Low-effort VR Headset User Authentication Using Head-reverberated Sounds with Replay Resistance

Ruxin Wang, Long Huang, Chen Wang
School of Electrical Engineering and Computer Science, Louisiana State University, Baton Rouge, LA, USA
rwang31@lsu.edu, lhuan45@lsu.edu, chenwang1@lsu.edu

2023 IEEE Symposium on Security and Privacy (SP) | 978-1-6654-9336-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/SP46215.2023.10179367

***Abstract*—While Virtual Reality (VR) applications are becoming increasingly common, efficiently verifying a VR device user before granting personal access is still a challenge. Existing VR authentication methods require users to enter PINs or draw graphical passwords using controllers. Though the entry is in the virtual space, it can be observed by others in proximity and is subject to critical security issues. Furthermore, the in-air hand movements or handheld controller-based authentications require active user participation and are not time-efficient. This work proposes a low-effort VR device authentication system based on the unique skull-reverberated sounds, which can be acquired when the user wears the VR device. Specifically, when the user puts on the VR device or is wearing it to log into an online account, the proposed system actively emits an ultrasonic signal to initiate the authentication session. The signal returning to the VR device's microphone has been reverberated by the user's head, which is unique in size, skull shape and mass. We thus extract head biometric information from the received signal for unobtrusive VR device authentication.**

Though active acoustic sensing has been broadly used on mobile devices, no prior work has ever successfully applied such techniques to commodity VR devices. Because VR devices are designed to provide users with virtual reality immersion, the echo sounds used for active sensing are unwanted and severely suppressed. The raw audio before this process is also not accessible without kernel/hardware modifications. Thus, our work further solves the challenge of active acoustic sensing under echo cancellation to enable deploying our system on off-the-shelf VR devices. Additionally, we show that the echo cancellation mechanism is naturally good to prevent acoustic replay attacks. The proposed system is developed based on an autoencoder and a convolutional neural network for biometric data extraction and recognition. Experiments with a standalone and a mobile phone VR headset show that our system efficiently verifies a user and is also replay-resistant.

***Index Terms*—Virtual Reality, Authentication, Biometric.**

I. Introduction

Virtual reality (VR) technology has gained widespread recognition over the past few years. Bridging digital and physical worlds, VR has been increasingly adopted in entertainment, education, medical care, and social networking to provide immersive experiences. The international data corporation predicts that the shipments of VR headsets will reach 28.6 million in 2025, with the compound annual growth rate anticipated at 41.4% [1]. In particular, the standalone VR headsets, such as the Meta (Oculus) Quest, account for the vast majority of shipments [2], which features built-in processors and storage and is wire-free. The recent COVID-19 pandemic further elevates the demand for VR technology in people's daily life.

Similar to other personal devices like smartphones, VR devices are closely connected with user privacy, including private app content, browsing histories, and preferences. However, many successes in mobile device authentication can not be easily copied to protect VR device/account access. Current solutions ask users to enter passwords with handheld controllers or in-air hand gestures, which are time-consuming and far less convenient than typing on a physical keyboard or touchscreen. Moreover, while a user inputs a password in the virtual space, the sensitive hand movements could leak the password to a nearby adversary in the physical world. Additionally, because a VR device may be shared with close friends and family members, practicing the multi-account service would be a nightmare, given the tedious and insecure VR authentication process. In this work, we design an authentication method for VR devices based on acoustic domain human-VR interactions, which is fast and easy to use.

There has been active work on extracting behavioral biometrics from VR users' hand/head/body motions for authentication. For example, Mathis *et al.* replace the traditional virtual keyboard with a 3D Rubik's cube for the user to enter a PIN, and the hand motion patterns such as moving, pointing, and button-click dynamics are captured by the handheld controllers for authentication [3], [4]. In addition to relying on hand motions, recent studies use VR headsets' inertial sensors to extract biometric information from head movements when the user performs pointing, grabbing, walking, and typing in the virtual space [5]–[7]. Wang *et al.* further develop an authentication mechanism to allow AR/VR users to unlock their accounts with a nodding action [8]. But these methods all require users to actively interact with the VR device, which is obtrusive and slow. The authentication performance is also limited by behavioral inconsistency and the low fidelity of embedded sensors.

This work employs the less obtrusive acoustic sensing technique to simplify the human-device interactions required for VR user authentication. We propose a low-effort VR user authentication system based on extracting the acoustic-domain head biometrics that are naturally born with head-mounted devices. Specifically, when the user puts on a VR device, a unique rigid body is formed by the user's head and the device, containing two chambers, the skull and the hollow space enclosed by the VR device and the face. When an authentication session is initiated, the proposed system emits an ultrasonic signal using the VR device's speakers. The signals traveling on

DOI 10.1109/SP46215.2023.00127

this rigid body are reverberated (e.g., damped and reflected) by the individually unique head size, skull shape, mass, and face pattern. The resulting signals reaching the microphone thus carry the user's biometric information and can be analyzed for authentication.

However, deploying active acoustic sensing on off-the-shelf VR headsets is not trivial. Though many works have demonstrated such sensing techniques on mobile devices [9]–[13], to our best knowledge, no prior work has successfully applied them to commodity VR devices. The main reason is that the raw audio data before echo cancellation can not be directly accessed on these devices without kernel and hardware modifications [14]. Because VR devices are designed to provide immersion, the echo sounds, from its speaker to mic, are unwanted and canceled by default [15], making the received audio feedback hardly recognizable for sensing purposes. Additionally, the current VR headset model designs, such as placing speakers and microphones under the device's fabric cloth and strap, cause further acoustic attenuation and noises.

To address the above challenges and achieve low-effort user authentication on off-the-shelf VR devices, we develop a head biometric-based authentication system based on a Convolutional AutoEncoder (CAE) and a Convolutional Neural Network (CNN). The system sends millisecond-level ultrasonic signals and takes echo sounds as the input. We derive the 2D spectrogram of mic data to measure the impact on the speaker-to-mic channel caused by the user's head. Furthermore, we develop the CAE-CNN algorithm for user authentication, which counteracts the device's built-in echo cancellation and encodes individual head biometrics. In particular, the CAE algorithm encodes the head biometric information based on reconstructing the spectrogram from the surviving echo sound. Next, our three-convolutional-layer CNN model learns the reconstructed head biometric spectrogram to distinguish users. Both the single-user verification and the multi-user identification services are supported. Access is granted only when the user's identity is claimed or belongs to one registered user. We further prevent acoustic replay attacks by leveraging both the CAE algorithm and the device's echo cancellation.

Our contributions can be summarized as follows:

- We propose an efficient and low-effort authentication system for current VR headsets leveraging head-reverberated sounds. The engineering contribution is allowing the already shipped VR headsets to identify a user through millisecond-level acoustic signals without kernel or hardware modifications.
- We find that a head-mounted VR device forms a unique rigid body with the user's head, and an actively emitted acoustic signal can measure the individual rigid body to extract head biometric information for authentication.
- This work, for the first time, achieves active acoustic sensing on commodity VR headsets and solves the challenge of obtaining recognizable acoustic information under echo cancellation. We further show that the existing echo cancellation mechanisms are naturally good to address acoustic replay threats.
- We develop a CAE-CNN algorithm to encode the head biometric for each individual from acoustic signals, counteract the built-in echo cancellation effect, reduce noise impacts to serve long-term use, and increase difficulties for replay attacks.
- Extensive experiments with a standalone and a mobile phone VR headset show that our system efficiently verifies a user and identifies multiple users while the acoustic replays are prevented.

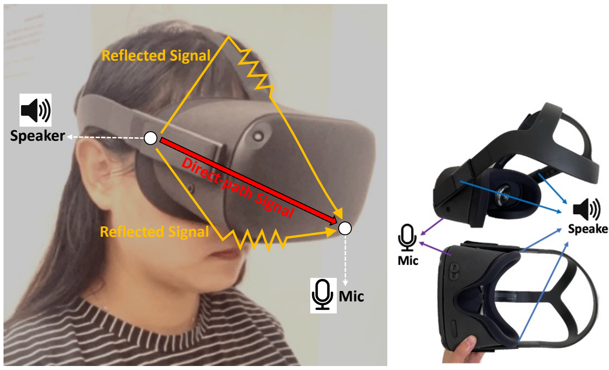


Fig. 1. Illustration of the acoustic signal interacting with the user's head.

II. Background and System Models

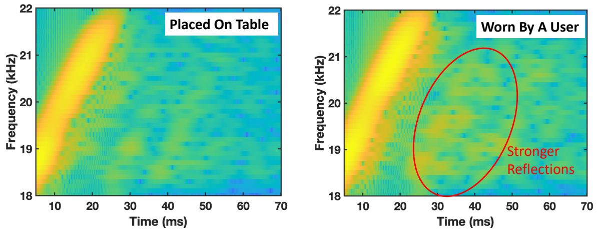
A. Head-reverberated Sound as Biometric

This work proposes to verify VR device users by capturing their head biometrics in the acoustic domain. We find that when users wear a VR headset, they tend to adjust the strap to wear it tightly and conveniently. The device and the head form a rigid body, whose deformation is relatively small, and the distance between any two points remains constant or nearly unchanged along time [16]. The VR device's speaker and microphone on this rigid body create an acoustic channel, and a sound traveling on this channel would be absorbed and reflected by it, as shown in Figure 1. Furthermore, the rigid body contains two chambers, the skull and the enclosed space between the face and the VR headset, which further capture the propagating signal beams, causing strong internal reflections and even amplifying the sound. Because each human head has a unique size, skull structure, mass, and facial pattern, the corresponding rigid body affects the sound differently before it reaches the microphone. We thus use the VR headset's speaker to emit acoustic signals and analyze its speaker-microphone channel responses to extract acoustically presented head biometrics.

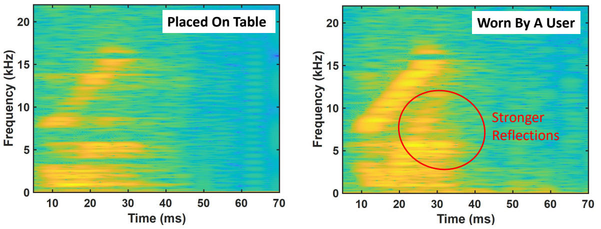
In particular, the relationship between the microphone sound $\hat{S}(f)$ and the original speaker signal $S(f)$ can be expressed in the frequency domain as

$$\hat{S}(f) = S(f)H_E(f) + N(f). \qquad (1)$$

$H_E(f)$ is the speaker-microphone channel response describing the echo effect, and $N(f)$ is the combination of ambient sounds. The channel response can be further divided into two



(a) Mobile phone VR headset.



(b) Standalone VR headset.

Fig. 2. Human head reverberating the sounds in the speaker-microphone channel of two types of VR devices.

components, the channel response related to the rigid body $H_{head}(f)$ and the environmental reflections $H_{env}(f)$. Thus the microphone sound can be further expressed as

$$\hat{S}(f) = S(f)\left[H_{head}(f) + H_{env}(f)\right] + N(f), \tag{2}$$

The principle of extracting head biometrics from acoustic signals is measuring $H_{head}(f)$ based on $\hat{S}(f)$ while reducing the impacts caused by $H_{env}(f)$ and $N(f)$. The intuition is that the environmental ultrasonic reflections suffer from higher attenuations due to traveling in the air, and the signals propagating on the rigid body could dominate the microphone data, though exposed to ambient noises [17].

To study the feasibility of using acoustic sensing to obtain users' head biometrics, we used two types of VR devices. Each device emitted a short 18-22kHz chirp signal when the device was placed on a table and worn by a user, respectively. The VR device's microphone recorded the echo sounds. Figure 2(a) compares the spectrograms obtained by the mobile phone VR headset (Samsung S8 phone) in the two scenarios. We find that when a user wears the VR headset, significant changes in the spectrogram are observed. In particular, stronger frequency components after the direct-path chirp signal (marked with a red circle) are observed when the user wears the device, indicating that more echo sounds return to the device's microphone. This is caused by the reflections of the head and the effects of the two formed chambers. Similar results can be observed on the standalone VR headset (Meta Quest) as shown in Figure 2(b). Moreover, we find that while the mobile phone VR keeps the original shape of the chirp signal in 18-22kHz, the standalone VR obtains the severely distorted signal non-linearly mapped within 0-16kHz, which is caused by its default echo cancellation mechanism and low sampling rate.

We further examine the head-reverberated sounds between two users using a mobile phone VR device, because its microphone keeps most frequency components of the original signal, making it easier to capture different physical impacts. Specifically, a stimulus sound consisting of five sinusoidal signals from 18kHz to 22kHz is used. Figure 3 shows the frequency responses of the sound, when it is reverberated by two users' heads. We observe that the received echo sounds present distinctive patterns for each user in the frequency domain. More specifically, the signal amplitude is amplified or suppressed differently at each frequency between the two users. The results confirm the potential of recognizing a user's head using acoustic signals and further motivate us to leverage the rich frequency components of a sound to describe head biometrics with fine granularity.

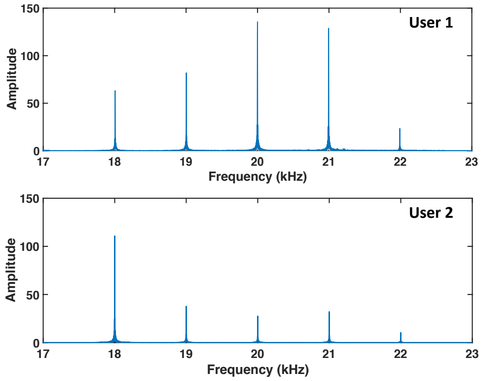


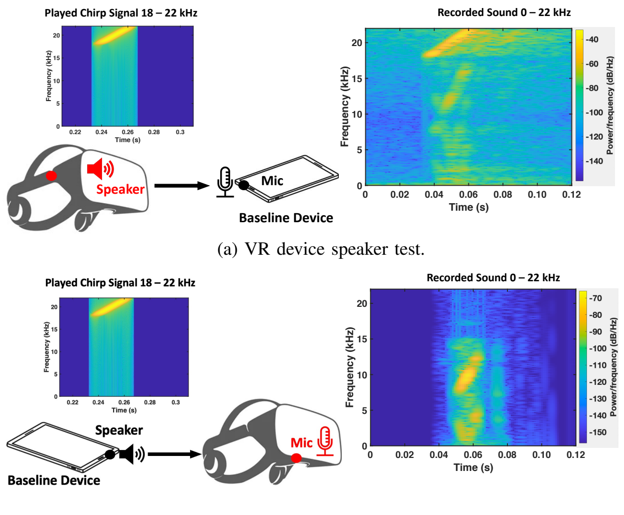
Fig. 3. Frequency responses of two VR users' head-reverberated sounds.

B. Challenges

As shown in Figure 2, the echo signal recorded by a standalone VR device is significantly different from the original signal, whose shape is not maintained or recognizable. Thus, it is hard to measure a target's physical impacts on the signal for sensing purposes. Additionally, the speakers and microphones of standalone VR devices are located either in the strap or under fabric cloth, whose influences on acoustic sensing are still unknown. To comprehensively understand the challenges of deploying active acoustic sensing on VR devices, we conduct experiments to test a Meta Quest and use a smartphone, Samsung Galaxy S8, as a baseline whose speaker-microphone audio feedback has been shown not to suffer from observable distortions.

VR Device Speaker Test. We use the Quest speaker to play an ultrasonic chirp signal (18-22kHz) and the phone mic to record the sound. As illustrated in Figure 4(a), the Quest speaker is able to play the ultrasonic sound with significant power maintained, though there are some frequency leakages below 18kHz. Such leakages cause the emitted ultrasonic sound slightly audible rather than truly inaudible. These sounds may be caused by the imperfect hardware and the vibration of the VR device frame and surface. We also observe a similar result on the Meta Quest 2 device, though with a different audio pattern. The speaker test indicates that the VR device speaker can support active acoustic sensing. It also motivates us to design the sensing signal in the ultrasonic band with a short duration, such as at the millisecond level, which further reduces the audibility of the signal to users.

VR Device Microphone Test. We use the phone speaker to play the same ultrasonic chirp (18-22kHz) and use the Quest's mic to record data, which is implemented with the APIs in



(a) VR device speaker test.

(b) VR device microphone test.

Fig. 4. Challenges of active acoustic sensing on standalone VR headset.

Unity [18]. Figure 4(b) shows the microphone test result, indicating **four critical challenges** of using the VR device for sensing: (1) The echo sound recorded by the Quest mic is severely distorted, and the sensing signal's original shape is unrecognizable, which makes it hard to analyze the impact of the sensing target. (2) The echo sound is also nonlinearly mapped into the low-frequency range ($\leq$ 16kHz), though the original sound is only at the inaudible frequency range. Although Unity supports over 48kHz microphone sampling rates, the Quest can only record audio with up to 32kHz sampling rate. The sounds above 16kHz would be nonlinearly projected to the lower frequency range with distortions. The data is also exposed to ambient audible noises. (3) Regarding the echo sound distortions, the default echo cancellation algorithm is the main cause, and there are also influences from the VR device's case and the speaker/mic locations. Some VR devices have such algorithms built in the hardware [14], [19], and thus, it requires rooting the VR device or modifying its hardware before applying the traditional active acoustic sensing methods. (4) Figure 4 also shows that the signal amplitudes of the Quest's mic data are significantly lower than that of smartphones, which adds more difficulties to using the signal for sensing. This is because such dedicated sensing signals are not recognized as meaningful media sounds or human voices and they are suppressed like noises [14]. As the noise suppression/removal algorithm is often integrated or concatenated with echo cancellation, we use echo cancellation to represent both in the rest of the paper.

Active Sensing Under Echo Cancellation. Active echo cancellation algorithms are widely used to remove the speaker-microphone-propagated sounds, such as least-squares FIR adaptive filters and frequency domain adaptive filters [20], [21]. These algorithms learn from short-term observations to estimate the speaker-microphone channel and predict a reference audio to deduct from the microphone data. Specifically, they need to consider multiple factors in generating the reference audio, including echo paths, circuit/signal propagation delays, and volume levels. The detailed designs of these algorithms on commodity VR devices are not released. But as shown in the above feasibility studies, these algorithms cannot completely remove the echo sounds due to the failure of perfectly generate the reference audio. We thus propose to recover useful sensing information from the surviving echo and attempt to counteract the effect of echo cancellation. By including the frequency response of echo cancellation $H_{EC}(f)$ and a noise suppression factor α to Equation 2, we model the active acoustic sensing under echo cancellation as

$$\hat{S}(f) = S(f)\left[H_{head}(f) + H_{env}(f) - H_{EC}(f)\right] + \alpha N(f). \quad (3)$$

Equation 3 presents the sensing model under echo cancellation by including its most significant component, the linear adaptive filter. The more precise model needs to consider the nonlinear mapping due to low sampling rates [22], device nonlinearities [23] and nonlinear filters [24].

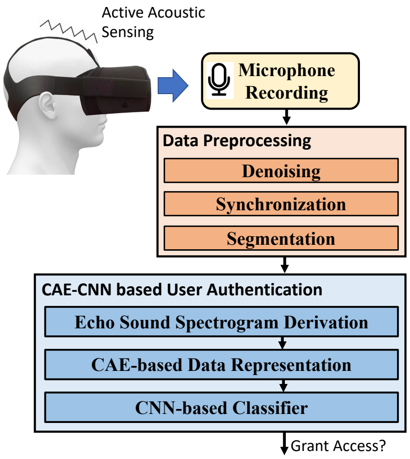


Fig. 5. The architecture of our system.

C. System Overview

The architecture of our system is shown in Figure 5. When the user puts on the VR device or is wearing it to log into an app or an online account, the system initiates an authentication session by playing a millisecond-level ultrasonic sound as the stimulus signal. The corresponding microphone data (i.e., echo sound) is obtained as the system input. We first perform *data preprocessing*, which consists of denoising, synchronization, and segmentation. Specifically, the system uses a bandpass filter to process the audio data and remove the acoustic noises outside the interested frequency band. Next, the system synchronizes the data based on the reference audio and locates the echo segment in the microphone data that corresponds to the stimulus signal and is expected to contain head-reverberated sounds.

The core of our system is a *CAE-CNN authentication algorithm*. The algorithm derives the spectrogram from the echo sound to describe the time-frequency characteristics of the head-reverberated signals. After that, the CAE model counteracts the echo cancellation effects and derives the stable head biometric encoding from the echo sound spectrogram, which further reduces the noise impacts, enables long-term use and adds difficulties to replay attacks. The reconstructed head biometric spectrogram is learned by a CNN to distinguish users. Based on the demanded authentication scenarios, the system can be deployed either locally at the VR headsets (e.g., for unlocking devices) or at the remote server (e.g., for logging into online accounts). For the remote server scenario, the system verifies the user against the claimed user identity, which is a binary classification. For the local authentication scenario, the system can identify more than one user to support multi-account features.

Multi-account Feature. Figure 6 illustrates the CAE-CNN algorithm for m-user account authentication, $m \geq 1$. Each row has a pair of CAE and CNN models bonded with one of m registered users (e.g., $user\ j$, $1 \leq j \leq m$), which stores the user's profile and estimates the probability of an authentication request to be from this user. When $user\ j$ enrolls in this system, a pair of CAE and CNN models are used to create two per-user profiles. Specifically, the $user\ j$'s sound spectrogram goes through the encoder of CAE to derive the head biometric encodings and create $user\ j$'s CAE profile. The head biometric encodings of $user\ j$ further go through the CAE decoder to reconstruct the head biometric spectrogram, which is fed into the following CNN model as $user\ j$'s training data. Moreover, a set of nonusers' head-reverberated sound audios are processed by both the encoder and decoder of $user\ j$'s CAE model, and the resulting spectrograms are further input into $user\ j$'s CNN model as the nonuser training data. The CNN model thus learns from the two classes of data to construct $user\ j$'s CNN profile. During the authentication phase, the spectrogram of a testing audio needs to be processed by each of the m users' CAE and CNN model pairs. Each CNN model outputs probabilities for two classes, the designated user and the nonuser, whose sum is 1. All user-class probabilities are further compared to find the maximum, which must be over 0.5 for an accept decision (i.e., one of the registered users).

D. Attack Models

The goal of an adversary is accessing the user's VR device to steal private information (e.g., browsing histories, preferences and sensitive App content) and perform unpermitted operations (e.g., deleting files, installing malware, making online payments and controlling the user's metaverse avatar). To achieve this goal, the adversary needs to spoof the user's identity to pass the VR device's authentication. We assume the adversary has gained physical access to the user's VR device and is familiar with our authentication system. Based on the specific professional knowledge and technical capabilities that an adversary could obtain, we consider the following attacks:

Zero-effort Attack. Rather than following the authentication procedure to present a biometric, an adversary may place the VR device on a table or a mount, attempting to break the authentication system with zero effort. It is worth noting that most VR headsets have the ability to detect the presence of the user's head based on a proximity sensor (usually placed on the top edge of the goggles). If not detecting the device worn by a user, the VR system goes to the sleep mode [25], [26]. Thus, to keep the VR system awake and complete authentication sessions, the adversary can simply attach a sticker to block the proximity sensor and fool the "liveness detection" mechanism.

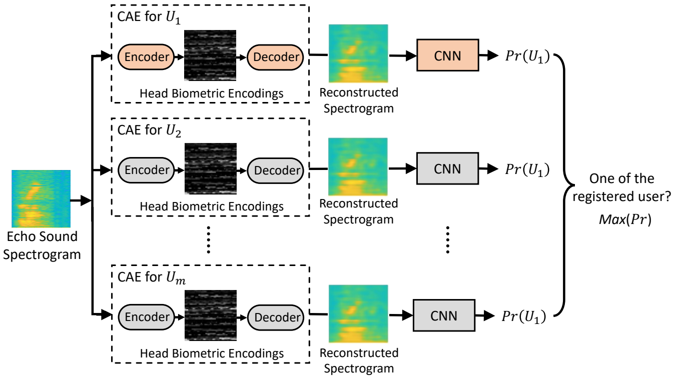


Fig. 6. The CAE-CNN authentication algorithm with m registered users.

Impersonation Attack. The attacker attempts to pass the authentication by wearing the VR headset in person, hoping similar head biometrics could be presented. The adversary may target victims with similar head sizes/mass and replicate the wearing position, strap height, and tightness. Moreover, as our system supports multi-user accounts, we divide impersonation attacks into two categories: 1) In *insider impersonation*, the attacker has been enrolled into the authentication system as one user but attempts to log into one other registered user's account; 2) In *outsider impersonation*, the attacker is not enrolled but attempt breaking into any one of the registered users' accounts.

3D Printed Head Attack. We consider a head biometric replay attacker, who could forge a physical head similar to the user based on 3D scanning & printing [27]. In particular, we use a commodity 3D scanner, Revopoint 3D Scanner-POP2 [28], to obtain the user's 3D head model and import it into a commodity 3D printer, Creality 3D Printer CR-10 V3 [29], to produce the fake head. Similar 3D printed heads have been used to break facial recognition systems [30]. We further add silicone [31], and the resulting head shows a similar head shape, size, weight, and face pattern.

Acoustic Replay Attack. Due to the open nature of acoustic channels, nearly all acoustic-based authentication systems are subject to replay threats. We consider two types of acoustic replay attacks based on how the adversary obtains the user's head-reverberated sounds: 1) The *side-channel eavesdropping replay* attacker places a hidden microphone in the user's proximity to record the user's authentication. The audio is later amplified and replayed by an external speaker to the VR authentication system. 2) The *leaked biometric replay* attacker is assumed to have obtained the audio files exactly the same as that used in the user's profile training. This replay audio is not impacted by additional noises (e.g., incurred during side-channel eavesdropping) and is expected to reflect the maximal replay attack performance.

Denial-of-Service Attack. The attacker who aims to disable or cause errors in the authentication process could use an

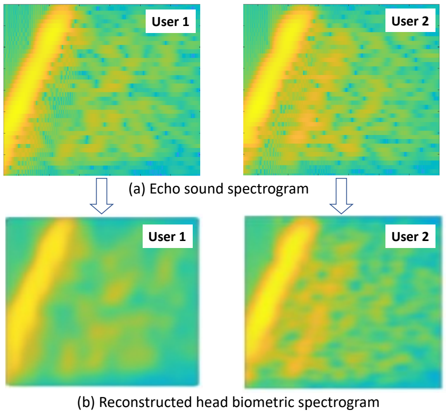


Fig. 7. Illustration of head biometric spectrograms of two users on mobile phone VR headset (before and after CAE reconstruction).

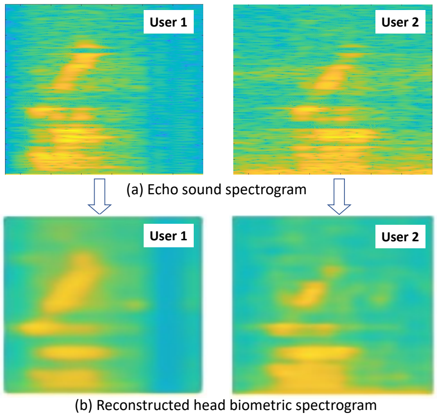


Fig. 8. Illustration of head biometric spectrograms of two users on standalone VR headset (before and after CAE reconstruction).

external speaker to play dedicated ultrasonic sounds near the target user. The ultrasonic attacking sounds may overwhelm the authentication signal to keep the user rejected without arousing attention.

III. Approach Design

A. Sensing Signal Design

We design an ultrasonic pulse signal to sense the VR user's head. In particular, the pulse signal is designed to sweep from 18kHz to 22kHz. This frequency band is barely audible to human ears and does not overlap with regular ambient noises, such as air conditioning, human voices, and media sounds. Moreover, the sweeping frequency provides frequency diversity to capture more aspects of the user's head biometric than a single frequency. The pulse signal lasts for a short period (i.e., 25ms), and we capture both the direct-path sound and the multi-path echoes that arrive later to analyze how the original speaker signal is absorbed and reflected by the user's head before returning to the device's microphone. We also apply a Hamming window to smooth the signal and reduce the spectral leakages and hardware noises, which are caused by sharp frequency jumps.

B. Data Pre-processing

Before analyzing the microphone data, we first calibrate the audio with denoising, synchronization and segmentation. Specifically, we apply a bandpass filter to remove the noise outside the interested frequency range. The passband is set as 18-22kHz for mobile phone VR, and the air conditioning noise, human voices and other environmental sounds can be removed [32], [33]. For standalone VR, because its mic can only record sounds up to 16kHz, we set the passband to be 1-16kHz and need to further use our CAE model to denoise and recover useful sensing information. It is worth noting that standalone VR devices' echo cancellation is often integrated with noise suppression. Thus, the above noises are suppressed together with the sensing signal.

Next, synchronization is performed to locate the pulse signal in the microphone data. We iteratively shift the microphone data $\hat{s}$ and compute its cross-correlation with a reference signal s. The shift length leading to the maximum cross-correlation coefficient indicates the time delay between the two signals as expressed by

$$delay = \underset{m}{argmax} \sum_{n=0}^{N-m-1} \hat{s}(n+m)s(n), \tag{4}$$

where m is the number of shifted samples. After subtracting this delay, we find the start of the pulse within the microphone data and obtain a 75ms audio segment containing both the pulse period (25ms) and a silent period. The purpose is to capture the user's head biometric features from both the direct-path signal and the reflected/refracted signals that arrive later. We find that received sounds attenuate over 20dB after 75ms, which shows low signal power and are discarded. Additionally, we normalize the amplitude of the audio segment to be within the range $[-1, 1]$.

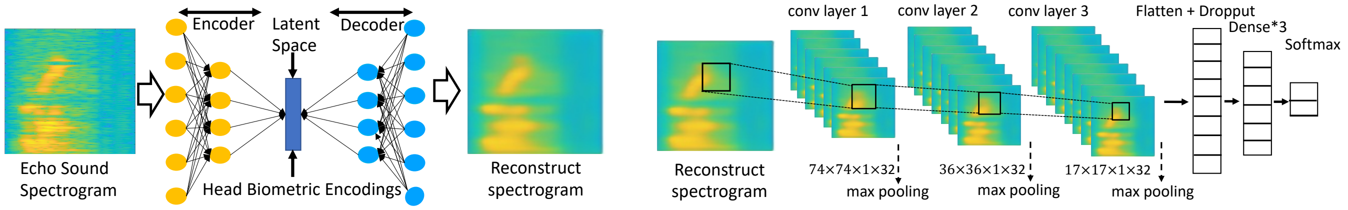
C. Echo Sound Spectrogram Derivation

We derive the echo sound spectrogram using the mic data, which is a time-frequency image to describe how each spectral point of the original signal is interfered with by the user's head along time. The echo sound spectrogram $s(\tau)$ is calculated based on a window function $w(\tau)$ with length T. Specifically, each pixel at the spectrogram position (t, f) is computed by Equation 5 and 6, where t and f are the time and the frequency index. The derived echo sound spectrogram is fed into our CAE-CNN authentication algorithm to extract head biometric information and perform biometric-based authentication.

$$STDTFT(t,f) = \sum_{\tau=t}^{t+T-1} s(\tau)w(\tau-t)e^{-j2\pi f\tau} \tag{5}$$

$$spectrogram(t,f) = |STDTFT(t,f)^2| \tag{6}$$

Figure 7 (a) and Figure 8 (a) show the derived echo sound spectrograms for two users using the mobile phone VR and the standalone VR, respectively. We observe that the same sensing signal results in slightly different echo spectrograms for the two users on both VR devices. Moreover, not only the direct-path signal but also the reflected signals present individually unique patterns in the time-frequency domain.



(a) CAE model for per-user head biometric encoding.

(b) CNN model for head biometric recognition.

Fig. 9. Detailed design of our CAE-CNN algorithm.

D. *CAE-CNN Authentication Algorithm*

We design a CAE-CNN algorithm to extract the user's head biometric encoding and perform user authentication. In particular, a pair of CAE model and CNN model is created for each registered user as introduced in Section II-C. The spectrogram derived from the echo sound is input into each user's consecutive CAE and CNN models. The probability of a head biometric belonging to each user is output and compared for authentication decision. We next introduce the detailed designs of our CAE and CNN models.

1) CAE-based Biometric Representation: We design a CAE model, which derives stable head biometric encodings under noises and counteracts the built-in echo cancellation to recover biometric spectrograms. The per-user CAE model is shown in Figure 9(a). It consists of three main parts, an encoder, the latent space, and a decoder. The input to the CAE model is a $64 \times 64 \times 3$ echo spectrogram image, which covers the frequency range 18-22kHz for mobile phone VR and 1-16kHz for standalone VR. The output of the CAE model has the same dimensions as the input. The encoding and decoding process filters out the noises, environmental reflections, and the echo cancellation's effect (as shown in Equation 3) and leaves only the head biometric information in the reconstructed spectrogram. Mean Squared Error (MSE) is used as the loss function to update the network weights during the training phase. In the CAE model, we choose the dimension of the latent space to be 2048.

Encoder. As the first major part of the CAE model, the encoder derives head biometric encodings from echo sound spectrograms and consists of four convolutional layers and three max pooling layers. Its detailed structure is presented in Table I. The encoder starts with a convolutional layer with 3×3 kernels of size $64\times64\times3$ to learn the large-scale features, followed by three convolutional layers with 3×3 kernels of size $32 \times 32 \times 48$, 3×3 kernels of size $16 \times 16 \times 192$ and 3×3 kernels of size $8 \times 8 \times 32$ to learn small-scale features. Additionally, a stride of 1 is applied, and ReLU is used as the activation function for all convolutional layers.

Decoder. The reconstructed spectrogram is decoded from the latent feature space (i.e., head biometric encodings) learned by the encoder. Its detailed structure is shown in Table II. It has five convolutional layers with 1×1 kernels of size $8 \times 8 \times 32$, 3×3 kernels of size $16 \times 16 \times 192$, 3×3 kernels of size $32 \times 32 \times 96$, 3×3 kernels of size $64 \times 64 \times 48$, and 3×3 kernels of size $64\times64\times3$, respectively. A stride of 1 is applied. ReLU and Sigmoid are used as the activation functions for the first four and the last convolutional layers, respectively. Three upsampling layers are used in the decoder with a stride of 2 for upsampling. The last convolutional layer is used to force the output from the previous layer to be interpreted as pixel intensity of an RGB image with dimension $64 \times 64 \times 3$. It is then fed into the following CNN model for user authentication.

CAE Profile. Each user has their own CAE model. During the registration phase, the user's training data are collected and input into the CAE to learn head biometric encodings. It is important to note that only the CAE encoder and the user's data are needed to create the user's CAE profile. There is no need for a nonuser dataset at this stage. We then use the user's CAE profile to reconstruct spectrograms during the authentication phase. Figure 7(b) and Figure 8(b) show the reconstructed spectrograms of two users with two types of VR devices. For both devices, we observe that the differences in the reconstructed spectrograms between the two users are "amplified". The reason is that the head biometric information is emphasized while the unrelated signal components, including the noises, environmental reflections, and the influence of echo cancellation, are removed. More importantly, though the mic data of standalone VR heavily suffers from ambient noises and echo cancellation, our CAE model is still able to extract stable biometric encodings from severely distorted sensing sounds.

TABLE I
THE STRUCTURE OF CAE MODEL ENCODER.

Layer	Output Shape	Param #
Input: Echo sound spectrogram	(64, 64, 3)	0
Conv2D + RecLineU	(64, 64, 48)	1344
Max Pooling 2D	(32, 32, 48)	0
Conv2D + RecLineU	(32, 32, 96)	41568
Max Pooling 2D	(16, 16, 96)	0
Conv2D + RecLineU	(16, 16, 192)	166080
Max Pooling 2D	(8, 8, 192)	0
Conv2D + RecLineU	(8, 8, 32)	6176

TABLE II
THE STRUCTURE OF CAE MODEL DECODER.

Layer	Output Shape	Param #
Input: Head biometric encodings	(8, 8, 32)	0
Conv2D + RecLineU	(8, 8, 192)	6336
Up Sampling 2D	(16, 16, 192)	0
Conv2D + RecLineU	(16, 16, 192)	331968
Up Sampling 2D	(32, 32, 192)	0
Conv2D + RecLineU	(32, 32, 96)	165984
Up Sampling 2D	(64, 64, 96)	0
Conv2D + RecLineU	(64, 64, 48)	41520
Conv2D + Sigmoid	(64, 64, 3)	1299

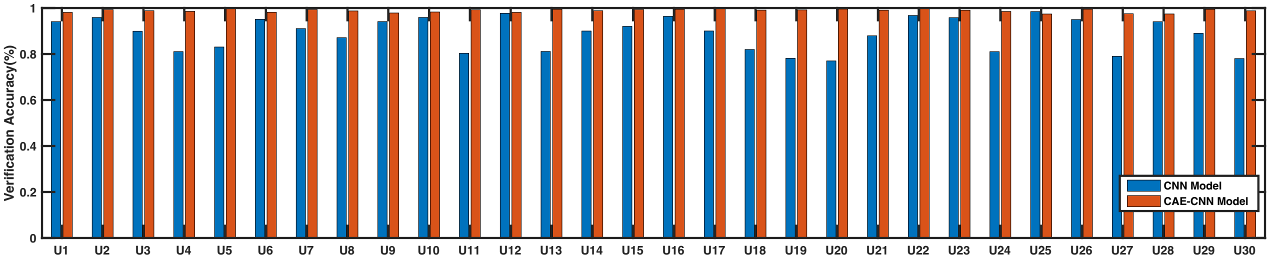


Fig. 10. User verification accuracy and the CNN vs. CAE-CNN comparison (standalone VR).

2) CNN-based User Identification: We develop a CNN model with three convolution layers and one fully connected layer to analyze reconstructed spectrograms and learn the head biometric features to differentiate users. The per-user CNN model outputs binary classification results, the probabilities for the *user* and *nonuser* class. The output dimensions in each layer are tuned to balance processing time and accuracy, which is calculated as

$$dimensions = (\tfrac{m-k+2d}{l} + 1) \times (\tfrac{m-k+2d}{l} + 1) \times t \qquad (7)$$

where m, k, l, d and t are the input image size, kernel size, step length, the number of padding, and number of filters.

The detailed structure of our CNN model is shown in Figure 9(b). In the first layer, the dimensions of the normalized input image are set as 150×150. After the input layer, there is a convolutional layer followed by a max pooling layer, where the convolutional kernel size is 3×3 and the pooling kernel size is 2×2. The step length is set as 1. The number of padding applied is set as 0, and the number of filters is 32. After the first convolution operation, the dimensions are calculated as $148 \times 148 \times 32$ by the above equation. Since the kernel size of the pooling layer is 2×2, the dimensions after the first pooling operation are $74 \times 74 \times 32$. We keep the same configuration for the rest of the convolution and pooling layers. At the end of the model, we utilize the softmax function to normalize the network output and obtain a probability for each class as the decision confidence or CNN score. We then utilize Adam as the optimizer leveraging the power of adaptive learning rates to find individual learning rates for each parameter. We use sparse categorical cross-entropy as the model's loss function since we expect class labels to be provided as integers instead of one-hot encoding.

CNN Profile. The training requests two classes of data labeled *user* and *nonuser*. Both classes' spectrograms first go through the user's CAE model to be encoded and decoded based on the user's CAE profile. The reconstructed spectrograms are then fed into the user's CNN model for training, which learns from two classes of data to create the user's CNN profile. When testing, if there is only one registered user, a high CNN score for the *user* class leads to a granted access permission. When a group of users shares one VR device, the testing data needs to go through all pairs of CAE and CNN models for multi-account authentication. The maximum CNN score is searched among all *user* classes. If it is greater than 0.5, the access permission of the corresponding user's account is granted.

IV. Performance Evaluation

A. Experimental Setup

Platforms. We evaluate our system with two types of VR headset devices, a standalone VR (i.e., Meta Quest) and a mobile phone VR (i.e., DESTEK V3 VR with a smartphone - Samsung Galaxy S8). The two devices also represent the VR devices, which have or have no default echo cancellation mechanism. We developed two experimental platforms for the study. Specifically, we developed an Android App for the mobile phone VR headset and installed it on the smartphone running Android 9.0, which emits an ultrasonic pulse signal and records the stereo sounds using the phone's two microphones simultaneously. For the standalone VR headset, we developed a VR App based on Unity 2019.4.4f1. We utilize its *AudioSource* package to let the standalone VR headset play the ultrasonic pulse signals via its two built-in speakers (embedded on the two side straps). It has only one built-in microphone to record the sound at the same time. The collected audios are processed offline.

Data Collection. We recruited 30 participants (7 females and 23 males) with ages $20 \sim 35$, heights $5'2'' \sim 6'4''$, weights $103 \sim 216 lbs$, fat ratios $15 \sim 28\%$, and hair lengths $0.6 \sim 22$ inches to conduct experiments with both VR devices. The IRB approval had been obtained. Before data collection, participants were given time to use the VR headset and adjust the headset straps to their convenient tightness. During experiments, each participant was asked to put on and take off the VR headset 20 times, and 10 chirp sounds were collected each time when the user was wearing the device. Thus, the slight change of the rigid body incurred by practical VR device uses and behavioral inconsistency is included in the collected data. Each participant also had the freedom to sit and stand during the experiment. Moreover, we conducted a long-term study with 11 participants and collected multi-session/day data over 15 months, and the impacts of weight changes and behavioral inconsistency over this long time and the cross-day variations (e.g., the ambient noises, clothes, and hair styles/lengths) are considered. For single session/day evaluation, we use 60% of the data for training and the rest for testing. For multi-session/day evaluation, we use the first day's 60% data for training and all other days' data for testing. Regarding the nonuser dataset, for every participant's CNN model, the other 29 participants are included to train the *nonuser* class.

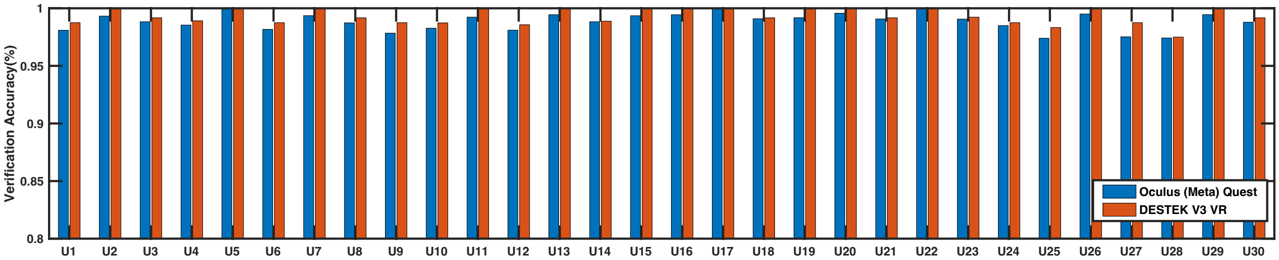


Fig. 11. Device comparison in user identification (Mobile phone VR headset vs Standalone VR headset).

B. User Authentication Performance

1) Single-user Verification: We first evaluate the authentication performance of our system in the single-user scenario. Figure 10 shows the user verification accuracy for each of the 30 participants using the standalone VR. We observe that our system based on CAE-CNN achieves high accuracy for all the participants. In particular, the system verifies the user with 98.87% accuracy on average, and nearly half of the participants achieve above 99% accuracy. Additionally, Appendix Table VI and Table VII show that our CAE-CNN algorithm achieves 98.90% True Positive Rate (TPR) and 98.82% True Negative Rate (TNR). The results confirm the high verification performance of our system.

CAE-CNN vs. CNN. To examine the security gain brought by the CAE model, we compare the verification performance of 30 participants using CAE-CNN and CNN, respectively. Figure 10 and Appendix Table VI and Table VII show that when using CNN alone, our system only achieves 88.90% accuracy, 89.00% TPR and 88.86% TNR. Adding the CAE model for each user significantly improves the verification performance by 11.2%. The reason is that encoding head biometrics helps remove noise, making the individually unique head biometrics become identifiable and more resistant to behavioral inconsistencies and environmental noise.

Device Comparison. We compare the performances of two VR devices, a standalone VR and a mobile phone VR. They represent two categories of popular VR models and record two types of microphone data, with and without default echo cancellation. Figure 11 shows the user verification accuracy achieved by the two devices. We find that our system performs well for both devices. In particular, our system achieves 98.87% accuracy on average for the standalone VR, and the average accuracy is 99.33% for the mobile phone VR. Moreover, for each of the 30 participants, the mobile phone VR device achieves higher accuracy. One reason is that it does not have the default echo cancellation mechanism and thus maintains good shapes of the sensing signal for analysis. The device model differences, including the speaker/microphone locations and the device surface materials (plastic or fabric cloth), are also reasons to cause the performance differences. The results demonstrate that our system is able to scale among different VR device models to provide reliable authentication services.

2) Multi-user Account Verification: We next evaluate the authentication system when the VR headset is shared among a group of m users (e.g., family members, lab mates, classmates). If a nonuser is recognized as any one of the registered users, or a user is identified as another registered user, the authentication fails. Figure 12 shows the multi-user account authentication results of the standalone VR headset when m changes from 1 to 5, covering most typical family sizes. We find that our system achieves a high user verification performance for all multi-user scenarios, though their average performance is slightly lower than the single-user scenario. In particular, when two users are registered, the system achieves a 98.92% TPR to identify the user and a 98.79% TNR to reject nonusers or misclassify the users. The performance drops slightly when more users are registered. When there are three registered users, the system achieves 98.88% TPR and 98.65% TNR, and when there are five registered users, the TPR and TNR are 97.56% and 97.50%. The results indicate the capability of our system to provide multi-user account verification services on VR devices.

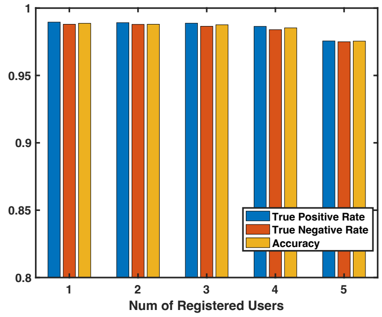


Fig. 12. Multi-user account verification performance (standalone VR).

Multiple CNNs vs. Single CNN. We compare the multi-user verification performance of using single and multiple per-user CNN models. Table III presents the performance of five registered user scenario. The multi-CNN model outperforms the single CNN for multi-user account verification. Specifically, using five CNNs improves the verification accuracy, TPR and TNR by 7.45%, 6.61%, and 8.6%, respectively. The reason is that we assign a CNN model to each user and leverage its multi-class classification capability to tolerate behavioral

TABLE III
MULTI-USER VERIFICATION (MULTI-CNNS VS. 1 CNN).

	Accuracy	TPR	TNR
One CNN	90.73%	91.95%	89.51%
Multiple CNNs	98.18%	98.56%	98.11%

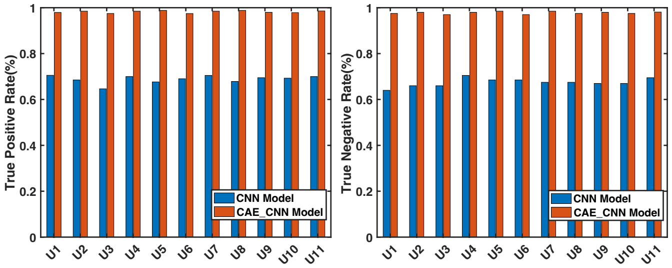
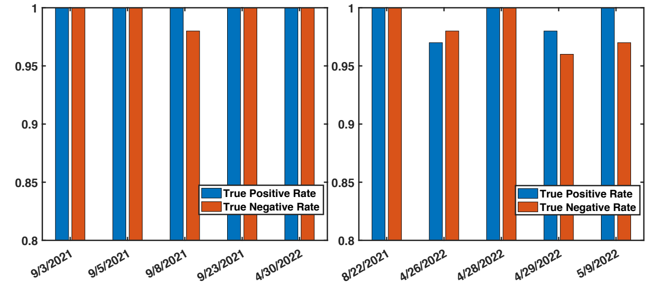


Fig. 13. Long-term evaluation with 11 participants and the CNN vs. CAE-CNN comparison (8/22/2021∼11/18/2022).



(a) Mobile phone VR headset

(b) Standalone VR headset

Fig. 14. Long-term tracking of two users with two VR devices.

inconsistency and noises.

C. Long-term Study

1) Eleven Users Over 15 Months: We conduct a long-term two-session study to examine the performance of our system. In 15 months, the participants may have varying hair lengths/styles, body weights, clothes, and the environmental factors, such as furniture and ambient noises, are not the same, reflecting a practical VR device-using scenario. Specifically, eleven participants' data on 8/22/2021 is used for training (60% data), and their data collected between 11/15/2022 and 11/18/2022 is used for testing. Meta Quest is used. Figure 13 shows that our system using the CAE-CNN model achieves high verification performance for all 11 participants, with 98.22% TPR and 97.78% TNR on average. The results indicate that our system is robust to diverse practical variations, and the acoustically represented head biometric is stable despite the normal human weight and clothes changes in the long run. In comparison, using the CNN model alone achieves 68.87% TPR and 67.59% TNR on average. The addition of the CAE model brings 42.6% performance improvement. The results confirm the ability of our CAE-CNN model to derive stable head biometrics and reduce the impacts of irrelevant factors.

2) Multi-day Tracking & Two VR Devices: To better understand the environmental impacts and the long-term changes of the user's head biometric, we track two participants' authentication performance on two VR devices over 8 months. We use 60% of the first-day data for training, the rest of the first-day and all other days' data for testing. Figure 14(a) shows the user verification accuracy achieved by the mobile phone VR on five different days, which are all very high. The minimum performance is achieved on the day 9/8/2021, which has a 98% TNR. Figure 14(b) presents the verification performance achieved by the standalone VR. The device achieves high accuracy for all five days, though its long-term performance is slightly lower than that of the mobile phone VR. In particular, the standalone VR device achieves 97% TPR and 98% TNR on 4/16/2022 and 98% TPR and 96% TNR on 4/29/2022. The slight performance fluctuations across different days are owing to the variations in the participant's status and the environmental noises. No significant performance degradation is observed over time. The results confirm our CAE-CNN model's effectiveness in providing daily VR authentication services.

D. Zero-effort and Impersonation Attack

Zero-effort Attack. We now evaluate our system against zero-effort attackers who attempt to pass authentication by presenting no head biometrics. We simulate this attack by placing the standalone VR headset on a table and blocking the device's proximity sensor with a small tape to pretend that a user wears the device. Then the attacker uses the handheld controller to start the authentication session and attack each participant. Table V presents the success rate of the zero-effort attack, which is 0%. The result reflects the effect of our system to reject authentication requests when the VR device is not worn by a human head.

Impersonation Attack. A robust and secure user authentication system needs to successfully detect both *insider* and *outsider* impersonation attackers. We simulate the two types of impersonation attacks to evaluate the security of our system. Specifically, each participant was respectively selected as the target user and two assumed attackers attempted to imitate the target user's way of wearing the VR headset. The attackers are beside each target user during data collection to learn and later repeat the user's wearing behaviors without changing the device's strap length. We consider two scenarios when the attackers are within or outside of the registered user group. Table V presents the success rates of the two impersonation attacks on standalone VR device, which is 0.66% for the *insider* impersonation and 0.82% for the *outsider* impersonation. The results show that our system effectively prevents in-person biometric imitations and confirms the robustness of acoustic head biometrics because an adversary is hard to imitate the skull structure, head mass and face patterns in person.

E. 3D Printed & Fake Head Attack

Though not being able to change head biometrics in person, an adversary may choose a silicone fake head with a similar size and shape to the target user's head to attack. A more advanced adversary could exploit the latest 3D scanning and printing technology to replicate a head with exactly the same head size/shape/weight and face patterns. We select one participant as the target user with the most similar head shape and size to our purchased silicone head. We further produce a 3D printed head. The attacks are illustrated in Figure 15. For each authentication session, the adversary puts the standalone VR headset on one fake head and uses the handheld controller to start the authentication. The attack performances are presented

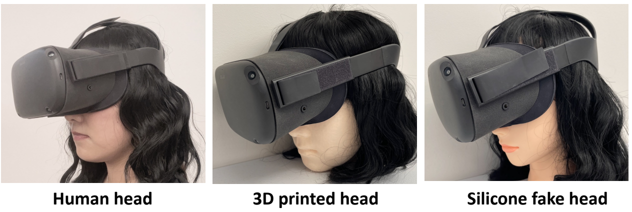


Fig. 15. Physically reproducing head biometrics with two fake heads.

in Table V. The silicone fake head and the 3D printed head achieve 0% and 0.50% success rate, respectively. We further test the 3D printed head with the mobile phone VR, which achieves a 1.03% success rate as shown in Table IV. The results indicate that only copying the head shape/size/weight and face patterns is still hard to break our system. The skull structure and the spatial distribution of head mass are more important biometric characteristics, which are hard to reproduce in practice.

F. Acoustic Replay Attack

We evaluate our system with two types of replay attacks. To generate the attacking sounds with sufficient signal power, we use an audio amplifier, Douk Audio Mini, with a loudspeaker, LU43PB 3-Way High-Performance Speaker, to replay the attacking audio. Based on the potential of an attacker to obtain the user's authentication audio, we consider the side-channel eavesdropping and the leaked biometric scenarios. The latter directly replays the authentication audio recorded by the VR device. For the mobile phone VR device, replaying leaked biometrics is expected to achieve maximal performance. For the standalone VR device, the side-channel eavesdropped sound is still barely inaudible, but the leaked biometric audio is audible, as shown in Figure 16.

Side-channel Eavesdropping Replay. We use a smartphone (i.e., Samsung Galaxy S8) to record during the target user's authentication process and use the above amplifier-speaker setup to replay the eavesdropped audio to the VR headset. Table V presents the replay success rate with the standalone VR, which is 0.5%. The result shows that our system performs well in preventing eavesdrop-based replays.

Leaked Biometric Replay. We simulate the attack by directly playing the authentication audio recorded by the user's VR device. Table IV and Table V present the success rates of this attack on the mobile VR and the standalone VR, which are 1.05% and 0.76%, respectively. The results show the high replay resistance of our system on different VR devices. Moreover, we notice that while mobile phone VR performs better than standalone VR in regular authentication scenarios, its performance under replay attacks is lower. The lack of echo cancellation majorly causes such differences.

TABLE IV
PERFORMANCE UNDER ATTACK (MOBILE PHONE VR).

Attack Scenarios	Attack Success Rate
3D Printed Head Attack	1.03%
Replay Attack-Leaked Biometric	1.05%

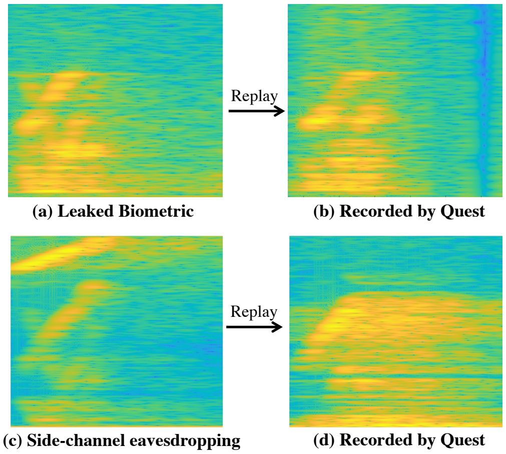


Fig. 16. Two types of acoustic replay attacks on standalone VR device.

CAE and Echo Cancellation. To understand the effect of the CAE model and the default echo cancellation in defending against replay attacks, we conduct a comparison study using different combinations of CNN, CAE and echo cancellation. Figure 17(a) presents the attack success rates of the leaked biometric replay. Specifically, when only using CNN, the replay attack achieves a 23.42% success rate, which is a serious security issue. But using CNN with echo cancellation, the attack success rate is reduced to 5.35%. CAE has a higher capability than echo cancellation to prevent replay sounds, and using CAE with CNN reduces the attack to a 1.05% success rate. Using CNN, CAE and echo cancellation altogether reduces the attack success rate to 0.76%. Better replay resistance performance against side-channel eavesdropping is presented in Figure 17(b). The results confirm the high replay resistance capability of our system, which involves both CAE and echo cancellation.

G. Noise Impact & Denial-of-Service Attack

When using a VR headset, the user tends to find a relatively quiet place with few people around. We thus evaluate our system under different ambient sounds with four-decibel levels: a typical room environment (30dB), music played in the next room (40dB), human conversation (50dB), and air conditioner noise (60dB). Two participants are involved in the ambient noise study. The standalone VR device is tested, as mobile devices have been tested by many prior works [13], [34], [35]. As illustrated in Figure 18, the system achieves the best performance under 30dB noise with 99.36% accuracy.

TABLE V
PERFORMANCE UNDER ATTACK (STANDALONE VR).

Attack Scenarios	Attack Success Rate
Zero-effort Attack	0%
Impersonation Attack *(insider)*	0.66%
Impersonation Attack *(outsider)*	0.82%
Silicone Fake Head Attack	0%
3D Printed Head Attack	0.50%
Replay Attack - Side-channel	0.50%
Replay Attack - Leaked Biometric	0.76%

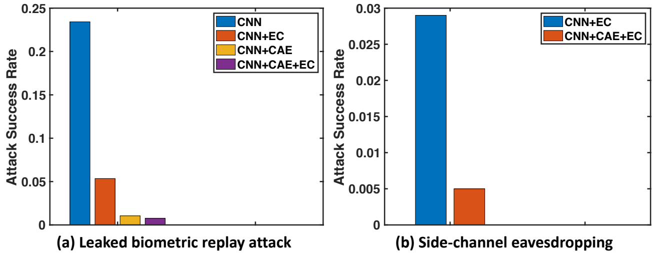


Fig. 17. Effects of CAE and echo cancellation to prevent replay sounds.

The accuracy performance is slightly degraded to 98.83% at 40dB and 97.52% at 50dB, which are still high. At 60dB, the user verification accuracy drops to 92.31%, which is still acceptable. The results show that our system can work well for most regular indoor scenarios.

We finally examine the Denial-of-Service attack, which generates ultrasonic sounds on purpose to block the authentication sounds without causing notice. The 17-22kHz white noises are used to generate ultrasonic interference. We find that the typical indoor ultrasound pressure level is usually lower than 10dB and thus choose five ultrasonic sound pressure levels from 10dB to 70dB. The performance is shown in Figure 19. We observe that our system achieves high accuracy with up to 50dB ultrasonic interference, though the performance slightly degrades when the ultrasound increases from 10dB to 50dB. In particular, our system achieves 95.55% TPR and 99.76% TNR under 10dB ultrasonic noises. The performance degrades to 92.82% TPR and 99.52% TNR when the ultrasonic noise increases to 20dB. We also find that the TNR is always higher than TPR at each ultrasound level, indicating that the system tries to reject all suspicious users in a noisy environment. When the ultrasound is at 50dB, our system achieves 92.24% TPR and 98.32% TNR. We thus choose 50dB to be the threshold for environment checking. If the ambient ultrasound is higher than it, the user has to use the traditional passwords.

V. Discussion & Future Work

Feature Importance. To investigate the importance of different head characteristics, we partition 30 participants into subsets according to head sizes and weights, and let assumed attackers only impersonate the subsets that show similar head sizes and weights. The attack success rates are between 0.63∼0.68%, comparable to the 0.66% reported in Section IV-D. Moreover, the 3D printed head's result in Section IV-E shows that it is still hard to break our system even presenting a similar head shape/size/weight and face pattern. The above studies indicate that the skull structure and the spatial distribution of mass within the head are more important and stable characteristics, which are also harder to reproduce in practice. Furthermore, the long-term study in Figure 14 shows that the normal changes of the user's weight, body-fat ratio, clothes and hair length do not obviously impact authentication performance. We further partition the long-term results regarding each user and find that their performances slightly drop on the same days, which indicates that environmental factors (e.g., ambient/device noises) have a greater impact than normal biometric feature changes.

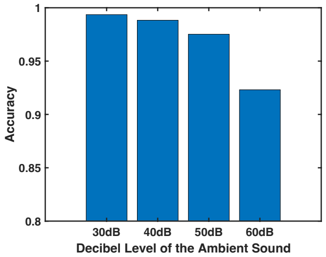


Fig. 18. Impact of ambient noise.

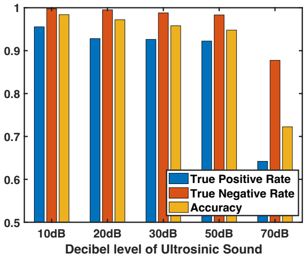


Fig. 19. DoS attack performance.

Practical Deployment. In our experiments, each participant puts on/off the VR headset 20 times (e.g., 3 seconds per time). For practical use, we need to know the impact of user enrollment time on verification performance. In particular, when the 30 participants wear the standalone VR headset 4, 6, 8, 10 and 12 times, our system achieves 86.37%, 90.25%, 92.44%, 96.84%, 98.87% verification accuracy on average, respectively. Thus, to achieve over 90% verification accuracy, a user needs to put on the device 6 times for enrollment. Furthermore, we conducted a user study to find that 2/30 participants noticed the sensing signal on a mobile phone VR device and 8/30 participants noticed the sound of standalone VR due to the frequency leakage. All participants feel the sound is acceptable, because the sound has a low volume and is at a millisecond level.

Future Work. We consider the following tasks for future work: (1) As we involve a nonuser data set for constructing the user's CNN profile, it is important to select a number of nonusers covering a wide range of races, ages, weights, and body fat ratios. (2) This work mainly studies Meta Quest and Quest 2, because of their low prices and high shipments. We will continue to evaluate other standalone VR devices, such as HTC VIVE Pro and HP Reverb. The potential of applying our system on AR devices will also be studied. Because we have solved the challenge of acoustic sensing under echo cancellation and noise suppression, it is expected that it will not be hard to deploy our system on these devices. (3) Due to hardware limitations, the authentication signal of our system is still slightly audible on commodity VR devices. Though all participants in our study think this sound is not disturbing, we will further reduce the audibility of the sensing signal by reducing its duration and trying other signal patterns. (4) Further studies are needed to reduce the user's training efforts, including simulating speaker-microphone channels to augment training data and using transfer learning to address the scenario when the user changes VR headsets. (5) The sensing signal pattern may be improved to address diverse ambient noises and escape echo cancellation. (6) The more advanced 3D scanning & printing technologies and filler materials will be used to study physical head forgery attacks. (7) we will explore other learning algorithms (e.g., FaceNet [36], [37]), which may replace the role of multiple CNNs in our system and balance complexity and security.

VI. Related Work

While not having a touch screen, current VR devices ask users to enter on a virtual floating keyboard with handheld controllers, VR pointers and hand gestures. During this process, the user's vision is confined in the virtual world, but the actions in the physical world could be observed by surrounding people or cameras, which are subject to authentication secret leakage [38]. To address such security issues, active works are on extracting behavioral biometrics from the user's motions to improve VR authentication security. For example, head movements are demonstrated to be identifiable when users listen to music beats [39] or perform a required task in the virtual space [5], [7], [40], [41], such as moving the VR pointer to follow a ball or walking. Such head behavioral biometrics are captured by the head-worn devices' inertial sensors. Wang *et al.* further develop a VR authentication technique that allows users to unlock their profiles with simple nodding actions, and the biometric features related to neck length and head radius are captured [8]. Furthermore, the VR headset can be used together with handheld controllers for enhanced VR authentication, which captures not only the head movements but also the hand motions, body motions and even the eye gaze [6], [42]. Additionally, Mathis *et al.* ask the user to enter a number password on a 3D Rubik's cube in the virtual space, where both the password and the handheld controller motion patterns are verified. However, these methods are all based on the slow motion-level human-device interaction, whose performance is limited by behavioral inconsistency and the low fidelity sensor data.

This work proposes to simplify the VR device authentication procedure using active acoustic sensing. There have been many studies exploring active acoustic sensing on mobile devices. For example, active acoustic sensing can be used to recognize finger/hand gestures performed on the mobile device or in the air [43], [44]. ForcePhone [45] emits repetitive chirp sounds and analyzes the structure-borne sounds to sense the finger force applied on the smartphone screen. VSkin [46] transmits modulated inaudible sounds to capture finger gestures on the back surface of the device. RobuCIR [47] develops a contact-free gesture recognition system for mobile devices based on active acoustic sensing. Furthermore, active acoustic sensing enables mobile devices to provide health monitoring. Nandakumar *et al.* [48] transmit 18-20kHz sound waves using the smartphone's speaker and capture the reflected-back signals to measure the chest/abdomen movements for apnea detection. Qian *et al.* [49] further generate an acoustic cardiogram using the acoustic reflections to monitor the user's heartbeats. Additionally, active acoustic sensing has been applied for indoor localization using either audible [50], [51] or inaudible sounds [52], [53].

Active acoustic sensing has also been widely adopted on mobile devices for low-effort user authentication. For example, EchoFace emits acoustic signals to detect the uneven stereo structure of the user's face to prevent 2D replay faces [9], and EchoPrint leverages the unique echoes bouncing off the user's facial contour for authentication [10]. EarEcho [54] leverages the audio played by the earpiece speaker to sense the user's ear canal, where acoustic features are extracted from the transfer function between the recorded echo and the played audio for user authentication. When a user holds a mobile device, acoustic signals are also used to recognize the user's hand for authentication [13], [55], [56]. However, no prior work has successfully copied the success of active acoustic sensing from mobile devices to head-mounted VR devices. Furthermore, the above works all require the recorded audio feedback to keep the major pattern of the original signal, while sensing with distorted signals after echo cancellation is still an unsolved challenge.

There are several studies using passive sensing for VR user authentication, such as verifying the user's voice commands [57] or capturing the user's subtle facial dynamics with inertial sensors when the user is speaking [58]. These methods still require the user's active participation and are more easily affected by environmental noises than active sensing. Schneegass *et al.* use the bone conduction speaker and microphone to achieve active acoustic sensing on Google Glasses and capture the bone-conducted sound through the user's skull as biometric [59]. Isobe *et al.* develop an eyeglass prototype with a pair of microphone and speaker on the nose pads, which extracts individual nose features using acoustic signals for authentication [11]. However, these methods require dedicated hardware and cannot be deployed on commodity VR devices. Some existing works use Head-Related Transfer Function (HRTF) and ear canal biometrics for authentication [60], [61]. HRTFs describe the speaker-to-ear channels with in-air propagated sounds, and the raw audio data is needed to estimate frequency responses. Differently, our sensing model uses onboard speakers and microphones and analyzes distorted signals, while ear information and earphones are not needed.

VII. Conclusion

This work proposes an efficient and replay-resistant VR user authentication system based on acoustic-domain head biometrics. The system interacts with the user via active acoustic sensing and captures the unique skull-reverberated sounds for authentication. To deploy the system on commodity VR headsets, we address the challenge of acoustic sensing under echo cancellation and develop the CAE-CNN algorithm. The CAE component reconstructs the spectrogram of the received echo sound to recover the head biometrics and counteract the effect of echo cancellation. Then the CNN component learns from the reconstructed head biometric spectrogram to build the per-user model and distinguish each user. We show that the current echo cancellation mechanism is not a hindrance to active acoustic sensing but is naturally a good mechanism to prevent acoustic relay attacks. Experiments with two VR device models (w/wo echo cancellation) and over a one-year long-term study show that our system efficiently verifies single or multiple users and is resistant to replay attacks.

Acknowledgments

This work was supported in part by LABoR LEQSF(2020-23)-RD-A-11, NSF CNS-2155131, and the LSU Provost's Fund for Innovation in Research - Faculty Travel Grant Program.

References

[1] V. R. H. S. E. Growth, "Smartphone apac market forecast 2014 - 2018," July. 2021. [Online]. Available: https://thejournal.com/articles/2021/07/01/virtual-reality-headsets-see-explosive-growth.aspx

[2] I. Trackers, "Worldwide quarterly augmented and virtual reality headset tracker," IDC, 2021. [Online]. Available: https://www.idc.com/tracker/showproductinfo.jsp?containerId=IDC_P35095

[3] F. Mathis, H. I. Fawaz, and M. Khamis, "Knowledge-driven biometric authentication in virtual reality," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–10.

[4] F. Mathis, J. Williamson, K. Vaniea, and M. Khamis, "Rubikauth: Fast and secure authentication in virtual reality," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–9.

[5] T. Mustafa, R. Matovu, A. Serwadda, and N. Muirhead, "Unsure how to authenticate on your vr headset? come on, use your head!" in *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics*, 2018, pp. 23–30.

[6] K. Pfeuffer, M. J. Geiger, S. Prange, L. Mecke, D. Buschek, and F. Alt, "Behavioural biometrics in vr: Identifying people from body motion and relations in virtual reality," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–12.

[7] M. Sivasamy, V. Sastry, and N. Gopalan, "Vrcauth: continuous authentication of users in virtual reality environment using head-movement," in *2020 5th International Conference on Communication and Electronics Systems (ICCES)*. IEEE, 2020, pp. 518–523.

[8] X. Wang and Y. Zhang, "Nod to auth: Fluent ar/vr authentication with user head-neck modeling," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–7.

[9] H. Chen, W. Wang, J. Zhang, and Q. Zhang, "Echoface: Acoustic sensor-based media attack detection for face authentication," *IEEE Internet of Things Journal*, vol. 7, no. 3, pp. 2152–2159, 2019.

[10] B. Zhou, Z. Xie, Y. Zhang, J. Lohokare, R. Gao, and F. Ye, "Robust human face authentication leveraging acoustic sensing on smartphones," *IEEE Transactions on Mobile Computing*, 2021.

[11] K. Isobe and K. Murao, "Person-identification methodusing active acoustic sensing applied to nose," in *2021 International Symposium on Wearable Computers*, 2021, pp. 138–140.

[12] R. Wang, L. Huang, and C. Wang, "Preventing handheld phone distraction for drivers by sensing the gripping hand," in *2021 IEEE 18th International Conference on Mobile Ad Hoc and Smart Systems (MASS)*. IEEE, 2021, pp. 410–418.

[13] L. Huang and C. Wang, "Pcr-auth: Solving authentication puzzle challenge with encoded palm contact response," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 1034–1048.

[14] A. Lovitt, A. J. Miller, P. Robinson, and S. Selfon, "Adaptive anc based on environmental triggers," Apr. 26 2022, uS Patent 11,315,541.

[15] O. VR, "Qsimulating dynamic soundscapes at facebook reality labs," Meta Quest, 10/25/2018. [Online]. Available: https://www.oculus.com/blog/simulating-dynamic-soundscapes-at-facebook-reality-labs/

[16] Wikipedia. (2021) Rigid body. [Online]. Available: https://en.wikipedia.org/wiki/Rigidbody/

[17] A. L. Ruina and R. Pratap, *Introduction to statics and dynamics*. Preprint for Oxford University Press, 2002.

[18] Unity, "Unity scripting api - microphone," Aug 4 2022. [Online]. Available: https://docs.unity3d.com/ScriptReference/Microphone.html

[19] S. V. A. Gari, P. Robinson, and J. R. Donley, "Sound level reduction and amplification," Apr. 6 2021, uS Patent 10,971,130.

[20] S. Dixit and D. Nagaria, "Lms adaptive filters for noise cancellation: A review," *International Journal of Electrical and Computer Engineering*, vol. 7, no. 5, p. 2520, 2017.

[21] S. Malik and G. Enzner, "State-space frequency-domain adaptive filtering for nonlinear acoustic echo cancellation," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 7, pp. 2065–2079, 2012.

[22] Y. Michalevsky, D. Boneh, and G. Nakibly, "Gyrophone: Recognizing speech from gyroscope signals," in *23rd USENIX Security Symposium (USENIX Security 14)*, 2014, pp. 1053–1067.

[23] F. Küch, W. Kellermann, E. Hansler, and G. Schmidt, "Nonlinear acoustic echo cancellation," *Topics in Acoustic Echo and Noise Control*, pp. 205–257, 2006.

[24] Vocal, "Nonlinear acoustic echo cancellation," October 2022. [Online]. Available: https://vocal.com/echo-cancellation/nonlinear-aec/

[25] R. U. u/thebossne55, "Oculus quest not detecting being worn," Reddit, 5/18/2021. [Online]. Available: https://www.reddit.com/r/OculusQuest/comments/i3mica/oculus_quest_not_detecting_being_worn/

[26] R. U. u/BaconSock, "The sensor in my quest 2 stops detecting my head sometimes," Reddit, 6/9/2021. [Online]. Available: https://www.reddit.com/r/oculus/comments/lv6han/the_sensor_in_my_quest_2_stops_detecting_my_head/

[27] J. R. Velasco, "3d scanning, 3d modeling and 3d printing a human head," ALL3DP, 3/28/2019. [Online]. Available: https://all3dp.com/2/3d-scanning-3d-modeling-3d-printing-a-human-head-how-to/

[28] Revopoint, "Pop 2 high-precision 3d scanner," Feb 25 2022. [Online]. Available: https://shop.revopoint3d.com/collections/3d-scanners/products/pop2-3d-scanner?variant=42240758546667

[29] Creality, "Creality3d cr-10 v3 3d printer - creality 3d," April 22 2020. [Online]. Available: https://creality3d.shop/products/creality3d-cr-10-v3-3d-printer

[30] N. Erdogmus and S. Marcel, "Spoofing face recognition with 3d masks," *IEEE transactions on information forensics and security*, vol. 9, no. 7, pp. 1084–1097, 2014.

[31] Amazon, "Silicone mold making kit," Jan 22 2022. [Online]. Available: https://www.amazon.com/Marvelous-Making-Liquid-Silicone-Rubber/dp/B08QV5ZR79

[32] R. Vasudevan and C. G. Gordon, "Experimental study of annoyance due to low frequency environmental noise," *Applied Acoustics*, vol. 10, no. 1, pp. 57–69, 1977.

[33] G. Leventhall, P. Pelmear, and S. Benton, "A review of published research on low frequency noise and its effects," 2003.

[34] Y.-C. Tung and K. G. Shin, "Echotag: Accurate infrastructure-free indoor location tagging with smartphones," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, 2015, pp. 525–536.

[35] J. Lian, J. Lou, L. Chen, and X. Yuan, "Echospot: Spotting your locations via acoustic sensing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 3, pp. 1–21, 2021.

[36] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[37] S. Eberz, G. Lovisotto, K. B. Rasmussen, V. Lenders, and I. Martinovic, "28 blinks later: Tackling practical challenges of eye movement biometrics," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019, pp. 1187–1199.

[38] C. George, M. Khamis, E. von Zezschwitz, M. Burger, H. Schmidt, F. Alt, and H. Hussmann, "Seamless and secure vr: Adapting and evaluating established authentication systems for virtual reality." NDSS, 2017.

[39] S. Li, A. Ashok, Y. Zhang, C. Xu, J. Lindqvist, and M. Gruteser, "Whose move is it anyway? authenticating smart wearable devices using unique head movement patterns," in *2016 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 2016, pp. 1–9.

[40] R. Miller, A. Ajit, N. K. Banerjee, and S. Banerjee, "Realtime behavior-based continual authentication of users in virtual reality environments," in *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. IEEE, 2019, pp. 253–2531.

[41] Y. Shen, H. Wen, C. Luo, W. Xu, T. Zhang, W. Hu, and D. Rus, "Gaitlock: Protect virtual and augmented reality headsets using gait," *IEEE Transactions on Dependable and Secure Computing*, vol. 16, no. 3, pp. 484–497, 2018.

[42] I. Olade, C. Fleming, and H.-N. Liang, "Biomove: Biometric user identification from human kinesiological movements for virtual reality systems," *Sensors*, vol. 20, no. 10, p. 2944, 2020.

[43] Y. Qifan, T. Hao, Z. Xuebing, L. Yin, and Z. Sanfeng, "Dolphin: Ultrasonic-based gesture recognition on smartphone platform," in *2014 IEEE 17th International Conference on Computational Science and Engineering*. IEEE, 2014, pp. 1461–1468.

[44] B. Van Dam, Y. Murillo, M. Li, and S. Pollin, "In-air ultrasonic 3d-touchscreen with gesture recognition using existing hardware for smart devices," in *2016 IEEE International Workshop on Signal Processing Systems (SiPS)*. IEEE, 2016, pp. 74–79.

[45] Y.-C. Tung and K. G. Shin, "Expansion of human-phone interface by sensing structure-borne sound propagation," in *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, 2016, pp. 277–289.

[46] K. Sun, T. Zhao, W. Wang, and L. Xie, “Vskin: Sensing touch gestures on surfaces of mobile devices using acoustic signals,” in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, 2018, pp. 591–605.

[47] Y. Wang, J. Shen, and Y. Zheng, “Push the limit of acoustic gesture recognition,” *IEEE Transactions on Mobile Computing*, 2020.

[48] R. Nandakumar, S. Gollakota, and N. Watson, “Contactless sleep apnea detection on smartphones,” in *Proceedings of the 13th annual international conference on mobile systems, applications, and services*, 2015, pp. 45–57.

[49] K. Qian, C. Wu, F. Xiao, Y. Zheng, Y. Zhang, Z. Yang, and Y. Liu, “Acousticcardiogram: Monitoring heartbeats using acoustic signals on smart devices,” in *IEEE INFOCOM 2018-IEEE conference on computer communications*. IEEE, 2018, pp. 1574–1582.

[50] J. N. Moutinho, R. E. Araújo, and D. Freitas, “Indoor localization with audible sound—towards practical implementation,” *Pervasive and Mobile Computing*, vol. 29, pp. 1–16, 2016.

[51] I. Rishabh, D. Kimber, and J. Adcock, “Indoor localization using controlled ambient sounds,” in *2012 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. IEEE, 2012, pp. 1–10.

[52] P. Lazik and A. Rowe, “Indoor pseudo-ranging of mobile devices using ultrasonic chirps,” in *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems*, 2012, pp. 99–112.

[53] F. Höflinger, J. Hoppe, R. Zhang, A. Ens, L. Reindl, J. Wendeberg, and C. Schindelhauer, “Acoustic indoor-localization system for smart phones,” in *2014 IEEE 11th International Multi-Conference on Systems, Signals & Devices (SSD14)*. IEEE, 2014, pp. 1–4.

[54] Y. Gao, W. Wang, V. V. Phoha, W. Sun, and Z. Jin, “Earecho: Using ear canal echo for wearable authentication,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, pp. 1–24, 2019.

[55] L. Huang and C. Wang, “Notification privacy protection via unobtrusive gripping hand verification using media sounds,” in *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, 2021, pp. 491–504.

[56] C. Wang, J. Mu, and L. Huang, “Protecting smartphone screen notification privacy by verifying the gripping hand,” in *Proceedings of the 2020 ACM Workshop on Information Hiding and Multimedia Security*, 2020, pp. 49–54.

[57] G. Peng, G. Zhou, D. T. Nguyen, X. Qi, Q. Yang, and S. Wang, “Continuous authentication with touch behavioral biometrics and voice on wearable glasses,” *IEEE transactions on human-machine systems*, vol. 47, no. 3, pp. 404–416, 2016.

[58] C. Shi, X. Xu, T. Zhang, P. Walker, Y. Wu, J. Liu, N. Saxena, Y. Chen, and J. Yu, “Face-mic: inferring live speech and speaker identity via subtle facial dynamics captured by ar/vr motion sensors,” in *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, 2021, pp. 478–490.

[59] S. Schneegass, Y. Oualil, and A. Bulling, “Skullconduct: Biometric user identification on eyewear computers using bone conduction through the skull,” in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016, pp. 1379–1384.

[60] Z. Yang and R. R. Choudhury, “Personalizing head related transfer functions for earables,” in *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*, 2021, pp. 137–150.

[61] S. Orike, B. Bakare, and C. Iyoloma, “A unique identification system using ear acoustics biometrics.”

APPENDIX

A. Security Gain of CAE

To understand how effectively our CAE-CNN algorithm addresses cross-day variations (e.g., noises) and counteracts the effect of echo cancellation, we present the user verification performance on the standalone VR device when CAE-CNN and CNN are used, respectively. The True Positive Rate (TPR) and True Negative Rate (TNR) of each of the 30 participants are shown in Table VI and Table VII. We find that the user verification performance is significantly improved by 11% TPR and 13% TNR after adding the CAE model, compared to using the CNN model alone to recognize the participant's head biometrics. Specifically, when using the CNN model alone, the system achieves 89.0% TPR and 88.86% TNR on average. In comparison, our CAE-CNN algorithm improves to 98.90% TPR and 98.82% TNR on average. The results confirm our proposed system's high accuracy and robustness in identifying users on commodity VR devices.

TABLE VI
META QUEST VERIFICATION PERFORMANCE (TPR) BEFORE AND AFTER APPLYING AN AUTO ENCODER.

Model	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	U12	U13	U14	U15
CNN	94.50%	96.19%	89.10%	82.18%	82.76%	95.0%	90.0%	86.25%	93.97%	96.83%	79.13%	97.48%	82.74%	90.78%	92.96%
CAE-CNN	98.25%	99.45%	98.82%	99.01%	100%	98.75%	98.75%	98.75%	98.28%	98.41%	99.27%	98.32%	99.78%	99.31%	98.59%

Model	U16	U17	U18	U19	U20	U21	U22	U23	U24	U25	U26	U27	U28	U29	U30
CNN	97.22%	91.74%	82.06%	76.36%	77.50%	85.34%	97.92%	95.58%	81.00%	98.91%	94.97%	77.51%	94.74%	90.70%	78.57%
CAE-CNN	99.07%	100%	99.71%	99.74%	99.17%	99.14%	100%	99.12%	98.00%	97.38%	99.74%	97.60%	97.37%	99.48%	97.62%

TABLE VII
META QUEST VERIFICATION PERFORMANCE (TNR) BEFORE AND AFTER APPLYING AN AUTO ENCODER.

Model	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	U12	U13	U14	U15
CNN	93.79%	95.69%	90.52%	80.86%	83.10%	95.13%	91.21%	87.24%	94.14%	95.86%	81.21%	97.76%	79.83%	89.56%	91.90%
CAE-CNN	97.93%	99.31%	98.79%	98.45%	100%	98.05%	99.48%	98.79%	97.76%	98.28%	99.14%	98.10%	99.14%	98.52%	99.48%

Model	U16	U17	U18	U19	U20	U21	U22	U23	U24	U25	U26	U27	U28	U29	U30
CNN	96.21%	88.62%	81.90%	79.31%	76.90%	88.45%	96.55%	95.86%	81.03%	98.10%	95.00%	80.17%	93.97%	87.93%	77.94%
CAE-CNN	99.48%	100%	98.79%	98.79%	99.66%	99.14%	100%	98.97%	98.62%	97.41%	99.31%	97.41%	97.41%	99.48%	98.86%