SoundLock: A Novel User Authentication Scheme for VR Devices Using Auditory-Pupillary Response

Huadi Zhu*, Mingyan Xiao†, Demoria Sherman*, and Ming Li*
* The University of Texas at Arlington, Email: {huadi.zhu, demoria.sherman}@mavs.uta.edu, ming.li@uta.edu
† California State Polytechnic University, Pomona, Email: mxiao@cpp.edu

Abstract—**Virtual Reality (VR) has shown promising potential in many applications, such as e-business, healthcare, and social networking. Rich information regarding users' activities and online accounts is stored in VR devices. If they are carelessly unattended, adversarial access will cause data breaches and other critical consequences. Practical user authentication schemes for VR devices are in dire need. Current solutions, including passwords, digital PINs, and pattern locks, mostly follow conventional approaches for general personal devices. They have been criticized for deficits in both security and usability. In this work, we propose SoundLock, a novel user authentication scheme for VR devices using auditory-pupillary response as biometrics. During authentication, auditory stimuli are presented to the user via the VR headset. The corresponding pupillary response is captured by the integrated eye tracker. User's legitimacy is then determined by comparing the response with the template generated during the enrollment stage. To strike a balance between security and usability in the scheme design, an optimization problem is formulated. Due to its nonlinearity, a two-stage heuristic algorithm is proposed to solve it efficiently. The solution provides necessary guidance for selecting effective auditory stimuli and determining their corresponding lengths. We demonstrate through extensive in-field experiments that SoundLock outperforms state-of-the-art biometric solutions with FAR (FRR) as low as 0.76% (0.91%) and is well received among participants in the user study.**

I. Introduction

Motivation. The rapid development of virtual reality (VR) has been seen in the past few years with a consistently growing popularity. According to a recent report [83], the VR market is around $28 billion in 2022; by 2030, the number is forecast to reach over $87 billion with a constant annual growth rate of 15%. With the capability of providing an immersive and interactive experience, VR has revolutionized gaming and entertainment and permeated a variety of applications, including e-commerce, education, healthcare, and military [102]. For example, retailers can bridge physical and online stores via VR to provide an immersive shopping experience for customers [60]; medical practitioners may communicate with patients in a VR environment for remote diagnosis [67]; military actions can be simulated and practiced in a virtual battlefield [53]. In the above applications, tremendous amounts of sensitive data are collected, processed, and stored on VR devices, such as customers' credit card information, patients' health status, and military secrets. Adversarial access to VR devices would cause data breaches and other critical consequences. Therefore, implementing user authentication mechanisms in VR is a crucial step in resisting unauthorized access.

However, user authentication on VR devices is still at the infant stage. Current solutions, including passwords, digital PINs, and pattern locks, mostly follow conventional approaches for general personal devices. Users have to use some external hand controllers to enter the credentials. They have been criticized for the usability deficit: It takes users substantial effort to select correct keys from the virtual keyboard using the controller [95]. What's worse, they are shown to be vulnerable to shoulder-surfing attacks. As the user enters her credential, the hand movement leaves a trajectory that can be easily mapped to the entered secrets with the keyboard layout [36, 95, 118]. Per the statistics from prior work [36], the success rate of shoulder-surfing attacks towards PINs and drawing patterns in VR is as high as 18%.

To address the above issues, great efforts have been devoted to exploring practical alternatives. Existing approaches can be generally categorized into the following classes: *knowledge-based methods* [35–37, 62, 116], *physiological biometrics* [9, 25, 51, 88], *behavioral biometrics* [56, 69, 75, 90, 117], *token-based methods* [23], and a mixture of above [61, 113, 118]. Among them, physiological biometrics attract the most attention due to its high usability and authentication accuracy. Nonetheless, its wide deployment is still faced with several challenges. First, to access the user's biometrics, such as electroencephalogram (EEG), electrocardiogram (ECG), electromyography (EMG), and iris patterns, dedicated and costly sensors are needed. These sensors are mostly unavailable in current VR headsets. While iris scans have been deployed on HoloLens 2, a high-end augmented reality (AR) device costing at least $3,500, they are less likely to integrate into an even broader set of medium-/low-end terminals with much lower budgets. Second, most physiological biometrics are irrevocable. Once a biometric credential is compromised or stolen, it cannot be reset. This property is also called cancelability.

Our approach. In this paper, we propose to leverage a new kind of biometric, *auditory-pupillary response*, for user authentication on VR headsets. By presenting users with auditory stimuli, the pupil's reaction, in the form of size changes, is universally observable among human beings [14, 40, 63, 70, 85]. The auditory-pupillary response is an autonomic reflex that dilates or constricts the pupil, mediated by the sympathetic and parasympathetic nervous systems, which are both parts of the autonomous nervous system. The biological uniqueness in the complex neural pathways and structure of iris muscles present particular features that make it possible to explore auditory-pupillary responses for user identification. As vali-

Network and Distributed System Security (NDSS) Symposium 2023
27 February - 3 March 2023, San Diego, CA, USA
ISBN 1-891562-83-5
https://dx.doi.org/10.14722/ndss.2023.24298
www.ndss-symposium.org

dated in our preliminary study (see Section II-B), inter-subject pupillary responses exhibit distinguishable patterns under the same stimulus, whereas intra-subject pupillary responses are consistent in multiple trials. These observations motivate us to develop SoundLock, a novel reflex physiological biometric authentication method for VR devices based on the auditory-pupillary response. During authentication, carefully designed auditory stimuli are rendered to the user via the VR device's audio channel. The corresponding pupillary response is captured by the eye tracker integrated into the device. The user's legitimacy is then determined by comparing the response with the template generated during the enrollment stage.

Compared with conventional authentication methods for VR, such as passwords, digital PINs, and drawing patterns, our scheme has the following prominent advantages. First, its usability has been greatly enhanced as it significantly reduces user effort for credential entry. A user's biometric, i.e., the auditory-pupillary response, is automatically gathered by the device. The entire process is hand-free and relieves users from memory burdens. Second, since the user's eyes are completely blocked by the VR headset, it is impossible for an adversary to gain visual observation of the authentication process to launch shoulder-surfing. Meanwhile, SoundLock, as a new kind of reflex physiological biometric for VR, outperforms existing static biometric [21, 48, 56, 69, 84] in the following aspects: First, auditory-pupillary responses are revocable. In the case of having one pupillary response stolen or counterfeited, a new credential can be easily generated by changing its associated stimulus. Second, SoundLock can be implemented on many mainstream VR headsets, e.g., HTC VIVE Pro Eye, Pico Neo series, Varjo VR-3, and Fove VR [1–5], which are already equipped with eye trackers. It is well accepted that incorporating eye-tracking technology is a trend in VR to assist in simulating depth of field and focus and providing users a more realistic and natural visual experience [26, 44, 94].

Despite these attractive properties, the design of SoundLock is faced with several non-trivial challenges. First, while pupillary response exhibits prominent inter-subject distinguishability, identifying essential features out of raw pupil size measurement for accurate user authentication is not an easy task. No prior research has been conducted on this topic. We thoroughly investigate 60 features, including morphological features that are pupillary response-specific and general statistical features, and narrow them down to 20 that best represent the uniqueness of each individual. We validate through a comprehensive evaluation that the selected features effectively produce high authentication accuracy. Second, to enlarge the credential pool, we adopt multiple auditory stimuli. However, the multi-stimuli prolong the authentication time and thus impair usability. To mitigate this issue, we model the problem into an optimization problem that maximizes the authentication accuracy while satisfying a hard constraint on the authentication time (see Section IV). It aims to balance security and usability. Realizing that it is challenging to directly solve the problem optimally owing to its non-linearity, we devise a two-stage heuristic algorithm to find the approximate solution efficiently. Lastly, like other biometrics, the auditory-pupillary response may exhibit variations over time. As a result, its authentication performance may degrade over a long time span. To deal with this issue, we adopt an adaptive biometric strategy to consistently update the classification model with the coming of new samples.

To evaluate the performance of SoundLock, we implement it on a VR device and carry out extensive experiments involving 44 participants. It achieves an F1-score of 0.984, FAR of 0.76%, and FRR of 0.91%, outperforming state-of-the-art solutions. Besides, our scheme can be performed within a practical authentication time of 7 s. SoundLock also demonstrates satisfactory consistency under various testing conditions. Finally, the user study manifests that our scheme is well received among the participants; especially, 72.7% of them are willing to adopt SoundLock as the authentication scheme on their (future) VR devices.

To summarize, the contributions of this paper include:

- We investigate a new kind of reflex physiological biometric, auditory-pupillary response, for user authentication on VR devices. We validate its feasibility through a measurement study.
- To model the response for user authentication, we investigate a set of morphological and statistical features, which are proven effective in producing high authentication accuracy.
- To strike a balance between security and usability in the design, we formulate an optimization problem. A two-stage heuristic algorithm is proposed to efficiently solve the problem with an approximate solution.
- We perform extensive in-field experiments to evaluate SoundLock. Results demonstrate that the proposed scheme outperforms state-of-the-art biometric authentication solutions and is well received among participants in the user study.

II. Preliminaries

A. Background on Auditory-Pupillary Response

The pupil size has been proven sensitive to a wide variety of auditory stimuli [14, 40, 63, 70, 85]. Fig. 1 exhibits pupil size, measured in pixels, changes as a subject is presented with an auditory stimulus, a white noise that starts at 1 s and stops at 5 s. This sample is randomly selected from our collected dataset. Measures from only one eye are collected since pupillary responses in both eyes have been confirmed to be consensual [45]. The presentation of an auditory stimulus results in a multi-phasic pupillary response. The initial phasic response is evoked with transient pupil dilation shortly after the stimulus onset, followed by a constriction. This process is followed by the second round of, and sometimes more rounds of, dilations and constrictions with attenuated amplitudes. After the stimulus offset, the pupil gradually returns to its baseline, i.e., the pupil size under the no-stimulus condition, accompanied by minor fluctuations [79].

Physiologically, the pupillary response is controlled by two muscles: the *iris radial muscle* (IRM) increasing the pupil size and the *iris sphincter muscle* (ISM) reducing the pupil size [18]. The balance between the sympathetic and parasympathetic nervous systems determines pupil size. The underlying mechanisms are complex; the relative contribution of the two systems depends on a variety of factors, such as stimulus characteristics and cognitive activities. Pupil dilation is controlled by the IRM. IRM consists of fibers that are oriented radially and connect the exterior of the iris with the interior. When IRM contracts, it pulls the interior of the iris

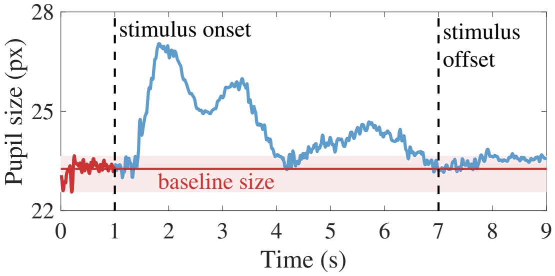


Fig. 1: Pupillary response to auditory white noise.

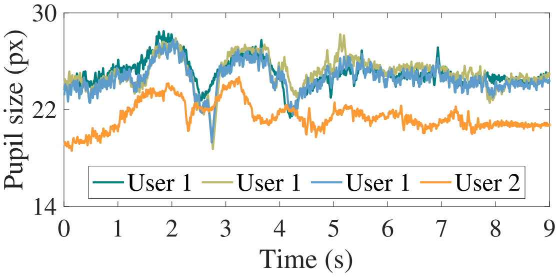


Fig. 2: Intra-/Inter-subject pupillary response.

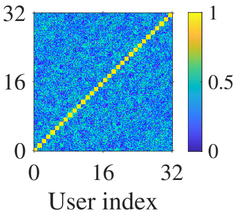


Fig. 3: Confusion matrix of Pearson correlation coefficients among 32 subjects.

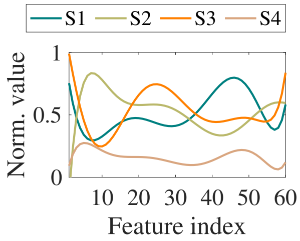


Fig. 4: Normalized values of 60 features extracted from pupillary responses.

outward, thus increasing the size of the pupil. Upon perception of auditory stimuli, psycho-sensory arousals are first triggered at the hypothalamus and the locus coeruleus. The activities on the hypothalamus and the locus coeruleus reflect arousals and project to the intermediolateral column of the spinal cord. The arousals finally reach IRM via a complicated network of nerves and cause contraction. In contrast, pupil constriction is controlled by ISM, which encircles the pupil like a cord that reduces pupil size when it contracts. As shown in Fig. 1, the pupil constricts once it dilates to a large extent. This process operates through the opposite action of pupil dilation. ISM is directed through the parasympathetic pathway. The activated Edinger-Westphal nucleus transmits information via the oculomotor nerve to the ciliary ganglion, which is located behind the eyeball. The information is further sent via the short ciliary nerve to innervate the ISM to contract. In short, the pupil dynamics observed under auditory stimuli are a joint effect delivered by IRM, ISM, and their corresponding neural pathways [57, 63, 70, 98, 105, 106].

B. Measurement Study

While the phenomenon of auditory-pupillary response is well recognized, whether it can be exploited for user authentication remains unclear. Our measurement study intends to answer this question by carrying out extensive experiments. A total of 32 subjects are invited. They listen to auditory stimuli of different types via the HTC VIVE Pro VR headset. A total of 20 stimuli are adopted, including white noise, monotones, prompt sounds, natural sounds, and human voices. They have been widely adopted in prior works on auditory-pupillary response [14, 40, 50, 63, 70]. Each auditory stimulus is a 6-second audio track. Subjects' pupillary responses are captured by a Pupil Labs eye tracker that is integrated into the headset. To facilitate the data collection, a specialized app is built using Unity, a cross-platform engine for VR development. To avoid impact from visual stimuli, participants are exposed to a dark VR environment, i.e., no image is displayed. The above process is repeated 20 times for each participant. The following analysis is conducted based on the collected 12,800 samples, i.e., time-resolved pupil size sequences.

Intra- and inter-subject pupillary response. Fig. 2 shows pupillary responses from four trials under the same stimulus. Three of them are collected from the same subject. The three intra-subject responses exhibit similar patterns, although they are from different trials. It indicates that pupillary response is relatively consistent for the same user. Meanwhile, inter-subject responses exhibit distinguishable patterns. To better quantify the intra-/inter-subject response relationship, Fig. 3 further plots the confusion matrix (160×160) of pupillary responses among the 32 participants in response to one stimulus. 5 samples are randomly selected from each participant. The Pearson correlation coefficient (PCC) is adopted. The PCC values on the diagonal line ($\mu = 0.91, \sigma = 0.02$) are significantly higher than those off the line ($\mu = 0.36, \sigma = 0.14$). It implies that individuals exhibit diverse pupillary responses when presented with the same auditory stimulus, while those from the same subject are consistent.

Pupillary response under various stimuli. We then play a variety of auditory stimuli to the subject. It is observed in Fig. 5 that the corresponding pupillary responses vary across the stimuli. We further extract 60 features out of the raw measures. Fig. 4 depicts their normalized values. Polynomial regression is applied for better illustration. The feature vectors are distinguishable with respect to various stimuli. Intuitively, it is possible to generate a large number of credentials for a user from her pupillary responses by applying various auditory stimuli. More importantly, these credentials can be easily revoked: In the case of having one pupillary response stolen, a new credential can be generated by changing its associated stimulus, which is called *cancelability* [80]. In contrast, this property does not exist in conventional biometrics, such as fingerprints, irises, and faces, which are static to human beings. Once their credentials are damaged or counterfeited, the user cannot cancel the pre-stored credentials or reset them.

Summary. Our findings are encouraging. First, given the same auditory stimulus, intra-subject pupillary responses exhibit consistent patterns in multiple trials, while inter-subject pupillary responses are distinguishable. This property lays the foundation for our idea that utilizes auditory-pupillary response as a new kind of biometric for user authentication. Second, the responses are diverse with respect to various stimuli. It thus motivates us to employ a sequence of stimuli to enlarge the pupillary response-based credential pool. More importantly, the property that the induced credential is stimuli-dependent offers the potential to design a cancelable biometric. An in-use pupil credential can be revoked and updated by simply applying new auditory stimuli. Lastly, we observe in the

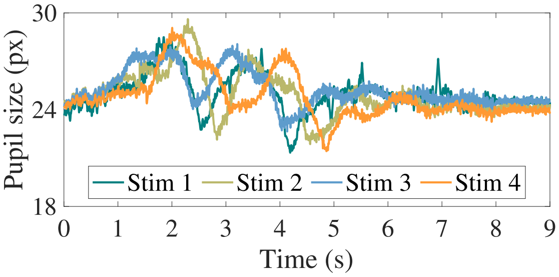


Fig. 5: Pupillary response under different stimuli.

measurement that the pupil demonstrates a stable behavior in response to auditory stimuli: It first dilates with the stimulus onset and then constricts, followed by a couple of more rounds of dilation-constriction until the stimulus offset. The transitional changes in the pupil size generate consecutive waveforms bearing rich information for authentication. We will investigate in Section III-B how to extract essential features.

C. Problem Statement

System model. We consider a general user authentication scenario on VR devices, where a user has to provide a correct credential to log in. We assume that the headset is equipped with an eye tracker for pupil detection and pupil size measurement. The proposed authentication scheme is composed of two stages. In the enrollment stage, the headset plays carefully designed audio stimuli and records users' corresponding pupillary responses. A set of relevant features are extracted upon which a classification model is trained and optimized. In the login stage, a user is presented with the same stimuli. The collected pupillary response is compared with the enrolled ones to determine the user's legitimacy.

Many mainstream VR headsets are equipped with eye trackers nowadays, such as Meta Quest Pro, HTC VIVE Pro Eye, PlayStation VR2, Pico Neo series, Varjo VR-3, and Fove VR. The list continues to grow. It is well recognized that eye tracking benefits VR in the following aspects: a) delivering a higher-quality graphics experience through foveated rendering, b) improving wearing comfort by automatically adapting the device to the user via calculating the user's inter-pupillary distance, and c) enhancing the interactions among virtual avatars to better reflect the user's visual attention. It is well accepted that incorporating the eye-tracking technology is a trend in the future development of VR [44, 94].

Adversary model. The adversary's goal is to impersonate the legitimate user and log into the VR headset. The adversary is assumed to have physical control of the headset and sufficient time to perform the attack. For example, the VR device is lost or stolen. We primarily consider the impersonation attack [58] throughout this work. The adversary intends to use its own biometric credential, i.e., pupillary response, under the auditory stimuli to get authenticated. Other common attacks will be discussed in Section V-A.

III. Basic Scheme Design

We start by introducing a basic scheme that renders a single auditory stimulus. It consists of three main components: preprocessing, feature extraction and selection, and classification. Upon the acquisition of a pupillary response, it is first preprocessed for signal cleaning. Then a set of response-specific features are extracted as well as selected. In the enrollment stage, these features are used to train the classifier; in the login stage, they are fed into the trained classifier for authentication.

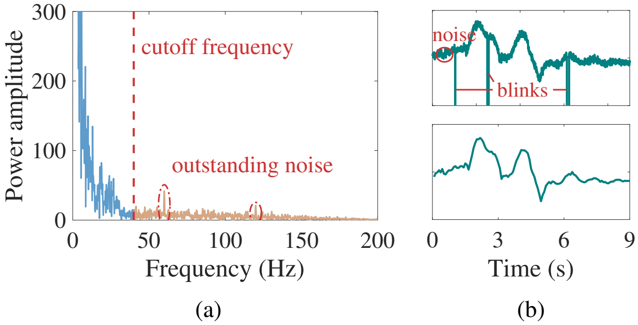


Fig. 6: (a) Raw measurement of pupillary response in the frequency domain. (b) Raw measurement (top) and processed (bottom) pupillary response in the time domain.

A. Preprocessing

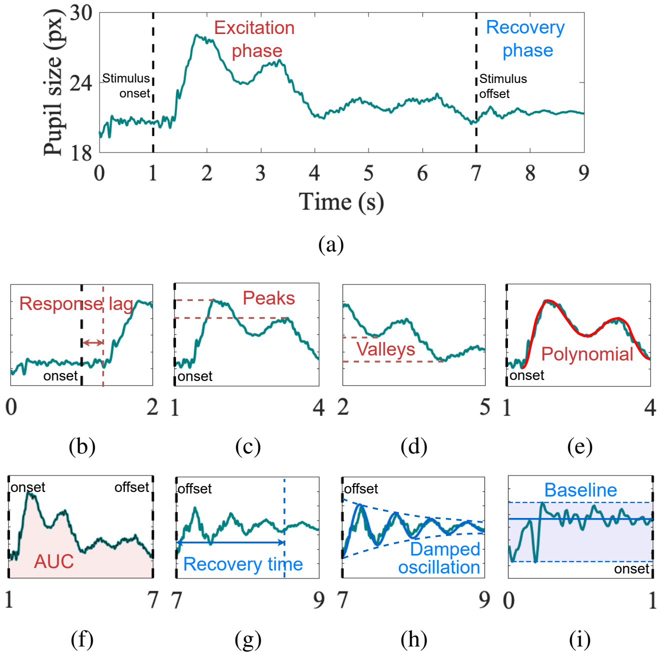
The pupillary response is acquired by an embedded eye tracker sampling at 200 Hz. Fig. 6b (top) plots the raw measurements, which are mixed with noise and zero-readings. This component aims to eliminate them and extract useful information from the raw measurement. The background noise is mainly caused by internal and external electromagnetic radiations (e.g., VR display refreshing, power line emanation, and their harmonics) that primarily exist above 50 Hz. In opposition, the frequency components of pupil size variations mainly reside at the lower end of the frequency band, as shown in Fig. 6a. Hence, we apply a low-pass filter with a cutting frequency of 40 Hz to eliminate the above-mentioned noise. The intermittent zero-readings exist in the measurement due to spontaneous blinks. We apply the classic interpolation method to smooth the pupillary response signals. Fig. 6b (bottom) plots the pupillary response after preprocessing.

B. Feature Extraction and Selection

We extract two types of features from the processed pupillary response: morphological features and statistical features. The former is features specifically proposed to outline the morphology of the auditory-pupillary response patterns; they reveal the intrinsic geometrical characteristics in the multiphasic signals. The latter is provides a more general description of the signal statistics. As demonstrated in Fig. 7a, a pupillary response can be divided into two phases: *excitation phase* and *recovery phase*. In the following, we provide details of extracting the candidate morphological features from both phases.

Excitation phase. It is between the stimulus onset and the stimulus offset. In this phase, the pupil is provoked by the stimulus and experiences transitional dilations and constrictions.

- *Response lag.* It is defined as the latency between the stimulus onset and the moment the pupil reacts to it, as shown in Fig. 7b. Prior studies show that this value is mostly determined by the neural pathways while less affected by mechanical properties of the iris muscles [63, 106]. Differences in response latency among individuals have been reported [12, 16, 33, 38,



(a)

(b) (c) (d) (e)

(f) (g) (h) (i)

Fig. 7: (a) Illustration of excitation and recovery phase. (b) Response lag. (c) Peak magnitudes. (d) Valley magnitudes. (e) Polynomial coefficients. (f) AUC. (g) Recovery time. (h) Damped oscillation. (i) Baseline size.

100, 115]. In general, senior people tend to exhibit longer response lag [38, 100].

- *Peak/Valley magnitudes.* Upon the stimulus onset, the pupil size increases as the pupil dilates and reaches a large extent. Thereafter, the pupil size decreases as it constricts. Multi-round dilations and constrictions generate a series of waveforms. The corresponding peak (valley) magnitudes then serve as the features as shown in Fig. 7c (Fig. 7d). A classic peak detection technique [20] is applied to identify peaks and valleys in the response waveforms.
- *Dilation/Constriction rates.* Apart from the peak and valley magnitudes in the response waveforms, we are also interested in the dilation/constriction rates. They are manipulated by a complex mechanism involving the iris muscles and many components along the neural pathways such as the nerve fibers in the intermediolateral column, the super cervical ganglion, and the ciliary nerves [57, 63, 98]; these rates reflect the biological heterogeneity in the human nervous systems and iris muscles. The dilation rate is calculated as the pupil size change in one dilation divided by the associated time duration. The definition of the constriction rate follows similarly.
- *Polynomial coefficients.* n-degree polynomials are applied to approximate the response waveforms during the excitation phase. We mainly focus on the first two waveforms as the rest tend to attenuate mixed with more noise. n is set to 4 empirically. Fig. 7e depicts derived approximate polynomials; they align well with the ground truth. The corresponding coefficients in the polynomials are treated as a subset of features.
- *Area under the curve (AUC).* It is the area of the response curve during the excitation phase, as illustrated in Fig. 7f. In general, the AUC tends to be larger when a user is more agile with the auditory stimulus. AUC is derived by taking the integral of the pupillary response over time.

Recovery phase. It starts from stimulus offset and lasts until the response cutoff.

- *Recovery time.* It is the time the pupil takes to return to its baseline. As depicted in Fig. 7g, it denotes the time interval between the stimulus offset and when the pupil stabilizes with negligible deviations from its baseline.
- *Damped oscillation.* With the stimulus offset, the pupil size gradually returns to its baseline, accompanied by oscillatory behavior, as illustrated in Fig. 7h. We propose to approximately characterize this pattern using a classic damped sine wave model: $y(t) = Ae^{-\lambda t}\cos(\omega t - \phi) + C$ [39]. The function parameters, A, λ, ω, and ϕ, are taken as a subset of features.
- *Pupillary unrest index (PUI).* Human eyes exhibit continuous pupil size fluctuations, known as pupillary unrest [41, 71, 89]. Although its origins are complex, this phenomenon is mediated by fluctuating inhibitory activity within the parasympathetic Edinger Westphal nucleus, possibly driven indirectly by the locus coeruleus [43, 82, 92]. The pupillary unrest index (PUI) has been proposed in prior work to characterize the pupillary unrest behavior [54]. It is defined as cumulative changes in the average pupil size in consecutive observation windows. We thus adopt PUI as part of the features.
- *Baseline size.* The pupil baseline size, depicted in Fig. 7i, has been well recognized as a user-specific biometric trait [22, 72]. It is the eye's natural status when no external stimulus is applied. In this work, several baseline-related parameters are considered, including the average size, maximum, minimum, standard deviation, and interquartile range. The baseline can be estimated once the pupil is recovered from the excitation status or before stimulus onset.

Aside from the above-mentioned morphological features, we also take into account general statistical features of pupil size from both phases, such as average, variance, median, skewness, and kurtosis. Since these statistical features have been widely adopted in signal characterization [8, 72, 118], we do not expand the discussion here. A full list of candidate features is given in Table VIII of Appendix A.

Feature selection. This step selects from the candidate features the most relevant ones for user authentication. The refined feature set helps to reduce the computation complexity and avoid model overfitting. To this end, we calculate the Fisher score for each feature, which is defined as the ratio between the feature's inter-class and intra-class variances; a higher ratio indicates a more significant role in contributing to classification accuracy. All candidate features are sorted according to their normalized Fisher scores in Fig. 16 (see Appendix A). Finally, the top 20 features are selected to feed into the classification model. These selected features include morphological features such as the dilation rates, the peak

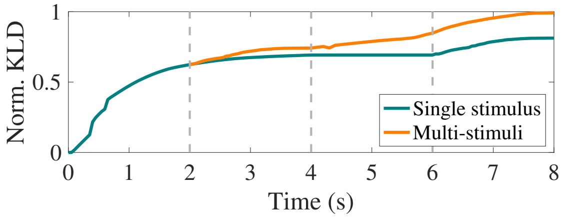


Fig. 8: Normalized KLD with respect to time. Comparison between KLD of response features under a single auditory stimulus and under multi-stimuli.

magnitudes, and the second valley magnitude; the only selected statistical feature is the average pupil size. The reason that morphological features rank relatively higher is probably that they precisely characterize the dynamics in pupillary response, whereas statistical features that are more generic and abstract.

C. Classification

The remaining task is to apply a classification method over the selected features for user authentication, i.e., to discriminate between the legitimate user and imposters. Two types of classifiers are adopted and evaluated in this work: one-class classifiers and binary classifiers. The former is trained only with samples from the class of interest, i.e., the enrolled legitimate user. The latter is trained on an explicitly labeled dataset of both classes, i.e., the legitimate user's samples and imposters' samples. The following representative machine learning models are employed. *k-nearest neighbor (k-NN):* It measures the similarity between testing samples and training samples and makes the decision by comparing the similarity with a threshold. It has been proven effective especially in cases with small training datasets. *Support vector machines (SVM):* Its main idea is to find a hyperplane in a multi-dimensional space that distinctly separates data points from different classes. Aside from k-NN and SVM, other common classification methods, including *logistic regression (LR)*, *Gaussian Naive Bayes (GNB)*, and *random forest (RF)*, are also evaluated in this work.

IV. Advanced Scheme with Multi-stimuli

The basic scheme utilizes one auditory stimulus. Inspired by the strong password selection criteria, e.g., more characters and a mixture of numbers, letters, and special characters, we propose to present the user multi-stimuli to enhance the response feature diversity. Specifically, a series of stimuli are played sequentially. Then, all the responses are concatenated and serve as the user's credentials. While the idea is simple, a critical issue is to decide the duration of each stimulus, as a too-long overall duration would impair the usability.

To facilitate the discussion, we adopt the metric Kullback-Leibler divergence (KLD) [47]. It is an indicator of similarity between two probability distributions $P(x)$ and $Q(x)$

$$D_{KL} = \sum_{x \in X} P(x) \log \left(\frac{P(x)}{Q(x)} \right). \tag{1}$$

We let $P(x)$ be the feature distribution of the enrolled user, and $Q(x)$ be that of the reference users, i.e., all the other users from the dataset. X stands for the feature space, and $x \in X$ denotes any possible interval of a feature value. Here KLD represents the distinguishability of the enrolled user from all other users.

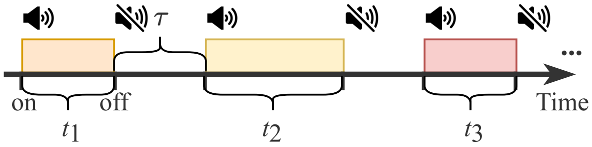


Fig. 9: The time sequence of multiple stimuli.

The larger the value is, the more distinguishable the user is, and the more accurately it can be identified. We further formulate KLD as a function of time. After the stimulus onset, more features are extracted from the measurement as time proceeds. For instance, features in the excitation phase are first derived, followed by features from the recovery phase. Fig. 8 shows the normalized KLD with respect to time. We first plot the KLD under a single stimulus (the cyan curve labeled as "Single stimulus" in Fig. 8). The stimulus starts at $t = 0$. KLD first rises quickly as t is between 0 s and 2 s. Its growth slows down as t passes 2 s. This implies the marginal benefit diminishes for involving more features under the same stimulus. We also show the KLD of employing two stimuli (the orange curve labeled as "Multi-stimuli" in Fig. 8). The first stimulus starts at $t = 0$ and stops at $t = 2$ s; then, the stimulus is off for 2 s, after which the second stimulus emerges from $t = 4$ s. The KLD experiences another significant increase shortly after the presence of the second stimulus. We make the following observations from Fig. 8. First, the features do not contribute equally in terms of user classification. The features identified earlier tend to play a more significant role than the ones identified later. Second, the involvement of multiple stimuli introduces more diversity in the pupillary response features and thus benefits classification accuracy.

Problem formulation. In the following, we discuss how to design auditory stimuli. An optimization problem is formulated, where the objective is to maximize the overall KLD in the corresponding pupillary response while keeping the entire authentication time within a practical threshold T_0, which sets a hard constraint on the authentication time. Formally, the optimization problem is expressed as follows

$$\begin{aligned} \max \quad & D_{KL}(P||Q) \\ \text{s.t.} \quad & \sum_{i=1}^{N} (t_i + \tau) \times m_i \leq T_0 \\ & m_i \in \{0, 1\} \end{aligned} \tag{2}$$

We aim to select a couple of (e.g., 2-4) auditory stimuli from the pool of size N, i.e., N different audio tracks. The binary variable m_i equals 1 if the i-th stimulus is picked and 0 otherwise. The stimuli selection is necessary as the pupil reacts differently toward various stimuli, as evidenced by our measurement study. Some stimuli are more effective in eliciting distinct patterns in pupillary responses than others. The variable t_i stands for the duration of the i-th stimulus. τ is a constant representing the interval duration between two adjacent stimuli, which manages the tradeoff between accuracy (the possibility that the pupil has returned to its baseline at the next stimulus onset) and usability (reasonable authentication time). After closely inspecting our collected data, we set it to 2 s empirically with 1.5% outliers. $\sum_{i=1}^{N}(t_i+\tau)\times m_i$ is thus the authentication time. The variables in the above-mentioned op-

Algorithm 1: Two-stage heuristic algorithm

```
input : $P_i$ ($i \in [1, N]$) and $Q$
output: Solution of $\boldsymbol{m}$ and $\boldsymbol{t}$
Calculate KLD for each stimulus $i$;
$K = \lceil \frac{T_0}{\tau} \rceil$;
Select top-$K$ stimuli with highest KLD;
for $j = 1$ to $2^K - 1$ do
    Formulate the simplified (2) given the $j$-th stimuli
      combination;
    Solve it via the approximate gradient descent
      algorithm;
Pick the stimuli combination with the highest KLD;
The corresponding $\boldsymbol{m}$ and $\boldsymbol{t}$ serve as the final solution.
```

timization problem include t_i's and m_i's, $i \in [1, N]$. Note that the problem formulation is user-specific, because the feature distribution in each individual's pupillary response is diverse. Correspondingly, the solutions of t_i's and m_i's are different across users; that is, each user is associated with a diverse optimum stimuli set and its duration. The problem formulation and calculation are performed during the enrollment stage.

A heuristic algorithm. The objective function and constraint of the above optimization problem are both non-linear. Besides, the two variable sets $\boldsymbol{m}$ and $\boldsymbol{t}$ are linked to each other. Hence, it is impractical to optimally solve it directly. In the following, we propose a heuristic algorithm to find the approximate solution with high computational efficiency. The algorithm is composed of two stages, each fixing the value of $\boldsymbol{m}$ and $\boldsymbol{t}$, respectively. The algorithm takes P_i ($i \in [1, N]$) and Q as inputs, where P_i is the user's feature distribution in the pupillary response under stimulus i and Q is the feature distribution of all reference users. In the first stage, we rank the KLD of each stimulus and select K candidate stimuli that generate the highest KLD. Here K is calculated as $\lceil \frac{T_0}{\tau} \rceil$. It represents the maximum number of stimuli that can be accommodated within T_0. Recall that τ is the interval duration between two adjacent stimuli. In the second stage, we exhaustively search for the maximum KLD among $2^K - 1$ possible stimuli combinations. To this end, we calculate KLD for each stimuli combination. Since m_i's are fixed under each combination as a result of the first stage, the original optimization problem is significantly simplified with t_i's as the only variables. Now the remaining question is how to solve the simplified optimization problem. For this, we employ the *approximate gradient descent* (AGD) algorithm [11, 59, 103]. It is an iterative method and useful especially when the derivative is hard to derive directly as in our case. The AGD algorithm finds an approximate solution for t_i's.

Dealing with long-term biometric changes. Like other biometrics, the auditory-pupillary response may exhibit variations over time [76, 86]. As a result, it can make the template acquired during the enrollment stage poorly representative of newly collected data samples, leading to degraded authentication performances. This phenomenon is known as *template aging* [42]. Many strategies have been developed to account for this issue [78, 81, 86]. Their main idea is to consistently update the classification model with new samples. In this work, we follow the existing approach to tackle the possible biometric pattern changes in the pupillary response. The core idea is to retrain the classification model with new samples from successful authentication trials. Our key steps are summarized as follows. 1) The system maintains a training dataset (reference set) of a fixed size after initial enrollment. The optimum training size is determined by the classifier, which is investigated in Section VII-A. Like traditional passwords, this training dataset is securely stored in the device. 2) When a new authentication sample arrives, it is labeled legitimate if the authentication is successful. 3) The dataset is updated with new samples in a first-in-first-out manner: these new samples are added into the reference set while the same number of outdated samples is discarded in the meantime. 4) The classification model is retrained over the updated dataset each several days or even more frequently, depending on the authentication frequency of the user. Since lightweight classifiers are employed in the proposed authentication scheme, the corresponding computation overhead of training is minimal. Note that there are even more sophisticated adaptive mechanisms (e.g., [55, 64, 77]). We plan to integrate them into our design in future studies.

V. Security Analysis

A. *Robustness Against Attacks*

We primarily consider the impersonation attack throughout this work. The adversary intends to use its biometric credential, i.e., pupillary response, under the auditory stimuli to get authenticated. To launch the attack, the attacker is assumed to have physical access to the victim's VR headset. It happens, for example, when the device is lost/stolen or temporarily possessed by the victim's roommate. Our evaluation results show that the success rate of such attacks is merely 0.76% on average. The robustness of SoundLock against the impersonation attack will be presented with details in Section VII-B.

Like other biometric methods, adversaries can also attack SoundLock via the *replay attack*, where the adversary injects a previously recorded sample of the pupillary response. Such an attack is extremely difficult to perform in our case. As the user's eyes are fully covered by the VR headset, it is impossible to record the target's pupillary response externally. On the other hand, it is possible for the adversary to access the victim's pupillary response samples via, say, pre-installing malware to the headset. Luckily, our scheme adopts the challenge-response authentication framework. With the interactive property, the attacker should know the auditory stimuli in advance to output the timely and correct response from the list of pre-recorded samples. It renders the attack very difficult to execute. Moreover, we argue that the device would be faced with an even more severe situation, if malware is pre-installed with access to the on-device authentication database.

Recent studies have also shown the feasibility of fabricating fake fingers and faces to bypass biometric authentication [13, 30, 110]. They are considered as a special kind of *mimicry attacks*. This attack is almost impossible to execute in our case, as the fabricated eyeball should be able to react to specific auditory stimuli. The pupil changes are subtle, smooth, dynamic, and unique to each individual. It is of great challenge, if not impossible, to build a mechanical device to mimic pupil dilation and constriction precisely. We are aware of some bionic eyes, which are essentially miniature cameras with necessary HCIs to optic nerves. Still, there is no "pupil"

TABLE I: Entropy of various authentication methods.

Work	Authentication method	Entropy (bits)
Wang et al. [108]	Password	$20 - 23$
Wang et al. [107]	PIN (4-digit[1], 6-digit[2])	8.41[1], 13.21[2]
Sae-Bae et al. [87]	Keystroke	$3.48 - 4.62$
Youmaran et al. [114]	Iris	$278 - 288$
Takahashi et al. [99]	Fingerprint	18.6
Adler et al. [7]	Face	$37.0 - 55.6$
SoundLock (this work)	Pupillometry	81

in bionic eyes. Besides, it costs around $150,000, which is extremely costly to deploy [104].

It is also possible that the auditory-pupillary response may be leaked, say, because of using a malicious (or compromised) device. Luckily, this new kind of biometric is revocable. In the case of having one pupillary response stolen or counterfeited, a new credential can be easily generated by changing its associated stimulus. It is also one of the prominent advantages of adopting auditory-pupillary response over other conventional biometrics for authentication.

B. *Entropy Analysis*

Entropy has been widely adopted to evaluate the security strength of authentication methods such as passwords [108] and PINs [107]. It is a measure of uncertainty in a random variable [29]. The classic entropy of a variable x with the distribution $P(x)$ is defined as $H = -\sum_{x \in X} P(x) \log P(x)$. In the context of biometric systems, however, the classic entropy overlooks intra-user variability by assuming each user has fixed biometric features and overestimates biometric information [97]. To tackle this issue, some prior works adopt an alternative metric *relative entropy* to measure the security of a biometric system [7, 97, 114]. We thus consider this metric too. Relative entropy is defined as the decrease in uncertainty about a person's identity due to a set of biometric features measurements [7]. It is measured under the framework of KLD, $K = \sum_{x \in X} P(x) \log \left(\dfrac{P(x)}{Q(x)} \right)$, where $P(x)$, $Q(x)$, and X represent the feature distribution of the target user, that of the reference set, and the feature space. It quantifies how much a single user's biometric feature distributions diverge from those of the population. It is noteworthy that the dataset plays an important role in the entropy computation as it defines the feature distributions $P(x)$ and $Q(x)$. According to the samples and their feature distributions collected in our dataset, K is calculated as 81 bits on average. Table I shows the relative entropy of SoundLock, keystroke, iris, fingerprint, and face, and the classic entropy of password and PIN. We can tell from the equations of these two kinds of entropy that classic entropy is an upper bound of relative entropy. In other words, the latter is a more conservative measure of authentication system security than the former [97]. The result shows that even the relative entropy of SoundLock (81 bits) largely exceeds those of passwords and PINs. SoundLock ranks second among all methods. It implies that dynamic pupillary response bears high uncertainty in the biometric information across individuals. It thus serves as a promising biometric for user identification. While the iris is associated with the highest relative entropy, the iris scanner is prohibitively costly to equip to a wide spectrum of VR devices.

VI. Experiment Methodology

A. *Experiment Setup*

Apparatus. We perform all experiments using an HTC VIVE Pro VR headset tethered to an Exxact TensorEX 1x Intel Core X-Series processor workstation. A Pupil Labs eye tracker is integrated into the VR headset. All virtual scenes and the prototype of SoundLock are implemented using Unity, a cross-platform engine for VR development, and scripted in C# and Python. The prototype is developed to render stimuli and capture the pupillary response (i.e., time-series pupil size) through the eye tracker's API. It includes functions of enrollment, optimization, authentication, and device lock/recovery.

Experiment setup. Before the experiment, participants receive an introduction to the concept of SoundLock as well as experimental instructions. After providing informed consent to take part in the study, they are asked to fill out a pre-study questionnaire based on the introduction to evaluate the expected usability of SoundLock. Then, participants are instructed to put on the VR headset. A student researcher assists in adjusting the device to ensure the wearing comfort and the correctness of eye tracker readings. Throughout the entire experiment, the lab environment is kept quiet by default. Next, the participant's pupillary response is recorded while performing the authentication tasks. Task details are presented in Section VI-B. There is a short break between authentication tasks. After the experiment, participants are asked to fill out a post-study questionnaire to evaluate the perceived usability of SoundLock through the tasks.

To facilitate evaluation, we adopt several commonly used metrics: false acceptance rate (FAR), false rejection rate (FRR), equal error rate (EER), F1-score, and area under the ROC curve (AUC).

B. *Experiment Design*

The entire experiment consists of two phases: a pilot study and an in-field study.

Pilot study. The purpose of the pilot study is a) to collect preliminary data for the measurement study (see Section II-B), b) to select from the candidate classifiers the one with the best overall performance, and c) to fix the classifier's training size and hyperparameters. In the pilot study, each participant is asked to listen to a set of 20 auditory stimuli consecutively. Their corresponding pupillary response is recorded. The auditory stimuli include white noise, monotones, prompt sounds, natural sounds, and human voices. Each stimulus is a 6-second audio track. Each stimulus is repeated 20 times for all participants. With the collected dataset, we carefully tune the training size and hyperparameters of each candidate classifier proposed in Section III-C. Then, we compare all candidate classifiers and select the one with the best performance. Results will be discussed in Section VII-A.

In-field study. The in-field study is performed with the prototype, the authentication system implemented with the best classifier as discussed above. The purpose is to evaluate the security and usability of the SoundLock prototype. Participants are asked to perform the following experiment tasks.

- *Enrollment*: Each participant is presented with a set of 20 auditory stimuli, with each stimulus 5 times. Their auditory-pupillary responses are recorded. All the samples are used to train the classifier as well as to determine the user-specific stimuli via the algorithm introduced in Section IV. In this way, each participant's biometric credential is enrolled.

- *Authentication*: In this task, the user-specific stimuli sequence is presented to the participant. Access is granted if the newly recorded pupillary response is classified as a legitimate one. A participant has three chances to pass the authentication. It is deemed successful if the biometric credential is verified in at least one in three trials.
- *Impersonation attack*: In this task, participants are asked to perform impersonation attacks. The attacker intends to use its own biometric credential, i.e., pupillary response, under the auditory stimuli to get authenticated. Specifically, each participant is randomly assigned with three other participants' biometrics to mimic. The attacker is presented with the victim's customized stimuli. The attack is deemed successful if the attacker gets authenticated in any one of three consecutive trials.
- Participants are asked to repeat the authentication task in a few follow-up sessions to examine the scheme performance under various conditions (see Section VII-C). Specifically, to investigate the impact of user motion, participants are asked to perform authentication under four types of motions: static (baseline), eye movement, head rotation, and body stretch. To evaluate the SoundLock performance across different time of day, a series of sessions are scheduled for the same group of people from 10 AM to 6 PM, with a 2-hour interval in between. To examine the impact of visual fatigue, authentication tasks are also conducted as participants are exposed to the VR environments for different time duration. We further carry out a longitudinal study. Participants are re-invited 7 days and 14 days after the main session to repeat the tasks. The purpose is to show if auditory-pupillary response as a biometric credential would change over time.

Attendance and time consumption. A total of 32 participants completed the pilot study. The average time spent is around 60 min, including 50 min for displaying all auditory stimuli samples and data recording with 10 min overhead. In the in-field study, 44 participants completed the main session, which consists of the enrollment, authentication, and impersonation attack tasks. They all participated in the impact of the user motion and the visual fatigue sessions right after the main session. The above sessions take around 50 min including necessary overhead, such as Q&A and reading/signing the consent form. 25 of them participated in the impact of time session. 28 and 18 of them completed the 7-day and 14-day longitudinal study, respectively. A user study is conducted; it consists of a pre-study and a post-study before and after the main session, respectively.

C. Recruitment and Ethical Aspects

Participant recruitment and demographics. The participants are recruited and informed through emails, social media postings (departmental Facebook website), and verbal communications. When a participant shows interest in participating in our study, we provide him/her a screening questionnaire asking about age, gender, race, and hearing and visual abilities. We screen participants with no hearing and visual impairments (corrected hearing ability with hearing aids and corrected visual ability with glasses and contact lenses will be allowed). Efforts have been made to recruit a diversified population based on age, gender, and race. After that, the participants are officially invited and asked to schedule a time and date with the researchers for the study. A total of 44 participants are recruited. They are all college students, faculty, and staff, aged between 17 and 40. Their demographic information is given in Table II. Each phase takes around 1 hour on average. Participants are compensated at a rate of $10 per hour.

Ethical aspects. The participants are provided with the Informed Consent document before the study. The document provides a detailed description of the study's procedure, benefits/risks, intentions, compensation, possible risks/discomforts, and rights. In order to make sure that participants are aware of the study procedures, the research team reads the summarized and important contents of the consent document at the beginning of each experiment and answers any questions the participant may have. The consent document is signed in person when the participants are in the lab. Subjects have the option to decide whether to participate in the experiments or not. During the experiment, they are free to take a break or quit at any time without penalty. They can ask any questions related to this research. The research team signs a confidentiality agreement with the participants regarding the protection of their biometric data, which are anonymized and securely stored, and will only be used for the purpose of this research. The entire study is IRB-approved.

VII. Results

A. Pilot Study–Classifier Selection

In the pilot study, the objective is to examine the candidate classification models and select the one that fits our scenario the best. The results will be used in the prototype development.

Classification model comparison. We implement different classification models as discussed in Section III-C, namely k-NN, OC-SVM, B-SVM, LR, GNB, and RF. 10-fold cross-validation is performed with the collected dataset. Specifically, the dataset is randomly split into two subsets, a training set and a testing set. Then, the classifier is trained and tested, with each user iteratively regarded as legitimate and the rest being imposters. This process is repeated 10 folds to prevent overfitting. We plot FAR and FRR in Fig. 10 by tuning the hyperparameters of the classification models.

For k-NN, it measures the distance between the testing sample and k training samples and compares it to a threshold α: if the distance is below α, the testing sample is deemed legitimate; otherwise, it is adversarial. Therefore, a larger α implies a looser detection rule that more likely considers an input sample legitimate and vice versa. By controlling the hyperparameter α, i.e., the distance threshold, we obtain the EER of k-NN equal to 1.5% at $\alpha = 1.0$ (see Fig. 10a).

For SVM, its idea is to find an optimal hyper-plane in high-dimensional space to perform classification. We adopt the radial basis function (RBF) kernel, a popular kernelized function, to transform the non-linear data to higher dimensions. A critical hyperparameter for the RBF kernel is γ, the standard deviation of the kernel function that defines the decision boundary qualitatively; a larger γ indicates a more relaxed

TABLE II: Participants' demographics. Auth exp* is about the participant's prior experience with authentication on VR.

Gender	#	%	Age	#	%	Iris color	#	%	Eye wear type	#	%	VR usage	#	%	Auth exp*	#	%
Female	16	37	≤18	4	9	Brown	34	77	None	28	64	Frequent	5	11	Proficient	3	7
Male	27	61	19-24	24	55	Hazel	6	14	Glasses	13	29	Occasional	8	18	Limited	5	11
Other	1	2	25-30	12	27	Blue	2	5	Contact lenses	3	7	Rare	13	30	None	36	82
			31-36	3	7	Green	1	2				Never	18	41			
			≥37	1	2	Other	1	2									

TABLE III: Performance comparison among different classification models.

Classification type	Model	EER (%)	F1-score	AUC
One-class	k-NN	**1.5**	**0.983**	**0.996**
	OC-SVM	3.4	0.956	0.989
Binary	B-SVM	4.3	0.935	0.986
	LR	4.6	0.929	0.990
	GNB	7.8	0.909	0.956
	RF	3.9	0.944	0.984

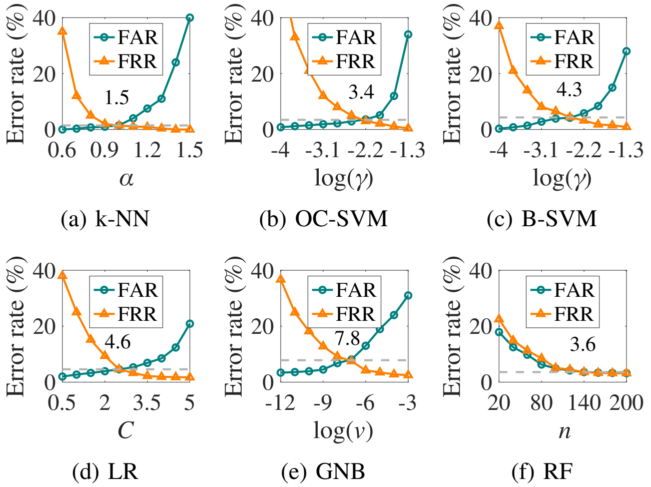
decision criterion to avoid the hazard of overfitting, resulting in a higher possibility that the input is accepted; a smaller γ implies a strict and sharp decision boundary. Figure 10b (10c) illustrates the FAR and FRR of the OC-SVM (B-SVM) as γ changes, with other parameters optimized. We find the lowest EER for OC-SVM as 3.4% at $\gamma = 6.3 \times 10^{-3}$. For B-SVM, the lowest EER is 4.3%, obtained at $\gamma = 3.2 \times 10^{-3}$.

Similarly, for LR, which uses a logistic function to model the dependent variable to generate a classification output, an essential hyperparameter is C, the inverse of regularization strength; a larger C corresponds to less regularization and vice versa. As depicted in Fig. 10d, the lowest EER of LR is obtained as 4.6% by tuning C to be 2.5.

As a widely adopted probabilistic machine learning algorithm, GNB works by calculating each data point and assigning the point to the higher class probability that it belongs to. An important hyperparameter is the variance smoothing v, which indicates the portion of the largest variance of all features added to variances for calculation stability. By setting $v = 10^{-7}$, we obtain the lowest EER of GNB as 7.8%, as shown in Fig. 10e.

RF consists of many decision trees and uses bagging and feature randomness when building each tree to create an uncorrelated forest of trees whose prediction by committee is the most accurate. An important hyperparameter is n, the number of trees. A larger n leads to more accurate predictions at the cost of higher computation time and power consumption. We plot in Fig. 10f the FAR and FRR curves as a function of the n. We find the EER converges to 3.6% as n approaches 140.

Table III compares all the classification models in terms of EER, F1-score, and AUC. Among them, k-NN produces the optimal FAR-FRR tradeoff with the lowest EER of 1.5% as well as the highest F1-score (0.983) and AUC (0.996). Its superior performance is primarily due to its robustness with respect to the data size. Compared with other models that generally require a large training dataset, k-NN better fits our scenario, where only a limited number of training samples (around 5) are collected.



(a) k-NN (b) OC-SVM (c) B-SVM

(d) LR (e) GNB (f) RF

Fig. 10: FAR, FRR, and EER of each classification model.

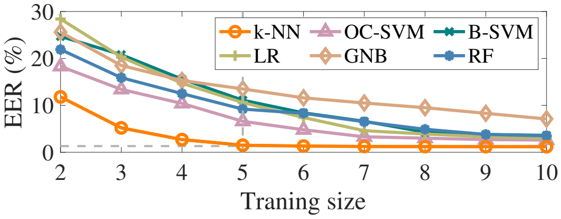


Fig. 11: Impact of training data size on EER.

Training data size. Fig. 11 shows the EER with respect to the training data size, i.e., the number of enrolled samples. Given the same training data size, k-NN achieves the lowest EER among the six classifiers, while GNB exhibits the worst performance. This is because the latter relies on the assumption that each class follows a Gaussian distribution. A larger dataset is thus needed to properly model the distribution. Empirical study indicates that it typically takes at least tens to hundreds of samples, depending on the task, to deliver a satisfying performance [10, 68, 74]. In contrast, only 5 samples are needed for k-NN to obtain EER as low as 1.5%. It indicates that k-NN attains a promising authentication accuracy with much fewer training samples.

To sum up, k-NN outperforms the other five models in classification accuracy, given the same training data size in our case. More importantly, it takes as few as 5 samples to sufficiently train the classifier. Hence, the enrollment stage can be performed efficiently.

B. *In-field Study–System Performance*

As a proof-of-concept implementation, we develop the prototype of SoundLock. Motivated by the results from the

TABLE IV: Performance comparison with state-of-the-art schemes. *Values are obtained from [36].

Approach	FAR (%)	FRR (%)	F1-score	Auth time
PIN*	-	>1.14	-	2.54-2.95
Drawing pattern*	-	>5.19	-	2.82-3.87
OcuLock [56]	3.55	3.55	0.983	≤10
SkullConduct [88]	6.90	6.90	-	≤23
Brain Password [51]	2.50	2.50	0.955	≈4.80
ElectricAuth [25]	0.83	2.00	-	≈**1.30**
SoundLock (this work)	**0.76**	**0.91**	**0.984**	≤7

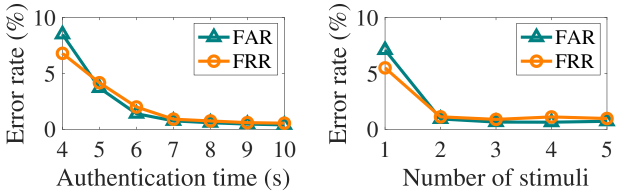
pilot study, we implement k-NN as the classifier and fix its hyperparameters as discussed. A total of five training samples are collected from each participant in the enrollment stage. A series of in-field tests are conducted to evaluate the system's performance.

Authentication accuracy vs. authentication time. We first examine the authentication accuracy of SoundLock with respect to authentication time in Fig. 12a. Authentication time is defined as the span from stimulus onset until the response cutoff. In other words, it includes the time to present stimuli and the time for the pupil to react. Both FAR and FRR drop given a longer authentication time. This is because more features are extracted and thus enhance the classification accuracy. We also notice that the benefit of a longer duration becomes marginal if it is beyond 7 s, with the average FAR and FRR as low as 0.76% and 0.91%, respectively. Fig. 12b depicts the authentication accuracy by adopting different numbers of stimuli; the error rate decreases with more stimuli presented. It complies with the result in Fig. 8–more stimuli enhance the distinguishability of the target user. Based on these observations, we adopt the multi-stimuli scheme and set the authentication duration threshold T_0 as 7 s in the optimization formulation to strike a balance between security and usability. Table IV compares the authentication time between SoundLock and existing works. Classic PINs and drawing patterns generally require a shorter time according to evaluation results from [36]. However, it demands relatively high motor skills for users to quickly enter these credentials in VR. They have been criticized as unfriendly to the elderly population and new users. Besides, relying on visual cues may hinder their usage for people with visual impairments. Among biometric schemes, SoundLock exhibits reasonable authentication time. Note that all these schemes need extra sensors, such as an EEG, to acquire the biometric signals.

Authentication accuracy comparison with state-of-the-arts. We further compare overall authentication accuracy between SoundLock and state-of-the-art solutions. Table IV shows that SoundLock almost achieves the best performance among all biometric methods in terms of FAR (0.76%), FRR (0.91%), and F1-score (0.984). Besides, it outperforms PIN and drawing pattern in FRR. It means a legitimate user gets denied by PIN or drawing pattern at a higher chance. This is because these two methods require necessary motor skills to perform especially on VR terminals. Errors would occur during credential entry when controllers are not operated properly.

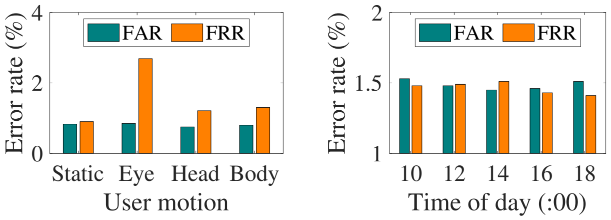
C. Performance Under Various Scenarios

In practical scenarios, a user may perform the authentication under different conditions, such as motion, time of day, and exposure time to VR environments. It is critical to evaluate if SoundLock is susceptible to these conditions.



(a) (b)

Fig. 12: Authentication performance.

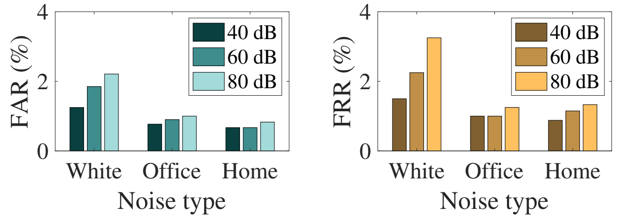


(a) Impact of user motion (b) Impact of time of day

Fig. 13: Performance under various conditions.

Impact of user motion. Since a user may make some movements during the authentication, it is important to show that the proposed scheme is motion-insensitive. In the experiments, participants are asked to perform four types of motions: sitting (static), eye movement, head rotation, and body stretch. The corresponding authentication accuracy is depicted in Fig. 13a. We find that the best performance is achieved at the static status with averaged FAR = 0.83% and FRR = 0.90%. Eye movement is associated with the highest FRR. This is because it introduces errors in the eye tracker calibrating the pupil size. Still, the authentication accuracy is practically acceptable with FAR = 0.85% and FRR = 2.69%. Based on the above results, users would be recommended to minimize their eye movement for the login duration. Other moving actions such as head rotation and body stretch also marginally increase the FRR, possibly due to the slight displacement of the eye tracker. Nevertheless, the increase is negligible; besides, the FAR remains consistent among various types of user motions ($0.80 \pm 0.05\%$), which suggests that user motions would not impact the security property of SoundLock.

Impact of noisy environments. We evaluate the impact of ambient noise on the performance of SoundLock. Three kinds of noises have been considered: white noise, office noise, and home noise. In particular, the white noise is synthesized with all the audible frequencies at the same intensity. The office noise is composed of people chatting, typing, phone ringing, computer fans, paperwork, etc. The home noise is a mixture of air conditioning, laundry, door locking, repairing, and TV sounds. All these soundtracks are downloaded from Mixkit [66]. In the experiments, the sounds are played as background noises by a pair of external speakers connecting to a second PC in the lab. We thus simulate the VR usage scenarios in generic, office, and home environments, respectively. Results are shown in Fig. 14. We find that the performance, FAR and FRR, degrades slightly as the sound level increases from 40 dB to 80 dB. Note that sound levels are in the 40-80 dB range in most offices and homes [27]. It indicates that the ambient noise does influence the pupillary response. On the other hand,



(a) FAR

(b) FRR

Fig. 14: Performance under various noise types and levels.

the influence is limited. Take home noise as an example. FAR = 0.67% and FRR = 0.88% as the sound level is 40 dB. They become 0.83% and 1.33%, respectively, if the noise is at 80 dB. It may be attributed to the fact that the stimulus audio is played via the VR headset, which is much closer to the user's ears than the noise sources. The former is thus dominant in the perceived sound.

Impact of time of day. We further examine if SoundLock is subject to the time of the day it is performed. A series of tests are scheduled over the same group of participants from 10 AM until 6 PM, with a 2-hour interval in between. We find in Fig. 13b that the performance is relatively stable throughout the day. To quantify the statistical difference in the FAR and FRR across different time of the day, we employ the *Kruskal-Wallis test* [46]. The test result indicates there is no significant difference on both FAR ($\chi^2 = 2.56$, $p > 0.05$) and FRR ($\chi^2 = 6.05$, $p > 0.05$) with respect to the time of the day.

Impact of exposure time to VR environments. It is well known that wearing VR for long periods can cause visual fatigue and motion sickness due to vergence-accommodation conflict [24]. It is therefore interesting to evaluate the performance of SoundLock with respect to users' exposure time to VR environments. In the experiment, each participant is asked to stay in the immersive environment for various periods of time, i.e., 10, 20, or 30 min, before performing the authentication. A participant can freely quit the experiment whenever they report discomfort or at any time they desire. In particular, users can choose to watch VR videos, play VR games, or browse online via the device. Table V summarizes the results. We observe that both FAR and FRR slightly increase under a long exposure time. The increase of FRR is relatively more prominent, by 0.74% from 0 min to 30 min. Conversely, FAR only sees a minor increase of 0.16% over time. This indicates the security of SoundLock is not influenced much, since incorrectly accepted adversarial authentications are limited; however, there is a moderately increasing chance that a legitimate user is wrongly classified. It indicates that pupillary response drifts slightly as the user is exposed to the VR environment for a while.

Longitudinal study. To investigate the long-term performance of SoundLock, participants are invited to attend two follow-up sessions, 7 days and 14 days after the main session, to repeat the authentication process. 28 and 18 participated in the two follow-up sessions, respectively. The adaptation strategy introduced in Section IV is adopted. For comparison, we also test in the last session the performance of SoundLock without adaptation. The result is summarized in Table VI. The error rate increases as time proceeds without adaptation, with FAR (FRR) rising from 0.79% (0.91%) to 8.89% (5.56%), after a 14-day duration. It implies that the biometrics, i.e., the auditory-pupillary response, drifts slowly over time. In comparison, the long-term performance becomes stable with the integration of our adaptation strategy. Specifically, the FAR (FRR) is 2.22% (1.48%), which merely exhibits a performance change of +1.46% (+0.57%). It suggests that our approach effectively deals with the temporal variation in pupillary response. Note that participants do not perform authentication in between sessions. We optimistically expect an even better long-term performance when SoundLock is under daily usage as the adaptation can be executed more frequently.

TABLE V: Performance under different exposure time to VR environments.

Expos. time	FAR (%)	FRR (%)
0 (baseline)	0.76	0.91
10 min	0.81	1.11
20 min	0.88	1.54
30 min	0.92	1.65

TABLE VI: Longitudinal study results. *Without the adaptation strategy.

Time span	FAR (%)	FRR (%)
0 (baseline)	0.76	0.91
7 days	1.19	2.14
14 days	2.22	1.48
14 days*	8.89	5.56

D. User Study

The goal of the user study is to evaluate the usability of SoundLock from participants' subjective perceptions.

Design. The study consists of a pre-study and a post-study, conducted before and after the main session of the experiment, respectively. To investigate the impact of the in-field experiments on user perception, the same questionnaire is used in both studies. In part-I of the questionnaire, all participants are asked to provide their perception of SoundLock by responding to 9 questions on a 5-point Likert scale (with 1 = strongly disagree and 5 = strongly agree). These questions cover multiple aspects of security and usability. Table VII lists all the questions. Part-II of the questionnaire includes three open-ended questions regarding overall experience *"What's your overall experience with SoundLock?"*, concerns *"Do you have any concerns or did you notice any potential issues of SoundLock?"*, and suggestions *"Do you have any suggestions to improve SoundLock in the future?"*.

Results. All 44 participants respond to the questions. Fig. 15b displays the distribution of answers to part-I questions in post-study. In general, participants express their satisfaction with SoundLock, especially in Q1 ($\mu = 4.32, \sigma = 0.97$, median = 5), Q2 ($\mu = 4.07, \sigma = 1.16$, median = 4), Q4 ($\mu = 4.20, \sigma = 0.92$, median = 4), Q5 ($\mu = 4.43, \sigma = 0.86$, median = 5), and Q6 ($\mu = 4.48, \sigma = 0.81$, median = 5). The least rated one is Q3 ($\mu = 3.55, \sigma = 1.25$, median = 4). As reported by several participants in the open-ended questions, this is caused by a couple of audio tracks in the stimuli pool. After close examination, listening discomfort is observed in audio tracks with bursting and high-pitch sound.

We then compare the survey results between the pre-study and the post-study using the Student's t-test [96], to investigate whether there is a significant statistical difference between the two studies. According to the test result, the most significant difference between the two studies lies in Q1 ($t(86) = 2.06, p = 0.021 < 0.05$), Q4 ($t(86) = 1.87, p = 0.032 < 0.05$), Q5 ($t(86) = 1.73, p = 0.044 < 0.05$), and Q6 ($t(86) = 1.70, p = 0.046 < 0.05$). Q7 has the least significant inter-study difference ($t(86) = 0.05, p = 0.480$). In general,

TABLE VII: Part-I questions.

	Question
Q1	SoundLock is a secure authentication scheme.
Q2	The authentication result is accurate.
Q3	There is no discomfort using SoundLock.
Q4	SoundLock is easy to use.
Q5	SoundLock is easy to learn.
Q6	SoundLock does not introduce much cognitive load.
Q7	The login time is acceptable.
Q8	SoundLock can be used on a daily basis.
Q9	I am willing to use SoundLock on my (future) VR device.

the average rating is higher in the post-study than the pre-study for all questions; even Q7 sees a slight improvement (μ_{pre} = 3.59 vs. μ_{post} = 3.61). These results indicate that SoundLock exceeds users' expectations after they have real experience with it. Q9 reflects the user's overall attitude towards SoundLock for real-world usage. The result for Q9 in the pre-study (post-study) is $\mu_{pre} = 3.75$, $\sigma_{pre} = 1.20$ ($\mu_{post} = 4.07$, $\sigma_{post} = 1.04$). Meanwhile, 63.6% (72.7%) users report a score larger than 3, i.e., agree or strongly agree, in the pre-study (post-study). This means that most users are willing to adopt SoundLock as the authentication method for VR devices. To summarize, SoundLock is well perceived by users, primarily due to its high security (Q1), ease to use (Q4), ease of learning (Q5), and low cognitive load (Q6).

Subjective feedback. A total of 24 participants respond to the open questions in part-II of the questionnaire. 13 participants leave feedback on the overall experience of SoundLock. Among them, 4 deem the authentication process in SoundLock to be fun, e.g., "*It was a fun experience!*" (P9). 3 appreciate the idea and logic behind SoundLock, e.g., "*The idea of using pupil for authentication is smart.*" (P35). 5 participants report satisfactory usability of SoundLock, e.g., "*I don't need to do anything and the authentication is automatically done.*" (P40).

Questions are raised by 9 participants. 3 of them question the robustness against consanguinity, e.g., "*Will twins or siblings be able to hack into each other's profile?*" (P35). This question is mainly due to the observation that identical twins or even siblings tend to share certain similar biometrics. Since there are no twins or siblings in our hired participants, we are unable to answer the question. We plan to investigate this as part of our future work with an extended group of subjects. 2 participants express privacy concerns, e.g., "*Will my pupillary response be used to infer what I'm thinking?*" (P38). Since the auditory-pupillary response is a reflex biometric, the pupillary response is stimulus-dependent. Basically, it reflects how human eyes react to an audio sound rather than the user's cortical processing, i.e., mental behavior. So far, we are unaware of any existing results on this topic. 3 participants mention some discomfort in listening to a couple of auditory stimuli with bursting and high-pitch sound. As a solution, we plan to investigate an even larger auditory stimuli pool in our future work. Volunteers will be invited to listen to and rate those stimuli. Any unpleasant ones will be discarded.

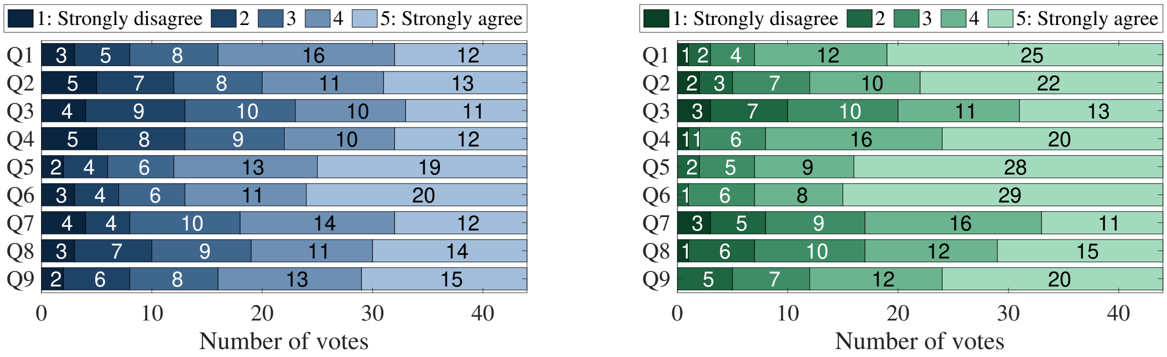
10 participants provide their comments for potential improvement. Among them, 3 suggest lowering the sound volume. It is worth mentioning that participants exhibit different tolerance of sound volume. 3 participants suggest combining with other forms of stimuli, such as colors, images, and videos. 4 propose to generalize SoundLock to the AR platform and other terminals, e.g., "*I think the system can be extended to smartphones, which will prove a valuable addition. The speaker can emit a sound and the eye image can be captured by the camera.*" (P1). Many of the comments are valid and inspire us with potential future work.

VIII. Related Work

User authentication on VR. While password and PIN serve as the most popular authentication mechanisms on VR devices [73], they have been criticized for these usability deficits: it takes users substantial effort to select correct letters/digits/characters from the virtual keyboard using the virtual laser extended from their controllers. Moreover, such authentication schemes have been proven highly vulnerable to shoulder-surfing attacks. Due to the occlusion of the VR headset, the user is unaware of the surroundings, rendering it easier for an adversary to acquire the entered credential through observation [32]. To address these issues, both industry and academia have been actively searching for practical alternatives. The existing methods can be broadly categorized into five classes: *knowledge-based methods* [6, 35, 37, 62, 116], *physiological biometrics* [9, 25, 51, 88], *behavioral biometrics* [56, 69, 75, 90, 117], *token-based methods* [23], and a mixture of them [61, 113, 118]. A recently published SoK paper provides an extensive discussion on this topic [95]. Please refer to it for more details. According to its discussion, physiological biometrics seem to outshine their peers so far due to their high usability and accuracy. Nevertheless, they bear at least two limitations for broad deployment. First, to capture users' biometric information, such as electroencephalogram (EEG), electrooculography (EOG), electrical muscle stimulation (EMS), and iris patterns, sophisticated sensors are required. For example, iris scan has been deployed for user authentication on HoloLens 2 [65], a high-end AR device costing at least $3,500. Due to its high price, the iris scanner is less likely to equip to general VR devices[1] in the near future [101]. Besides, biometrics are unique to an individual. Once such credentials are damaged or counterfeited, the user cannot cancel the pre-stored credentials or reset them with different biometric input. This property is also called *cancelability*. In contrast, our approach exhibits the following advantages. First, it is free from additional high-end sensing devices; instead, it only needs an eye tracker, which has been integrated into many prevalent commercial VR headsets [1–5]. It is well accepted that incorporating eye-tracking technology is a trend in VR to assist in simulating depth of field and focus and providing users with a more realistic and natural visual experience [26]. Second, auditory-pupillary responses are cancelable. In the case of having one pupillary response stolen, a new credential can be generated by changing its associated stimuli. A comprehensive comparison with prior user authentication schemes for VR is provided in Table IX in Appendix B.

Pupillary biometrics for user authentication. The idea of exploiting pupillometry for user authentication has been around for a decade [17, 22, 72]. Most efforts have been devoted to enhancing authentication accuracy. For example, Bednarik et al. [15] proposed combining pupillary biometrics with eye movements for user authentication. A similar idea is adopted in [31]. However, implementing these schemes is faced with several practical challenges: eye movements and

[1] The price of Meta Quest 2, the most popular VR device so far, ranges from $299 to $399.



(a) Pre-study results

(b) Post-study results

Fig. 15: Participants' subjective response distributions.

pupillary behaviors are task-dependent and light-sensitive. To overcome these limitations, researchers proposed leveraging pupillary light reflex (PLR) for user authentication [28, 111, 112]. PLR is an involuntary reaction of the human eyes to an external light stimulus: as a user is presented with lights of various combinations of chromas and intensities, her pupils will constrict and dilate accordingly. Typically, to elicit prominent and acute changes in pupil size (to create distinguishable features for classification), users are presented with strong light stimuli. It may lead to snow-blindness and flash-blindness effects [19, 34, 93]. Performing it on a daily basis could potentially bring severe health issues, e.g., temporary or even permanent vision impairment. Alternatively, SoundLock avoids the above concern as it employs auditory stimuli.

Challenge-response protocols for biometric authentication. Challenge-response has recently emerged as a popular authentication protocol and is frequently combined with biometrics for user identification. It leverages a user's physiological response to a given stimulus, i.e., challenge, injected by the interactive device. The assumption is that each user's response to a given challenge is unique. Examples of challenge-response biometrics include: the palm's/finger's response to vibrations [49, 52], EEG response to visual stimuli [9, 51, 109], or muscle response to electrical stimulation [25]. For example, Velody [49] makes use of the unique and nonlinear hand-surface vibration response for user identification. Similarly, VibWrite [52] enables user authentication via finger inputs on ubiquitous surfaces through physical vibration. It is implemented using a pair of vibration motors and a receiver that can be attached to any surface. Lin et al. [51] proposed a psychophysiological authentication protocol using carefully designed visual stimuli to acquire brain response signals. A similar idea is adopted in [9, 109]. Compared to conventional biometric authentication, the credentials created under challenge-response protocols are revocable–once a credential is counterfeited, it is convenient to reset it. Nonetheless, all the above schemes either rely on sophisticated sensors for response data acquisition or require actuators for challenge generation (e.g., motor vibrator), which do not exist in VR headsets. Hence, they are inapplicable here. Recently, reflexive eye behaviors in response to visual stimuli [91] have been exploited for user authentication. Their stimulus consists of presenting a single red dot on a dark screen that changes position multiple times. Then reflexive saccades are triggered; the distinctive gaze path is treated as the unique signature. This scheme requires explicit action, i.e., eye movement, from the user. Instead, SoundLock elicits users' involuntary pupil size changes in response to auditory stimuli with bare cognitive effort.

IX. Limitations and Future Work

In this section, we discuss several limitations of this work and present our future research directions.

Enrollment time. SoundLock is associated with a relatively long enrollment time. Under the current design, it ranges between 800 to 820 seconds. This is because SoundLock collects user's pupillary responses to the entire stimuli pool which consists of dozens of audio clips in the enrollment stage. The user-specific optimization is applied to find the best stimuli sequence for an individual. Note that the enrollment is only performed once for each user. To further reduce it, we can replace the current online user-specific optimization with offline optimization on the population scale, that is, an optimal stimuli sequence is derived for a large population group. In this way, only one stimuli sequence is rendered in the enrollment stage rather than the entire pool. The enrollment time would be substantially reduced accordingly. If a user's credential is counterfeited, a new stimuli sequence should be requested. As another possible approach, rather than presenting a user with the whole stimuli pool, we can reasonably present a subtset. We will carefully select the stimuli that generate the highest entropy among users. Besides, analysis is necessary to evaluate its impact on authentication accuracy.

Multi-modality stimuli. SoundLock only makes use of auditory stimuli. In fact, visual stimuli, such as lights, images, and moving objects, would also evoke pupillary response. In our future work, we plan to investigate biometric authentication methods exploiting multi-modality stimuli. Hopefully, it would introduce new feature dimensions and thus further enhance the system entropy. There are several research questions deserving thorough investigation. First, how to combine visual and auditory stimuli? There are at least two strategies, to display the two kinds of stimuli sequentially or concurrently. Different strategies would lead to distinctive pupillary response patterns (and thus entropy) and time efficiency. Second, under the new design, a new set of prominent and reliable features should be extracted from the raw data to optimize the accuracy. Third, the user-specific stimuli optimization will be revisited to balance security and usability with multi-modality stimuli.

Scalability. SoundLock has been tested among 44 subjects. In our future work, we plan to find out whether the proposed biometric works for a larger and more diverse population. Besides, the current benchmarking of system entropy is based on the dataset collected so far. With extended participation, the calculation result would reflect the ground truth better. Besides, SoundLock is only prototyped and evaluated on one kind of VR model (HTC VIVE Pro) and has been exclusively focused on the VR platform. Next, we plan to evaluate SoundLock on a broader set of VR headsets and examine the impact of device heterogeneity. Additionally, we will also examine the feasibility of generalizing our idea to other platforms, such as AR terminals and smartphones.

X. Conclusion

In this paper, we present SoundLock, a novel user authentication scheme designed for VR devices. SoundLock recognizes legitimate users by extracting carefully designed features from pupil size changes in response to auditory stimuli. We first introduce a basic scheme using a single stimulus, followed by an advanced scheme with multi-stimuli. A proof-of-concept prototype of SoundLock is implemented on a VIVE Pro VR headset. Extensive in-field experiments are performed involving 44 participants. Results show that SoundLock offers high authentication accuracy, which outperforms state-of-the-art biometric authentication solutions for VR. SoundLock also exhibits consistent performances under various testing conditions. Our user study reveals that SoundLock is well received; 72.7% of the participants are willing to adopt SoundLock as the authentication mechanism on their (future) VR devices.

References

[1] High-resolution vr headset for professionals - varjo vr-3. varjo.com/products/vr-3/, 2022.

[2] Htc vive pro eye. vive.com/eu/product/vive-pro-eye/overview/, 2022.

[3] Neo 2. pico-interactive.com/us/neo2.html, 2022.

[4] Neo 3 pro/pro eye. picoxr.com/us/neo3.html, 2022.

[5] Transform your business with eye tracking. fove-inc.com/, 2022.

[6] Abdelrahman et al. Cuevr: Studying the usability of cue-based authentication for virtual reality. In *Proc. 2022 Int. Conf. Adv. Vis. Interfaces*, pages 1–9, 2022.

[7] Adler et al. Towards a measure of biometric feature information. *Pattern Anal. Appl.*, 12(3):261–270, 2009.

[8] Altın. Comparison of different time and frequency domain feature extraction methods on elbow gesture's emg. *Eur. J. Interdiscip. Stud.*, 2(3):35–44, 2016.

[9] Arias-Cabarcos et al. Inexpensive brainwave authentication: new techniques and insights on user acceptance. In *30th USENIX Secur. Symp.*, pages 55–72, 2021.

[10] Arsioli et al. Machine learning applied to multifrequency data in astrophysics: blazar classification. *Mon. Notices Royal Astron. Soc.*, 498(2):1750–1764, 2020.

[11] Bai et al. From multitask gradient descent to gradient-free evolutionary multitasking: a proof of faster convergence. *IEEE Trans. Cybern.*, 2021.

[12] Barbur. Pupillary responses to grating stimuli. *J. Psychophysiol.*, 1991.

[13] Barral et al. Fake fingers in fingerprint recognition: Glycerin supersedes gelatin. In *Form. Pract. Secur.*, pages 57–69. Springer, 2009.

[14] Beatty. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychol. Bull.*, 91(2):276, 1982.

[15] Bednarik et al. Eye-movements as a biometric. In *Scandinavian Conf. Image Anal.*, pages 780–789, 2005.

[16] Bergamin et al. Latency of the pupil light reflex: sample rate, stimulus intensity, and variation in normal subjects. *Investig. Ophthalmol. Vis. Sci.*, 44(4):1546–1554, 2003.

[17] Boutros et al. Iris and periocular biometrics for head mounted displays: Segmentation, recognition, and synthetic data generation. *Image Vis. Comput.*, 104:104007, 2020.

[18] Encyclopedia Britannica. Iris. 2020.

[19] Brown. Flash blindness. *Am. J. Ophthalmol.*, 60(3):505–520, 1965.

[20] Cai et al. Touchlogger: Inferring keystrokes on touch screen from smartphone motion. In *6th USENIX Workshop Hot Top. Secur. (HotSec 11)*, 2011.

[21] Cantoni et al. Gant: Gaze analysis technique for human identification. *Pattern Recog.*, 48(4):1027–1038, 2015.

[22] Cascone et al. Pupil size as a soft biometrics for age and gender classification. *Pattern Recog. Lett.*, 140:238–244, 2020.

[23] Chan et al. Glass otp: Secure and convenient user authentication on google glass. In *Int. Conf. Financial Cryptography, Data Secur.*, pages 298–308, 2015.

[24] Chang et al. Virtual reality sickness: a review of causes and measurements. *Int. Journ. Hum.-Comput. Interact.*, 36(17):1658–1682, 2020.

[25] Chen et al. User authentication via electrical muscle stimulation. In *Proc. 2021 CHI Conf. Hum. Factors Comput. Syst.*, pages 1–15, 2021.

[26] Clay et al. Eye tracking in virtual reality. *J. Eye Mov. Res.*, 12(1), 2019.

[27] Cochary. Common noise levels – how loud is too loud? noiseawareness.org/info-center/common-noise-levels/, 2021.

[28] Da Costa et al. Dynamic features for iris recognition. *IEEE Trans. Syst, Man, Cybern.*, 42(4):1072–1082, 2012.

[29] Cover et al. *Elements of information theory*. Wiley-Interscience, 2006.

[30] Duc et al. Your face is not your password face authentication bypassing lenovo–asus–toshiba. *Black Hat Briefings*, 4:158, 2009.

[31] Eberz et al. 28 blinks later: Tackling practical challenges of eye movement biometrics. In *Proc. 2019 ACM SIGSAC Conf. Comput. Commun. Secur.*, pages 1187–1199, 2019.

[32] Eiband et al. Understanding shoulder surfing in the wild: Stories from users and observers. In *Proc. 2017 CHI Conf. Hum. Factors Comput. Syst.*, pages 4254–4265, 2017.

[33] Fink et al. From pre-processing to advanced dynamic modeling of pupil data. 2021.

[34] FotonVR. Is virtual reality headset harmful to the eyes? fotonvr.com/is-virtual-reality-headset-harmful-to-the-eyes/, 2020.

[35] Funk et al. Lookunlock: Using spatial-targets for user-authentication on hmds. In *Proc. 2019 CHI Conf. Hum. Factors Comput. Syst.*, pages 1–6, 2019.

[36] George et al. Seamless and secure vr: Adapting and

evaluating established authentication systems for virtual reality. NDSS, 2017.

[37] George et al. Investigating the third dimension for authentication in immersive virtual reality and in the real world. In *2019 IEEE Conf. Virtual Reality, 3D User Interfaces (VR)*, pages 277–285, 2019.

[38] Van Gerven et al. Memory load and the cognitive pupillary response in aging. *Psychophysiol.*, 41(2):167–174, 2004.

[39] HandWiki. Damped sine wave. 2022.

[40] Hoeks et al. Pupillary dilation as a measure of attention: A quantitative system analysis. *Behav. Res. Meth. Instrum. Comput.*, 25(1):16–26, 1993.

[41] Jain et al. Pupillary unrest correlates with arousal symptoms and motor signs in parkinson disease. *Mov. Disord.*, 26(7):1344–1347, 2011.

[42] Jain et al. 50 years of biometric research: Accomplishments, challenges, and opportunities. *Pattern Recog. Lett.*, 79:80–105, 2016.

[43] Joshi et al. Relationships between pupil diameter and neuronal activity in the locus coeruleus, colliculi, and cingulate cortex. *Neuron*, 89(1):221–234, 2016.

[44] Karlén. Eye-tracking is virtual reality's next frontier. venturebeat.com/games/eye-tracking-is-virtual-realitys-next-frontier/.

[45] Klingner et al. Measuring the task-evoked pupillary response with a remote eye tracker. In *Proc. 2008 Symp. Eye Tracking Res. Appl.*, pages 69–72, 2008.

[46] Kruskal et al. Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.*, 47(260):583–621, 1952.

[47] Kullback et al. On information and sufficiency. *Ann. Math. Stat.*, 22(1):79–86, 1951.

[48] Kupin et al. Task-driven biometric authentication of users in virtual reality (vr) environments. In *Int. Conf. Multimed. Model.*, pages 55–67, 2019.

[49] Li et al. Velody: Nonlinear vibration challenge-response for resilient user authentication. In *Proc. 2019 ACM SIGSAC Conf. Comput. Commun. Secur.*, pages 1201–1213, 2019.

[50] Liao et al. Correspondences among pupillary dilation response, subjective salience of sounds, and loudness. *Psychon. Bull. Rev.*, 23(2):412–425, 2016.

[51] Lin et al. Brain password: A secure and truly cancelable brain biometrics for smart headwear. In *Proc. 16th Annu. Int. Conf. Mob. Syst., Appl. Serv.*, pages 296–309, 2018.

[52] Liu et al. Vibwrite: Towards finger-input authentication on ubiquitous surfaces via physical vibration. In *Proc. 2017 ACM SIGSAC Conf. Comput. Commun. Secur.*, pages 73–87, 2017.

[53] Liu et al. Virtual reality and its application in military. In *IOP Conf. Ser.: Earth Environ. Sci.*, volume 170, page 032155, 2018.

[54] Lüdtke et al. Mathematical procedures in data recording and processing of pupillary fatigue waves. *Vis. Res.*, 38(19):2889–2896, 1998.

[55] Lumini et al. A clustering method for automatic biometric template selection. *Pattern Recog.*, 39(3):495–497, 2006.

[56] Luo et al. Oculock: Exploring human visual system for authentication in virtual reality head-mounted display. In *2020 Netw. Distrib. Syst. Secur. Symp. (NDSS)*, 2020.

[57] Lykstad et al. Neuroanatomy, pupillary dilation pathway. 2018.

[58] Mansfield et al. Best practices in testing and reporting performance of biometric devices. 2002.

[59] Mariño et al. An approximate gradient-descent method for joint parameter estimation and synchronization of coupled chaotic systems. *Phys. Lett. A*, 351(4-5):262–267, 2006.

[60] Martínez-Navarro et al. The influence of virtual reality in e-commerce. *J. Bus. Res.*, 100:475–482, 2019.

[61] Mathis et al. Knowledge-driven biometric authentication in virtual reality. In *Proc. 2020 CHI Conf. Hum. Factors Comput. Syst.*, pages 1–10, 2020.

[62] Mathis et al. Rubikauth: Fast and secure authentication in virtual reality. In *Proc. 2020 CHI Conf. Hum. Factors Comput. Syst.*, pages 1–9, 2020.

[63] S. Mathôt. Pupillometry: Psychology, physiology, and function. *J. Cogn.*, 1(1), 2018.

[64] Mhenni et al. Keystroke template update with adapted thresholds. In *2016 2nd Int. Conf. Adv. Technol. Signal Image Process. (ATSIP)*, pages 483–488. IEEE, 2016.

[65] Microsoft. Microsoft hololens — mixed reality technology for business. microsoft.com/en-us/hololens/, 2022.

[66] Mixkit. mixkit.co/, 2022.

[67] Moline. Virtual reality for health care: a survey. *Stud. Health Technol. Inform.*, pages 3–34, 1997.

[68] Murphy. Naive bayes classifiers. 18(60):1–8, 2006.

[69] Mustafa et al. Unsure how to authenticate on your vr headset? come on, use your head! In *Proc. 4th ACM Int. Workshop Secur. Priv. Anal.*, pages 23–30, 2018.

[70] Nakakoga et al. Pupillary response reflects attentional modulation to sound after emotional arousal. *Sci. Rep.*, 11(1):1–10, 2021.

[71] Nakamura. Measurement of pupillary unrest in eyestrain. *Jpn. J. Ophthalmol.*, 40(4):533–539, 1996.

[72] Nugrahaningsih et al. Pupil size as a biometric trait. In *Int. Workshop Biom. Auth.*, pages 222–233, 2014.

[73] Oculus. Work out your way with meta quest 2. meta.com/quest/products/quest-2/.

[74] Ontivero-Ortega et al. Fast gaussian naïve bayes for searchlight classification analysis. *Neuroimage*, 163:471–479, 2017.

[75] Pfeuffer et al. Behavioural biometrics in vr: Identifying people from body motion and relations in virtual reality. In *Proc. 2019 CHI Conf. Hum. Factors Comput. Syst.*, pages 1–12, 2019.

[76] Pisani et al. Adaptive biometric systems: Review and perspectives. *ACM Comput. Surv.*, 52(5):1–38, 2019.

[77] Poh et al. Model and score adaptation for biometric systems: Coping with device interoperability and changing acquisition conditions. In *2010 20th Int. Conf. Pattern Recog.*, pages 1229–1232. IEEE, 2010.

[78] Poh et al. Critical analysis of adaptive biometric systems. *IET Biom.*, 1(4):179–187, 2012.

[79] Purves et al. *Neurosciences*. De Boeck Supérieur, 2019.

[80] Rathgeb et al. A survey on biometric cryptosystems and cancelable biometrics. *Eurasip J. Inf. Secur.*, 2011(1):1–25, 2011.

[81] Rattani et al. Template update methods in adaptive biometric systems: A critical review. In *Int. Conf. Biom.*, pages 847–856, 2009.

[82] Reimer et al. Pupil fluctuations track rapid changes

in adrenergic and cholinergic activity in cortex. *Nat. Commun.*, 7(1):1–7, 2016.

[83] ReportLinker. Virtual reality market size, share trends analysis report, 2022 - 2030, 2022.

[84] Rigas et al. Biometric identification based on the eye movements and graph matching techniques. *Pattern Recog. Lett.*, 33(6):786–792, 2012.

[85] Van Rij et al. Analyzing the time course of pupillometric data. *Trends Hear.*, 23:2331216519832483, 2019.

[86] Roli et al. Adaptive biometric systems that can improve with use. *Adv. Biom.*, pages 447–471, 2008.

[87] Sae-Bae et al. Distinguishability of keystroke dynamic template. *PloS one*, 17(1):e0261291, 2022.

[88] Schneegass et al. Skullconduct: Biometric user identification on eyewear computers using bone conduction through the skull. In *Proc. 2016 CHI Conf. Hum. Factors Comput. Syst.*, pages 1379–1384, 2016.

[89] Schumann et al. Sympathetic and parasympathetic modulation of pupillary unrest. *Front. Neurosci.*, 14:178, 2020.

[90] Shen et al. Gaitlock: Protect virtual and augmented reality headsets using gait. *IEEE Trans. Dependable Secur. Comput.*, 16(3):484–497, 2018.

[91] Sluganovic et al. Using reflexive eye movements for fast challenge-response authentication. In *Proc. 2016 ACM SIGSAC Conf. Comput. Commun. Secur.*, pages 1056–1067, 2016.

[92] Smith et al. Single neuron activity in the pupillary system. *Brain Res.*, 24(2):219–234, 1970.

[93] Steam Community. Htc vive can cause snow blindness if you don't do this. steamcommunity.com/app/358040/discussions/0/365163686083238173/, 2016.

[94] Stein. Watching me, watching you: How eye tracking is coming to vr and beyond. cnet.com/tech/computing/watching-me-watching-you-how-eye-tracking-is-coming-to-vr-and-beyond/.

[95] Stephenson et al. Sok: Authentication in augmented and virtual reality. In *2022 IEEE Symp. Secur. Priv. (SP)*, pages 1552–1552, 2022.

[96] Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

[97] Sutcu et al. What is biometric information and how to measure it? In *2013 IEEE Int. Conf. Technol. Homeland Secur.*, pages 67–72, 2013.

[98] Szabadi. Functional organization of the sympathetic pathways controlling the pupil: light-inhibited and light-stimulated pathways. *Front. Neurol.*, 9:1069, 2018.

[99] Takahashi et al. A measure of information gained through biometric systems. *Image Vis. Comput.*, 32(12):1194–1203, 2014.

[100] Tekin et al. Static and dynamic pupillometry data of healthy individuals. *Clin. Exp. Optom.*, 101(5):659–665, 2018.

[101] Thakkar. Biometric devices: Cost, types and comparative analysis. 2017.

[102] Thompson. Vr applications: 23 industries using virtual reality. 2022.

[103] Vlatakis-Gkaragkounis et al. Efficiently avoiding saddle points with zero order methods: No gradients required. *Adv. Neural Inf. Process. Syst.*, 32, 2019.

[104] Wakefield. Bionic eyes: Obsolete tech leaves patients in the dark. bbc.com/news/technology-60416058/, 2022.

[105] Walls. The evolutionary history of eye movements. *Vis. Res.*, 2(1-4):69–80, 1962.

[106] Wang et al. Modulation of stimulus contrast on the human pupil orienting response. *Eur. J. Neurosci.*, 40(5):2822–2832, 2014.

[107] Wang et al. Understanding human-chosen pins: characteristics, distribution and security. In *Proc. 2017 ACM Asia Conf. Comput. Commun. Secur.*, pages 372–385, 2017.

[108] Wang et al. Zipf's law in passwords. *IEEE Trans. Inf. Forensics Secur.*, 12(11):2776–2791, 2017.

[109] Wu et al. An eeg-based person authentication system with open-set capability combining eye blinking signals. *Sens.*, 18(2):335, 2018.

[110] Xiao. Security issues in biometric authentication. In *Proc. 6th Annu. IEEE SMC Inf. Assur. Workshop*, pages 8–13, 2005.

[111] Yan et al. Using pupil light reflex for fast biometric authentication. In *Proc. ACM Turing Celeb. Conf.-China*, pages 139–143, 2020.

[112] Yano et al. Extraction and application of dynamic pupillometry features for biometric authentication. *Meas.*, 63:41–48, 2015.

[113] Yi et al. Glassgesture: Exploring head gesture interface of smart glasses. In *IEEE INFOCOM 2016 35th Annu. IEEE Int. Conf. Comput. Commun.*, pages 1–9, 2016.

[114] Youmaran et al. Measuring biometric sample quality in terms of biometric feature information in iris images. *J. Electr. Comput. Eng.*, 2012, 2012.

[115] Young et al. Pupil responses to foveal exchange of monochromatic lights. *J. Stoch. Anal.*, 70(6):697–706, 1980.

[116] Yu et al. An exploration of usable authentication mechanisms for virtual reality systems. In *2016 IEEE Asia Pacific Conf. Circuits Syst. (APCCAS)*, pages 458–460, 2016.

[117] Zhang et al. Continuous authentication using eye movement response of implicit visual stimuli. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(4):1–22, 2018.

[118] Zhu et al. Blinkey: A two-factor user authentication method for virtual reality devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(4):1–29, 2020.

Appendix A
Candidate Features

Table VIII lists all the 60 candidate features introduced in Section III-B, including their names, categories, phases, and notations. They are sorted by the normalized Fisher score as demonstrated in Fig. 16. The top 20 features are selected.

Appendix B
Comparison among User Authentication Schemes on VR

Table IX provides a comprehensive comparison among some representative user authentication schemes for VR. The existing schemes are categorized into knowledge-based authentication (white), physiological biometric authentication (light gray), behavioral biometric authentication (medium gray), and multi-factor authentication (dark gray). All schemes are compared from multiple aspects of usability and security.

TABLE VIII: List of all the 60 candidate features.

Index	Feature name	Category	Phase	Notation	Index	Feature name	Category	Notation
1	Response lag	Morphological	Excitation	rl	44	Average	Statistical	avg
2-3	Peak magnitudes	Morphological	Excitation	p_{1-2}	45	Maximum	Statistical	max
4-5	Valley magnitude	Morphological	Excitation	v_{1-2}	46	Minimum	Statistical	min
6-7	Dilation rates	Morphological	Excitation	dr_{1-2}	47	Variance	Statistical	var
8-9	Constriction rates	Morphological	Excitation	cr_{1-2}	48	Median	Statistical	med
10-19	Dilation polynomial coefficients	Morphological	Excitation	dpc_{1-10}	49	Root mean square	Statistical	rms
20-29	Constriction polynomial coefficients	Morphological	Excitation	cpc_{1-10}	50	Skewness	Statistical	$skew$
30	Area under curve	Morphological	Excitation	auc	51	Kurtosis	Statistical	$kurt$
31	Recovery time	Morphological	Recovery	rt	52	Roughness	Statistical	$rough$
32-35	Damped oscillation	Morphological	Recovery	do_{1-4}	53	Sharpness	Statistical	$sharp$
36-37	Pupillary unrest	Morphological	Recovery	pu_{1-2}	54	First quartile	Statistical	$q1$
38	Baseline average	Morphological	Recovery	$bavg$	55	Third quartile	Statistical	$q3$
39	Baseline maximum	Morphological	Recovery	$bmax$	56	Interquartile range	Statistical	iqr
40	Baseline minimum	Morphological	Recovery	$bmin$	57	Mean absolute deviation	Statistical	mad
41	Baseline variance	Morphological	Recovery	$bvar$	58	Slope sign change	Statistical	ssc
42	Baseline median.	Morphological	Recovery	$bmed$	59	Mean crossing	Statistical	mx
43	Baseline interquartile range	Morphological	Recovery	$biqr$	60	Willison amplitude	Statistical	wa

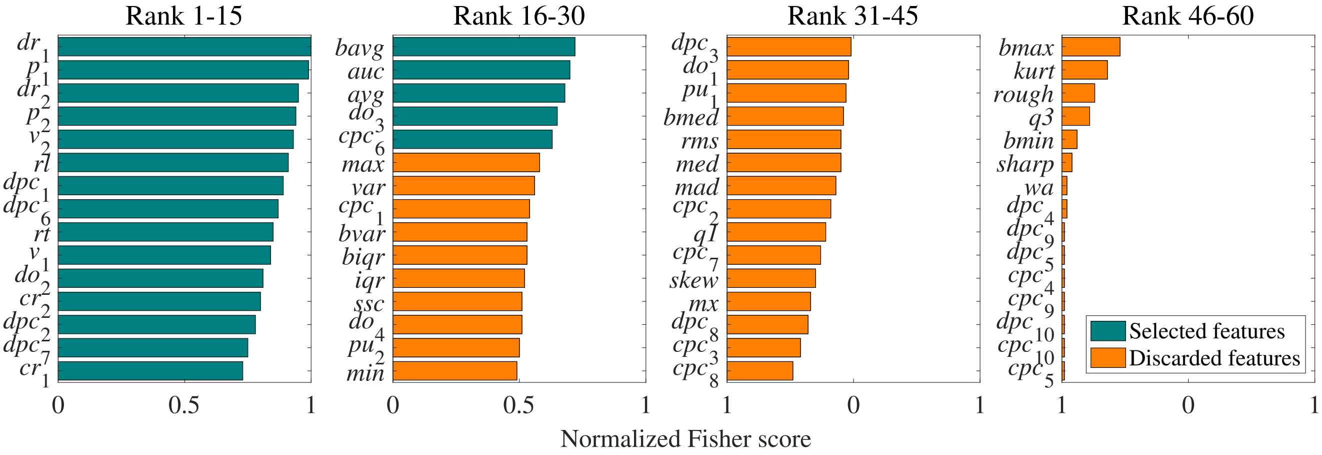


Fig. 16: All candidate features sorted by their normalized Fisher scores.

TABLE IX: Comparison among different user authentication approaches for VR. ●: method fulfills criterion. ◐: method quasi-fulfills criterion. ○: method does not fulfill criterion. –: not enough information.

Scheme	Extra sensor-free	Hand-free	Auth speed	Accuracy	Revocability	Against replay	Against shoulder-surfing	Against impersonation	Against guessing
PIN	●	○	⋆⋆⋆	⋆⋆	●	○	○	–	○
Drawing pattern	●	○	⋆⋆⋆	⋆⋆	●	○	○	–	○
3D pattern [116]	●	○	⋆	–	●	○	●	–	○
CueVR [6]	●	○	⋆⋆	⋆	●	○	●	–	○
LookUnlock [35]	●	●	⋆	–	●	○	◐	–	○
RoomLock [37]	●	○	⋆⋆	⋆⋆	●	○	◐	–	○
RubikAuth [62]	●	○	⋆⋆⋆	⋆⋆⋆	●	○	●	–	○
SkullConduct [88]	○	●	⋆	⋆⋆	○	○	●	●	●
Brain Password [51]	○	●	⋆⋆⋆	⋆⋆⋆	●	●	●	●	●
Arias-Cabarcos et al. [9]	○	●	⋆⋆	⋆	●	●	●	●	●
ElectricAuth [25]	○	○	⋆⋆⋆	⋆⋆⋆	●	●	●	●	●
SoundLock (this work)	●	●	⋆⋆	⋆⋆⋆	●	●	●	●	●
GaitLock [90]	●	●	⋆⋆⋆	⋆⋆⋆	○	○	◐	●	●
OcuLock [56]	○	●	⋆	⋆⋆⋆	○	●	●	●	●
Kupin et al. [48]	○	○	⋆⋆⋆	⋆⋆	○	–	–	●	●
Mustafa et al. [69]	●	●	–	⋆⋆	○	–	●	●	●
Pfeuffer et al. [75]	●	◐	–	⋆	○	○	○	●	●
Zhang et al. [117]	●	●	⋆⋆⋆	⋆⋆	–	–	●	●	●
GlassGesture [113]	●	●	–	⋆⋆⋆	●	–	●	●	●
RubikBiom [61]	●	○	⋆⋆⋆	⋆⋆	●	–	●	●	●
BlinKey [118]	●	●	⋆	⋆⋆⋆	●	–	●	●	●