. 2017. The Borg Rating of Perceived Exertion (RPE) scale. *Occupational Medicine* 67, 5 (2017), 404–405.
[49] Cheng Zhang, AbdelKareem Bedri, Gabriel Reyes, Bailey Bercik, Omer T. Inan, Thad E. Starner, and Gregory D. Abowd. 2016. TapSkin: Recognizing On-Skin Input for Smartwatches. In *Proceedings of the 2016 ACM International Conference on Interactive Surfaces and Spaces* (Niagara Falls, Ontario, Canada) *(ISS '16)*. Association for Computing Machinery, New York, NY, USA, 13–22. https://doi.org/10.1145/2992154.2992187
[50] 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.
[51] 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* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3173574.3174011
[52] Yang Zhang and Chris Harrison. 2015. Tomo: Wearable, Low-Cost Electrical Impedance Tomography for Hand Gesture Recognition. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (Charlotte, NC, USA) *(UIST '15)*. Association for Computing Machinery, New York, NY, USA, 167–173. https://doi.org/10.1145/2807442.2807480
[53] Tianming Zhao, Yan Wang, Jian Liu, Yingying Chen, Jerry Cheng, and Jiadi Yu. 2020. TrueHeart: Continuous Authentication on Wrist-worn Wearables Using PPG-based Biometrics. In *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*. IEEE Computer Society, USA, 30–39. https://doi.org/10.1109/INFOCOM41043.2020.9155526
[54] Yi Zhou, Yufan Chen, Yongbao Ma, and Hongqing Liu. 2020. A Real-Time Dual-Microphone Speech Enhancement Algorithm Assisted by Bone Conduction Sensor. *Sensors* 20, 18 (2020), 18 pages. https://doi.org/10.3390/s20185050

A PRE-PROCESSING RESPONSE SIGNALS

Response signal alignments are often performed using cross-correlation between response signals and stimulus signals [25, 26]; however, this approach is not applicable in our case since our transmitted signal involves white noise. Thus, we tested $H_0 : x(t) \sim \text{Gaussian}(0, \sigma_0^2)$ versus $H_1 : x(t) \sim \text{Gaussian}(0, \sigma_1^2)$ with $\sigma_1 \gg \sigma_0$ in order to extract the response signals exactly. The test statistic is given as $F = S_1^2/S_0^2$, where S_0 is the standard deviation computed from the data without application of the cue stimuli, and S_1 is the standard deviation of data at each time frame (e.g., 2 milliseconds). We detected the start and end of the response signal by finding the time frames where $F > C\chi^2_{0.95,m-1}/(m-1)$ with the number m of data points in each time frame and the 95th-percentile of the Chi-square distribution with $m-1$ degrees of freedom. We used $C = 10$ as σ_1 is at least 3 times larger than σ_0.

B STUDY DEMOGRAPHICS

Table 6 shows the demographics of study participants recruited for three studies we performed. In multi-recall session study, 10 participants were asked to come back for another five recall sessions after enrollment. These participants were used as the genuine set in the evaluation. The other 10 participants completed enrollment only; we used their data as the imposter train set.

C FIRST STUDY ACCURACY RESULTS

Tables 7 and 8 show the first study accuracy results for all feature sets, including PSD, transfer function, MFCCs, and the top performing PSD/MFCC concatenation feature. The average EER, FAR, and FRR of per-pose classifiers are presented in Table 7; those three accuracy reports for the multi-pose blenders are presented in Table 8.

D PHYSICAL CHARACTERISTICS OF USERS AND AUTHENTICATION ACCURACY

Figure 9 shows how the first study authentication accuracy changes by physical characteristics of genuine users. For this analysis, we picked the fist-based classifiers trained with the PSD features and $k = 4$, and HTER as the authentication accuracy metric. In each of the 100 permutation rounds, we grouped the 15 genuine users based on their "gender," "age," "wrist size," "weight," and "height." We performed two-sample t-tests to compare the HTER

Table 6. Demographics of study participants.

		Study 1 (N=25)	Study 2 (N=20)		Study 3 (N=15)	Study 4 (N=10)
		Single-session	Multi-recall session		Usability	Performance robustness
			Genuine users	Imposters		
Ethnicity	White	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (10%)
	Asian	25 (100%)	10 (100%)	10 (100%)	15 (100%)	9 (90%)
	Black	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Others	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Gender	Female	14 (56%)	0 (0%)	0 (0%)	7 (46%)	4 (40%)
	Male	11 (44%)	10 (100%)	10 (100%)	8 (53%)	6 (60%)
Age	18-20	20 (40%)	0 (0%)	0 (0%)	1 (7%)	1 (10%)
	21-30	12 (48%)	9 (90%)	10 (100%)	11 (73%)	8 (80%)
	31-40	3 (12%)	1 (10%)	0 (0%)	3 (20%)	1 (10%)
Handedness	Right-handed	22 (88%)	9 (90%)	6 (40%)	15 (100%)	8 (80%)
	Left-handed	2 (8%)	1 (10%)	3 (30%)	0 (0%)	2 (20%)
	Both-handed	1 (4%)	0 (0%)	1 (10%)	0 (0%)	0 (0%)
Weight (kg)	Not respond	0 (0%)	0 (0%)	0 (0%)	1 (7%)	0 (0%)
	[40, 45)	0 (0%)	0 (0%)	0 (0%)	1 (7%)	0 (0%)
	[45, 50)	1 (4%)	0 (0%)	0 (0%)	2 (13%)	0 (0%)
	[50, 55)	5 (20%)	0 (0%)	0 (0%)	0 (0%)	1 (10%)
	[55, 60)	4 (16%)	0 (0%)	0 (0%)	2 (13%)	1 (10%)
	[60, 65)	4 (16%)	2 (20%)	2 (20%)	0 (0%)	4 (40%)
	[65, 70)	6 (24%)	3 (30%)	0 (0%)	2 (13%)	2 (20%)
	[70, 75)	3 (12%)	3 (30%)	4 (40%)	1 (7%)	1 (10%)
	[75, 80)	2 (8%)	1 (10%)	1 (10%)	4 (27%)	1 (10%)
	Above 80	0 (0%)	1 (10%)	3 (30%)	2 (13%)	0 (0%)
Height (cm)	[150, 160)	6 (24%)	0 (0%)	0 (0%)	3 (20%)	1 (10%)
	[160, 170)	8 (32%)	1 (10%)	3 (30%)	4 (27%)	4 (40%)
	[170, 180)	8 (32%)	9 (90%)	5 (50%)	7 (47%)	3 (30%)
	[180, 190)	3 (12%)	0 (0%)	2 (20%)	1 (7%)	2 (20%)
Wrist circumference (cm)	[13, 14)	1 (4%)	0 (0%)	0 (0%)	3 (20%)	1 (10%)
	[14, 15)	5 (20%)	2 (20%)	1 (10%)	2 (13%)	3 (30%)
	[15, 16)	10 (40%)	3 (30%)	3 (30%)	4 (27%)	3 (30%)
	[16, 17)	6 (24%)	4 (40%)	4 (40%)	1 (7%)	2 (20%)
	[17, 18)	2 (8%)	1 (10%)	2 (20%)	4 (27%)	1 (10%)
	[18, 19)	1 (4%)	0 (0%)	0 (0%)	1 (7%)	0 (0%)

Table 7. Authentication accuracy for per-pose classifiers. Mean FAR, FRR and EER measured over 100 permutations (randomly selecting 15 genuine users and 10 imposters each time). k indicates the number of donnings used for training.

		EER (%)		Probability threshold = 0.5					
				$k = 3$			$k = 4$		
Features	Hand pose	$k = 3$	$k = 4$	FAR (%)	FRR (%)	HTER (%)	FAR (%)	FRR (%)	HTER (%)
		Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
PSD	Fist	1.72 (0.69)	0.88 (0.46)	1.68 (0.69)	9.41 (3.50)	5.55 (1.80)	1.91 (0.71)	4.55 (2.48)	3.23 (1.27)
	Open	2.66 (0.99)	2.05 (0.80)	2.29 (0.85)	10.69 (4.09)	6.49 (2.05)	2.60 (0.93)	5.51 (2.41)	4.06 (1.26)
	Relax	2.23 (1.17)	2.02 (1.15)	1.97 (0.90)	7.91 (4.06)	4.94 (2.08)	2.50 (1.09)	4.34 (2.71)	3.42 (1.44)
Transfer	Fist	1.68 (0.69)	0.70 (0.36)	1.75 (0.71)	9.65 (3.53)	5.70 (1.81)	1.98 (0.74)	4.59 (2.47)	3.28 (1.27)
	Open	2.66 (0.94)	1.89 (0.77)	2.33 (0.85)	11.70 (4.46)	7.02 (2.22)	2.65 (0.93)	5.76 (2.43)	4.21 (1.26)
	Relax	1.87 (1.03)	1.57 (0.90)	2.03 (0.91)	7.57 (3.98)	4.80 (2.04)	2.58 (1.11)	4.18 (2.63)	3.38 (1.41)
MFCCs	Fist	3.06 (0.97)	2.41 (0.85)	1.28 (0.63)	11.04 (3.18)	6.16 (1.61)	1.52 (0.71)	5.25 (1.92)	3.39 (1.02)
	Open	4.61 (1.31)	3.12 (0.99)	1.96 (0.93)	13.59 (3.51)	7.77 (1.58)	2.27 (1.00)	5.67 (2.45)	3.97 (1.13)
	Relax	5.57 (1.85)	3.57 (1.46)	1.81 (0.71)	15.80 (4.71)	8.80 (2.33)	2.21 (0.86)	11.71 (4.42)	6.97 (2.25)
MFCCs + PSD	Fist	1.48 (0.64)	0.72 (0.43)	1.99 (1.05)	8.02 (3.02)	5.01 (1.48)	2.22 (1.07)	3.99 (2.44)	3.10 (1.27)
	Open	2.23 (0.82)	1.59 (0.65)	2.76 (1.28)	8.84 (3.68)	5.80 (1.70)	3.09 (1.39)	4.52 (2.26)	3.81 (1.18)
	Relax	2.01 (1.06)	1.75 (1.02)	2.51 (1.43)	7.88 (3.68)	5.19 (1.92)	3.06 (1.66)	3.69 (2.57)	3.38 (1.56)

differences in the two gender groups. As for the others, we performed correlation tests to measure the strength of a relationship between HTER results and a given physical characteristic. The graphs show the frequency counts for the resulting p-values.

Table 8. Authentication accuracy for multi-pose ensemble blenders. Mean FAR, FRR, and EER measured over 100 permutations.

Features	Hand pose	EER (%)		Probability threshold = 0.5					
				$k = 3$			$k = 4$		
		$k = 3$	$k = 4$	FAR (%)	FRR (%)	HTER (%)	FAR (%)	FRR (%)	HTER (%)
		Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
PSD	Fist/Open	0.54 (0.42)	0.15 (0.16)	0.72 (0.63)	3.94 (2.19)	2.34 (1.17)	0.88 (0.69)	0.51 (0.68)	0.70 (0.53)
	Fist/Relax	0.30 (0.21)	0.21 (0.17)	0.63 (0.52)	5.19 (2.55)	2.91 (1.27)	0.92 (0.64)	1.21 (1.53)	1.06 (0.80)
	Open/Relax	0.62 (0.45)	0.42 (0.33)	0.49 (0.62)	6.42 (3.11)	3.46 (1.58)	0.74 (0.75)	4.26 (2.14)	2.50 (1.13)
	Fist/Open/Relax	0.18 (0.14)	0.03 (0.04)	0.30 (0.32)	2.99 (2.23)	1.64 (1.12)	0.44 (0.39)	0.37 (0.51)	0.40 (0.31)
Transfer	Fist/Open	0.59 (0.45)	0.16 (0.17)	0.74 (0.64)	4.41 (2.37)	2.58 (1.23)	0.91 (0.69)	0.58 (0.70)	0.75 (0.53)
	Fist/Relax	0.31 (0.21)	0.22 (0.17)	0.69 (0.56)	5.08 (2.54)	2.88 (1.27)	0.99 (0.68)	1.11 (1.48)	1.05 (0.76)
	Open/Relax	0.69 (0.48)	0.47 (0.35)	0.49 (0.61)	6.45 (3.08)	3.47 (1.56)	0.77 (0.78)	4.21 (2.14)	2.49 (1.14)
	Fist/Open/Relax	0.20 (0.16)	0.03 (0.04)	0.32 (0.33)	3.37 (2.46)	1.84 (1.24)	0.46 (0.40)	0.42 (0.54)	0.44 (0.33)
MFCCs	Fist/Open	0.32 (0.28)	0.11 (0.16)	0.63 (0.55)	7.27 (2.78)	3.95 (1.40)	0.82 (0.68)	0.34 (0.51)	0.58 (0.41)
	Fist/Relax	0.57 (0.32)	0.33 (0.25)	0.68 (0.53)	12.26 (3.95)	6.47 (2.01)	0.88 (0.58)	3.68 (2.73)	2.28 (1.40)
	Open/Relax	0.87 (0.43)	0.26 (0.21)	0.58 (0.55)	13.39 (3.74)	6.99 (1.86)	0.86 (0.70)	5.07 (2.28)	2.97 (1.17)
	Fist/Open/Relax	0.07 (0.09)	0.05 (0.07)	0.41 (0.44)	8.04 (2.84)	4.22 (1.40)	0.62 (0.58)	1.22 (1.41)	0.92 (0.79)
MFCCs + PSD	Fist/Open	0.44 (0.35)	0.11 (0.13)	0.62 (0.52)	3.54 (2.16)	2.08 (1.12)	0.79 (0.59)	0.54 (0.63)	0.67 (0.45)
	Fist/Relax	0.28 (0.19)	0.19 (0.16)	0.48 (0.41)	5.08 (2.61)	2.78 (1.29)	0.74 (0.54)	1.09 (1.42)	0.92 (0.72)
	Open/Relax	0.54 (0.40)	0.34 (0.28)	0.39 (0.30)	6.78 (2.06)	3.58 (1.03)	0.63 (0.40)	4.16 (1.61)	2.39 (0.80)
	Fist/Open/Relax	0.17 (0.14)	0.01 (0.02)	0.24 (0.25)	3.41 (2.47)	1.82 (1.24)	0.39 (0.34)	0.39 (0.52)	0.39 (0.31)

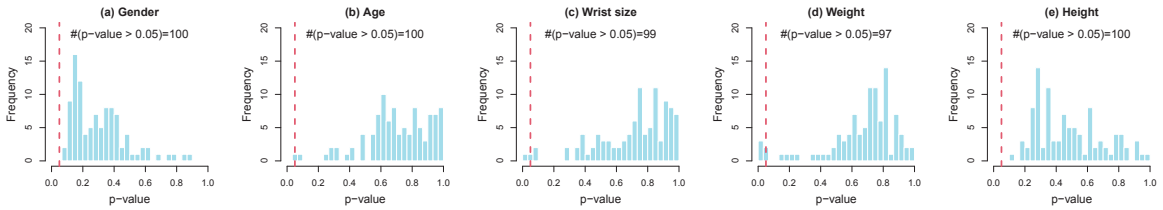

Fig. 9. Relationship between HTER and physical characteristics (gender, age, wrist size, weight, and height) of genuine users. For each of the 100 random user set permutations, we performed two-sample t-tests to compared the two gender groups; for the rest, we performed correlation tests. We picked the fist-based classifiers trained with PSD features and $k = 4$, and their HTER results for this analysis. The graphs show the frequency count of the resulting p-values.

E MULTI-RECALL SESSION RESULTS

We evaluated multi-recall session performance based on the best performing PSD and MFCC concatenated features. Table 9 shows the average EER, FAR, and FRR of the three pose-specific classifiers as well as the multi-pose ensemble blenders. The first row shows the first-visit enrollment performance to validate the second study results against the first (single-session) study results. The second row presents accuracy results for the classifiers that were trained with just the enrollment set (i.e., without any retraining). The rest of the table shows the retraining performance: we increased the number of recall sessions that were included in the genuine retrain set from one to three.

To show the effectiveness of combining PSD and MFCC features, we also show the performance of these individual features in Figure 10. It is clear that the two feature sets compliment each other, and show the best performance when they are concatenated and used together.

Table 9. Multi-recall session authentication accuracy based on the PSD and MFCC concatenated features. Mean FAR, FRR, and EER of the pose-specific classifiers as well as multi-pose ensemble blenders, computed across nine subjects. The first row shows the first-visit enrollment performance. The second row presents accuracy results without any retraining. The rest of the table shows the retraining performance: we increase the number of recall sessions that are included in the retrain set from one to three.

Feature	Training	Testing	Hand pose	EER (%)	Probability threshold = 0.5			
					FAR (%)		FRR (%)		HTER (%)	
				Mean	(SD)	Mean	(SD)	Mean	(SD)
MFCCs+PSD	Enrollment First four donnings	Enrollment Last donning	Fist	0.00	(0.00)	0.97	(1.56)	0.00	(0.00)	0.45	(0.78)
			Open	0.09	(0.28)	0.55	(0.91)	0.00	(0.00)	0.28	(0.45)
			Relax	0.00	(0.00)	3.70	(4.14)	0.00	(0.00)	1.85	(2.07)
			Fist/Open	0.00	(0.00)	0.35	(0.60)	0.00	(0.00)	0.17	(0.30)
			Fist/Relax	0.00	(0.00)	0.77	(1.41)	0.00	(0.00)	0.38	(0.70)
			Open/Relax	0.00	(0.00)	1.08	(1.67)	0.00	(0.00)	0.54	(0.83)
			Fist/Open/Relax	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)
	Enrollment	Recall 1-5	Fist	5.78	(6.57)	0.97	(1.56)	29.56	(31.98)	15.26	(15.52)
			Open	6.28	(7.25)	0.55	(0.91)	46.52	(31.60)	23.54	(15.69)
			Relax	10.23	(9.87)	3.70	(4.14)	38.59	(28.37)	21.15	(14.52)
			Fist/Open	2.76	(2.85)	0.35	(0.60)	40.30	(33.46)	20.32	(16.55)
			Fist/Relax	6.16	(8.68)	0.77	(1.41)	39.48	(33.54)	20.13	(16.51)
			Open/Relax	7.13	(8.36)	1.08	(1.67)	45.11	(30.47)	23.10	(14.82)
			Fist/Open/Relax	5.06	(6.89)	0.00	(0.00)	52.52	(36.28)	26.26	(18.14)
	Enrollment + Recall 1	Recall 2-5	Fist	1.83	(2.07)	3.10	(2.81)	11.94	(13.54)	7.52	(6.08)
			Open	2.62	(4.32)	2.87	(3.86)	19.35	(25.84)	11.11	(13.02)
			Relax	7.31	(9.84)	5.23	(5.42)	15.19	(22.51)	10.21	(12.06)
			Fist/Open	1.43	(3.28)	1.43	(2.19)	11.85	(11.98)	6.64	(5.74)
			Fist/Relax	1.94	(3.06)	2.78	(3.18)	12.13	(17.78)	7.45	(8.72)
			Open/Relax	3.44	(4.31)	1.97	(2.80)	18.80	(18.94)	10.38	(9.63)
			Fist/Open/Relax	1.36	(3.00)	1.50	(2.28)	8.98	(15.46)	5.24	(7.60)
	Enrollment + Recall 1-2	Recall 3-5	Fist	1.03	(1.20)	3.84	(3.50)	2.10	(3.40)	2.97	(1.55)
			Open	2.07	(3.84)	3.29	(4.41)	14.44	(30.37)	8.87	(15.70)
			Relax	5.51	(7.58)	6.25	(5.47)	7.04	(13.99)	6.64	(8.46)
			Fist/Open	0.50	(1.19)	1.81	(2.26)	4.20	(10.98)	3.01	(5.29)
			Fist/Relax	0.40	(0.60)	4.01	(4.19)	3.33	(10.00)	3.67	(4.68)
			Open/Relax	1.89	(3.20)	2.86	(3.27)	4.07	(11.03)	3.46	(5.28)
			Fist/Open/Relax	1.13	(3.02)	2.46	(3.70)	3.70	(11.11)	3.08	(5.51)
	Enrollment + Recall 1-3	Recall 4-5	Fist	1.25	(1.79)	5.42	(5.30)	1.11	(3.33)	3.26	(2.63)
			Open	2.04	(4.37)	4.54	(5.53)	12.41	(20.67)	8.47	(10.83)
			Relax	4.38	(7.16)	8.15	(6.73)	6.30	(16.54)	7.22	(9.07)
			Fist/Open	0.51	(1.08)	2.62	(3.46)	1.67	(4.41)	2.14	(2.32)
			Fist/Relax	0.17	(0.51)	5.17	(6.14)	3.33	(10.00)	4.25	(4.97)
			Open/Relax	1.66	(2.52)	3.32	(3.50)	5.56	(16.67)	4.44	(7.88)
			Fist/Open/Relax	0.79	(2.22)	3.47	(4.88)	4.44	(13.33)	3.96	(6.55)

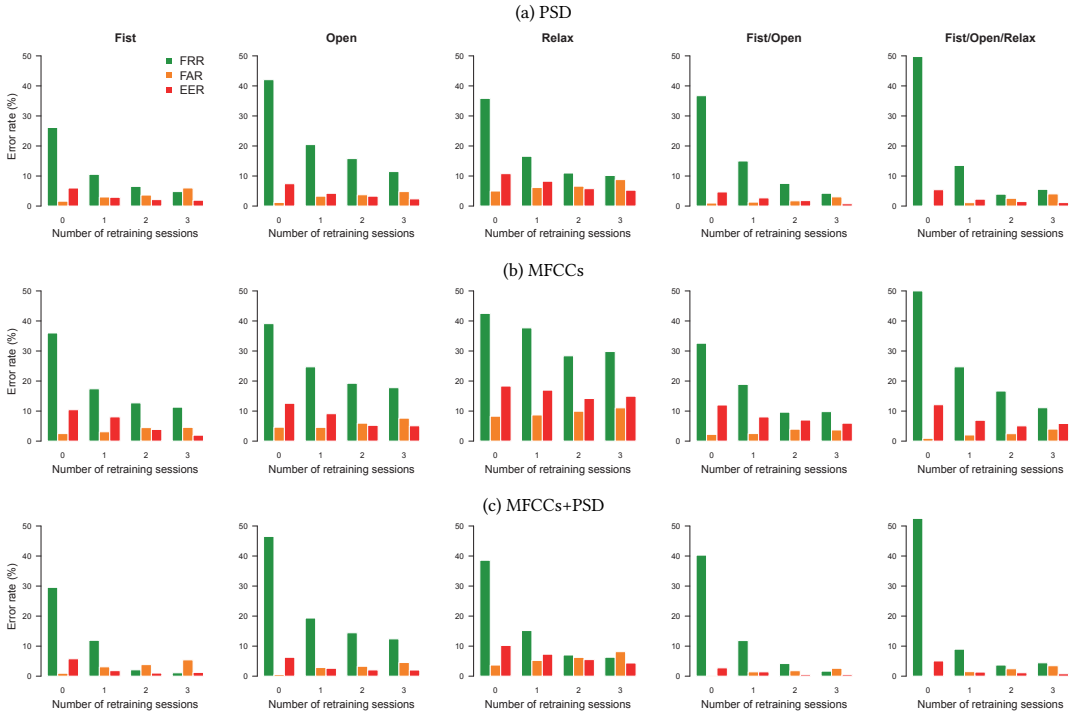

Fig. 10. Multi-session average FRR, FAR (threshold 0.5), and EER with respect to varying number of recall sessions included in the retrain set ("number of retraining sessions"). "0" indicates that classifiers trained only on enrollment data were used. We show the individual PSD and MFCC feature set performance through the graphs in (a) and (b), and the concatenated performance through the graphs in (c).