SAFARI: Speech-Associated Facial Authentication for AR/VR Settings via Robust VIbration Signatures

Tianfang Zhang*
Rutgers University
New Brunswick, New Jersey, USA
tz203@scarletmail.rutgers.edu

Qiufan Ji*
New Jersey Institute of Technology
Newark, New Jersey, USA
qj39@njit.edu

Zhengkun Ye
Temple University
Philadelphia, Pennsylvania, USA
zhengkun.ye@temple.edu

Md Mojibur Rahman Redoy Akanda
Texas A&M University
College Station, Texas, USA
redoy.akanda@tamu.edu

Ahmed Tanvir Mahdad
Texas A&M University
College Station, Texas, USA
mahdad@tamu.edu

Cong Shi
New Jersey Institute of Technology
Newark, New Jersey, USA
cong.shi@njit.edu

Yan Wang
Temple University
Philadelphia, Pennsylvania, USA
y.wang@temple.edu

Nitesh Saxena
Texas A&M University
College Station, Texas, USA
nsaxena@tamu.edu

Yingying Chen†
Rutgers University
New Brunswick, New Jersey, USA
yingche@scarletmail.rutgers.edu

Abstract

In AR/VR devices, the voice interface, serving as one of the primary AR/VR control mechanisms, enables users to interact naturally using speeches (voice commands) for accessing data, controlling applications, and engaging in remote communication/meetings. Voice authentication can be adopted to protect against unauthorized speech inputs. However, existing voice authentication mechanisms are usually susceptible to voice spoofing attacks and are unreliable under the variations of phonetic content. In this work, we propose SAFARI, a spoofing-resistant and text-independent speech authentication system that can be seamlessly integrated into AR/VR voice interfaces. The key idea is to elicit phonetic-invariant biometrics from the facial muscle vibrations upon the headset. During speech production, a user's facial muscles are deformed for articulating phoneme sounds. The facial deformations associated with the phonemes are referred to as visemes. They carry rich biometrics of the wearer's muscles, tissue, and bones, which can propagate through the head and vibrate the headset. SAFARI aims to derive reliable facial biometrics from the viseme-associated facial vibrations captured by the AR/VR motion sensors. Particularly, it identifies the vibration data segments that contain rich viseme patterns (prominent visemes) less susceptible to phonetic variations. Based on the prominent visemes, SAFARI learns on the correlations among facial vibrations of different frequencies to extract biometric representations invariant to the phonetic context. The key advantages of SAFARI are that it is suitable for commodity AR/VR headsets (no additional sensors) and is resistant to voice spoofing attacks as the conductive property of the facial vibrations prevents biometric disclosure via the air media or the audio channel. To mitigate the impacts of body motions in AR/VR scenarios, we also design a generative diffusion model trained to reconstruct the viseme patterns from the data distorted by motion artifacts. We conduct extensive experiments with two representative AR/VR headsets and 35 users under various usage and attack settings. We demonstrate that SAFARI can achieve over 96% true positive rate on verifying legitimate users while successfully rejecting different kinds of spoofing attacks with over 97% true negative rates.

CCS Concepts

• **Security and privacy → Authentication**.

Keywords

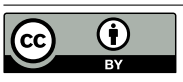
Authentication, AR/VR headsets, Speech vibrations

ACM Reference Format:
Tianfang Zhang, Qiufan Ji, Zhengkun Ye, Md Mojibur Rahman Redoy Akanda, Ahmed Tanvir Mahdad, Cong Shi, Yan Wang, Nitesh Saxena, and Yingying Chen. 2024. SAFARI: Speech-Associated Facial Authentication for AR/VR Settings via Robust VIbration Signatures. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24), October 14–18, 2024, Salt Lake City, UT, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3658644.3670358

1 Introduction

With the capability to deliver immersive and interactive experiences, face-mounted AR/VR devices have emerged as prominent contenders to personal computers. Leading technology companies (e.g., Apple [1], Meta [2]) are at the forefront of promoting spatial computing [47], a paradigm where users can interact with digital media/programs displayed in a 3D virtual space through gestures and voice. This paradigm shift inevitably migrates a large volume of sensitive data and functionalities from computers and

*Co-primary authors.
†Yingying Chen is the corresponding author.



CCS '24, October 14–18, 2024, Salt Lake City, UT, USA

ACM ISBN 979-8-4007-0636-3/24/10
https://doi.org/10.1145/3658644.3670358

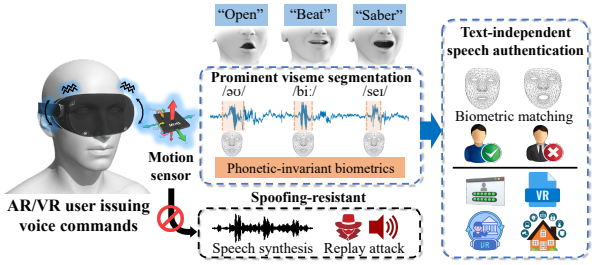


Figure 1: Illustration of the proposed AR/VR speech authentication system based on facial vibrations of visemes (i.e., the facial couterparts of phonemes).

mobile devices into AR/VR headsets, such as accounts, photos, financial records, and medical information. To protect the security and privacy of AR/VR users, voice authentication [44, 46, 56, 60] has emerged as a promising technology. The authentication mechanism leveraging the voice biometrics can be applied to voice commands to access the sensitive data or control the AR/VR programs. The user can also be swiftly authenticated during voice communication (e.g., virtual meetings, virtual social interactions) without interrupting the workflow or communication. Therefore, voice authentication is considered as both transparent and intuitive for AR/VR users.

However, the adoption of voice authentication in AR/VR platforms faces two key problems. *(i) Susceptibility to Voice Spoofing Attacks:* The open nature of sound propagation leaves voice authentication extremely vulnerable to voice spoofing attacks. In AR/VR settings, where remote voice communications are prevalent, an adversary can easily obtain the victim's voice samples through a shared audio channel (e.g., a VR voice chat, a virtual meeting, a social interaction session) or the victim's public speeches. The voice samples can then be used to reproduce or synthesize speech with the user's voice biometrics, bypassing the authentication mechanism [12, 27, 41]. *(ii) Variability in Phonetic Patterns:* Voice biometrics heavily depend on speech content, specifically the phonemes, which are the smallest sound units in speech. The biometric representations of a phoneme (e.g., formant frequencies and spectral characteristics) can vary significantly depending on its phonetic context (i.e., the other phonemes around it), which is called the co-articulation effect [21]. Therefore, text-independent voice authentication that identifies a user regardless of the speech content typically requires extensive training voice data [22, 45].

In this work, we introduce SAFARI, the first spoofing-resistant and text-independent speech authentication system for AR/VR headsets. Our system can be seamlessly integrated into mainstream headsets to secure voice inputs, such as those used in voice dictation, navigations, and app controls. The key idea of SAFARI is to capture facial geometry deformations during speeches by leveraging minute facial vibrations upon the headset. We illustrate SAFARI in Figure 1. During speech production, a user's face geometry is deformed due to the movements of facial muscles for articulating phoneme sounds. Such facial deformations are referred to as visemes, the facial counterparts of phonemes [11]. As the headset is mounted on the user's head, these deformations can induce minute vibrations upon the headset, thereby encoding the viseme patterns into the motion sensor readings. Visemes are consistent across speech content at two levels [8]: First, visemes have less diversity compared to phonemes, as multiple phonemes that appear visually similar when spoken are grouped under a single viseme [8, 11]. For instance, the phonemes /p/, /b/, and /m/ have the same facial deformations where the mouth closes at the beginning, followed by a release of air pressure. The facial deformations of the three phonemes are mapped into one viseme $V_{P,M,B}$. Second, the same viseme shows highly consistent facial deformations across different speech content (e.g., viseme /b/ in "begin", "browse", and "battery"). These two properties of visemes allow SAFARI to profile facial vibrations on only a small set of visemes (i.e., 11 visemes in English) instead of extensive voice data covering various speech content [22, 45]. More importantly, visemes carry rich biometric characteristics of the user's facial muscles, tissues, and bones, which are confined to the human body. The internal propagation mechanism of viseme-associated facial vibrations makes our system resilient to voice spoofing attacks relying on voice biometric theft via the air media or an audio channel. Even when the facial vibrations can be acquired (e.g., through a malicious AR/VR app), it is difficult for the adversary to reproduce the same vibrations on the headset.

Capturing effective viseme patterns poses significant challenges in AR/VR headsets. Traditional vision-based methods [4, 17] that rely on cameras to record images/videos of faces are impractical in AR/VR headsets equipped only with outward-facing cameras. Therefore, SAFARI utilizes the headset's built-in motion sensors to capture facial vibrations, thereby sensing visemes. While motion sensors are insensitive to airborne sounds [5], they can pick up conductive facial vibrations induced by visemes. This conductive property also makes SAFARI resilient to acoustic interferences in the environment (e.g., ambient noises and speeches of nearby people). The design of AR/VR headsets covers only the upper region of the face (e.g., cheeks, nose bridge, and forehead). The lower facial areas, particularly near the mouth, lower jaw, and chin, tend to produce weaker and less consistent vibration patterns, which compromises viseme sensing via facial vibrations. To overcome this problem, SAFARI focuses on identifying facial vibration segments that exhibit pronounced and consistent viseme patterns. We find that the formation of a vowel viseme coupled with its adjacent consonant viseme often involves a complete mouth open-and-close cycle, such as /bɪ/ in "begin" and /məʊ/ in "motion". We refer to these viseme combinations as prominent visemes, characterized by substantial deformations in the upper face region and a richer array of biometric properties (e.g., muscle movement patterns and facial structures). Centered on the prominent viseme segments, we develop a correlation learning scheme that contrasts different frequency components within the prominent viseme, deriving phonetic-invariant facial biometrics for speech authentication.

We face several challenges to realize SAFARI: *(1) Significant distortions caused by body motions:* In AR/VR environments, users may engage in strong and continuous body motions while utilizing SAFARI (e.g., playing games, exploring the virtual world). Our system should recover subtle viseme patterns from significant distortions to ensure reliable authentication. *(2) Difficulty in prominent viseme segmentation:* To enable reliable authentication, SAFARI needs to identify regions of prominent visemes. However, the smooth and continuous transitions between visemes make the segmentation particularly challenging. We must develop algorithms that accurately detect the starting and ending points from the prominent

visemes. *(3) Unknown biometrics related to visemes:* The influence of viseme-related biometrics in facial vibrations has not been studied in prior work. It is necessary to extract reliable viseme-associated biometric representations from the vibration patterns.

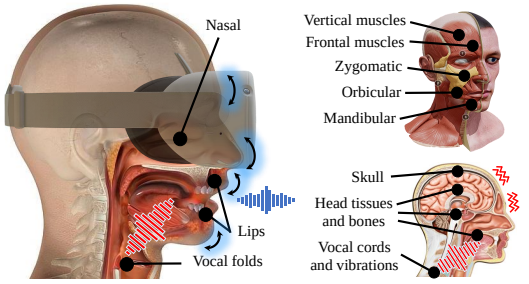
To mitigate the motion-induced distortions, we design a diffusion-based generative model for viseme reconstruction. This model is trained by iteratively adding the noises of different body motions in AR/VR settings, and it reconstructs the motion sensor data to the original state unaffected by body motions. In addition, we design a two-step scheme to identify and segment temporal regions of prominent visemes with rich biometrics, each containing a complete mouth open-and-close cycle. The first step involves detecting vowel sounds by analyzing high-frequency facial vibrations captured in the motion sensor readings, which occurs predominantly during speech. This approach effectively distinguishes visemes from non-speech facial expressions (e.g., smiling, showing anger). In the second step, SAFARI locates the temporal positions of mouth opening/closing around these vowel sounds based on low-frequency facial movements. Furthermore, SAFARI utilizes correlation learning for deriving facial biometrics that are phonetically consistent. This strategy contrasts low-frequency facial movements with high-frequency facial vibrations, providing a dual-perspective analysis of each prominent viseme. The correlation will dynamically weight the stable components within the facial vibrations. Based on the weighted vibrations, a transformer model is utilized to extract phonetic-invariant facial biometrics for text-independent authentication. Our main contributions are summarized as follows:

- We present SAFARI, a spoofing-resistant and text-independent speech authentication system that can be seamlessly integrated into mainstream AR/VR devices. It is the first work that shows distinctive phonetic-invariant and viseme-associated facial biometrics can be extracted using built-in AR/VR motion sensors.
- We design a generative diffusion model, which reconstructs viseme patterns to a state unaffected by human motions. We further develop a scheme to identify and segment prominent visemes with rich viseme patterns for speech authentication.
- We develop a correlation learning strategy with a transformer architecture to reliably link viseme patterns with users' unique facial biometrics. Through contrasting different frequency components within the prominent visemes, the strategy extracts biometric representations that are distinctive for individual users while being invariant to the phonetic context.
- We validate SAFARI by conducting extensive experiments using two commercial AR/VR headsets on 35 users with ages ranging from 18 to 37, including native and non-native English speakers. Through training on 20 short voice commands, SAFARI can achieve over 96% true positive rates in authenticating enrolled users. The system can also successfully defend against blind attack, vibration replay attack, and observe-and-mimic attack with over 97% true negative rates.

2 Preliminaries

2.1 Kinetics of Viseme Production

Visemes are produced during speech articulation. As depicted in Figure 2, they are shaped by the movements and vibrations of facial muscles, synchronized with the distinct phoneme sounds in speech.



(a) Speech-induced facial muscle vibrations (b) Facial muscle and head structure

Figure 2: Illustration of facial muscle vibrations of visemes during human speech articulation.

Thus, visemes are the facial counterparts of phonemes. The viseme patterns are unique to each user, distinguished by facial characteristics such as facial landmarks, cheekbones, and nose structure. The deformations primarily result from the facial muscle movements. Specifically, five categories of muscles contribute to viseme production: orbicular, zygomatic, mandibular, frontal, and vertical muscles [8]. During speech production, the frontal and zygomatic muscles in the upper face areas influence the headset's position and orientation, thus affect the motion sensor readings. The orbicular, mandibular, and vertical muscles in the lower face areas may indirectly affect the motion sensor readings by altering the tension and shape of adjacent muscles and tissues. However, the visemes only involving these muscles, particularly many consonant visemes (e.g., /f/, /r/, /w/), tend to produce weak and inconsistent facial vibration patterns. Besides, the facial muscles also carry minute vibrations originated from vocal organs (e.g., vocal cords and vocal tract), which serve as an auditory element of visemes. While these internal muscle vibrations are not visible on the user's face, they are closely correlated with the visemes at the user's face. These vibrations also carry unique biometric information about the user's vocal organs and muscle structures, resulting in distinctive vibration patterns for different users producing the same viseme.

2.2 Sensing Viseme-associated Facial Vibrations

To study the feasibility of viseme sensing through facial vibrations, we conduct preliminary experiments using Meta Quest, which is one of the most popular VR headsets. The headset is equipped with a three-axis accelerometer and gyroscope. In SAFARI, we mainly leverage the accelerometer given its better sensitivity to vibrations [57]. The sampling rate is set to 1000Hz. In the experiments, a participant wearing the Meta Quest is asked to pronounce four distinct phonemes (visemes): /p/ ($V_{P,B,M}$), /d/ ($V_{D,T,S}$), /əʊ/ (V_A), and /ɪ/ (V_I). We simultaneously collect data from both the headset's accelerometer and microphone. We visualize the time-frequency spectrograms of the accelerometer and audio data of the phonemes in Figure 3. Particularly, the temporal regions of viseme articulation have markedly higher spectrum energy compared to those where the viseme is absent. Therefore, we utilize the spectrum energy as a baseline to locate the viseme patterns in the motion sensor readings. Moreover, we observe that the spectrograms of accelerometer exhibit strong energy at a low-frequency range (i.e., $\leq$ 100Hz) during phoneme production. Different from phoneme sounds at higher frequency, these patterns are associated with visemes, specifically

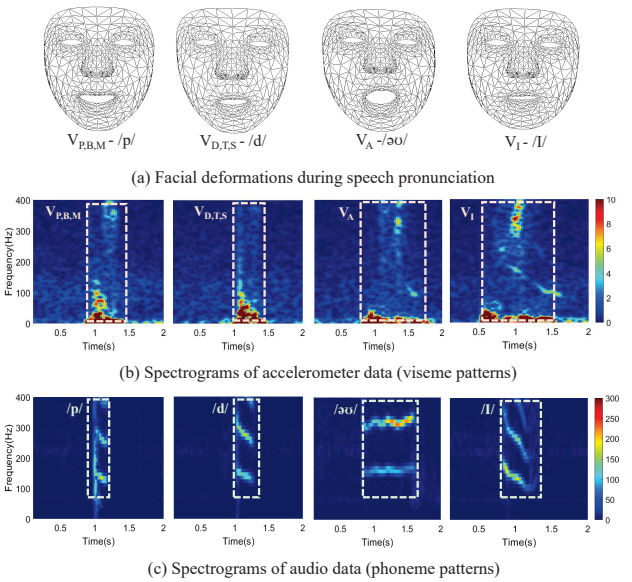


Figure 3: Spectrograms of viseme-associated facial vibrations captured by the accelerometer (Z-axis) and phoneme-related voice sound recorded by the microphone. The facial vibrations cannot be recorded by the microphone.

the facial muscle movements to articulate the phonemes. These findings show that the headset's accelerometer can indeed respond to individual visemes. Further comparisons between accelerometer readings and audio spectrograms reveal that these low-frequency responses to visemes are not captured by the headset's microphone, which is designed to pick up air pressure changes rather than conductive movement patterns. In addition to muscle movements, the accelerometer also detects high-frequency vibrations (i.e., $\geq$ 100Hz) transmitted through the muscles. Originated from the vocal cords, these vibrations undergo complex attenuation, reflection, and refraction within the head, thus containing unique facial muscle characteristics. It exhibits substantial different frequency responses compared to those recorded in the microphone. These observations confirm the feasibility of capturing visemes through facial vibration sensing on AR/VR headsets. These studies also validate the internal propagation characteristics of facial vibrations, which prevents the leakage of viseme biometrics via the audio channel.

3 Threat Model

We consider an adversary who targets private information (e.g., accounts, photos, financial records) or unauthorized operations (e.g., making payments, installing malware) on the user's AR/VR devices. We assume the adversary is familiar with the authentication mechanism of SAFARI and can wear the user's AR/VR headset. Based on the prior knowledge and techniques available to the adversary, we categorize the following three attack types:

Blind Attack. The adversary does not have any prior knowledge on the viseme patterns of legitimate users. To deploy the attack, the attackers wear the user's headset and use voice interactions with their own visemes, with the expectation that the random facial expression might bypass the user authentication scheme.

Table 1: 11 different visemes with their corresponding phoneme sets and representative word examples.

Consonants		
Viseme	**Phoneme**	**Word examples**
$V_{J,C,H}$	/ʤ/, /ʧ/, /ʃ/, /ʒ/	**j**ump, **ch**at, mo**t**ion, vi**s**ion
$V_{P,M,B}$	/p/, /b/, /m/	**p**ick, **b**it, **m**ake
$V_{F,V}$	/f/, /v/	**f**at, **v**alue
$V_{R,W}$	/r/, /w/	**r**un, **w**ater
$V_{D,T,S}$	/d/, /t/, /s/, /z/, /θ/, /ð/	**d**esk, **t**ake, **s**ad, **z**oom, **th**ink, **th**at
$V_{G,K,N}$	/g/, /k/, /n/, /ŋ/, /l/, /y/, /h/	**g**ap, **c**at, **n**et, pi**ng**, **l**ip, **y**es, **h**as
Vowels		
Viseme	**Phoneme**	**Word examples**
V_A	/aː/, /aʊ/, /aɪ/, /ʌ/	**c**ar, **ou**t, **fl**y, **c**up
V_E	/e/, /eɪ/,/æ/, /ə/, /eə/, /ɜː/	**e**gg, **s**ave, **a**pple
V_I	/iː/, /ɪə/, /ɪ/	**b**e**a**t, **sh**ip
V_O	/ɔː/, /ɔɪ/, /əʊ/, /ɑ/	d**oo**r, b**oy**, n**o**se
V_U	/ʊ/, /ʊə/, /uː/	b**oo**k, b**oo**t

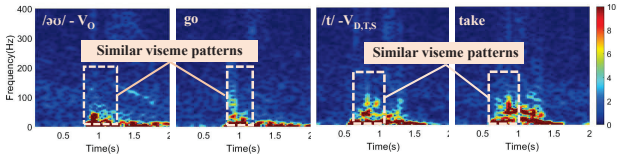
Vibration Replay Attack. We consider the situations in which the adversary can obtain the users' voice samples. This can be achieved by stealthily recording the user's sound using a microphone. He/she replays the voice recordings using a playback device (e.g., a smartphone) in direct contact with the headset. The adversary hopes the conductive vibrations of the playback device can result in patterns similar to facial vibrations. Although the motion sensors do not show significant response to air-borne sounds [5], they can capture the direct vibrations. The adversary can also leverage voice synthesis techniques [41, 50] to generate voice samples with the same speech content and voice biometrics of the legitimate users for deploying vibration replay attack.

Observe-and-mimic Attack. For this attack, we assume that the adversary can observe the visemes of legitimate users while they use SAFARI, which can be realized via observation or video tapping. The adversary then attempts to imitate the facial deformations of legitimate users while they pronounce speech. Note that the adversary can also have a similar face shape with the victim user.

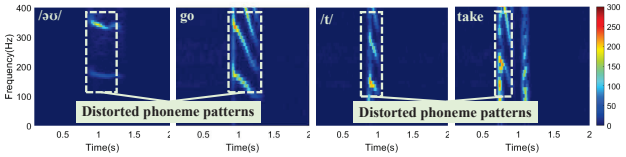
4 System Design

4.1 Enabling Text-independent Speech Authentication via Viseme Profiling

The idea of SAFARI is to learn phonetic-invariant facial biometrics associated with visemes. The utilization of visemes is beneficial in this task on two levels. On the first level, visemes are inherently less diverse than phonemes due to their many-to-one relationship. As illustrated in Table 1, facial deformations corresponding to 44 phonemes are typically categorized into 11 distinct visemes, based on the similarities among the facial landmarks of visemes [28, 30]. This reduced diversity in visemes makes it easier for SAFARI in profiling the entire range of viseme-related facial biometrics. For instance, the viseme V_E includes six phonemes (/e/, /eɪ/, /æ/, /ə/, /eə/, /ɜː/), all sharing similar facial deformations and associated biometrics. On the second level, visemes exhibit strong consistency in facial deformations across different speech contexts. To demonstrate this, we analyze the spectrograms of isolated visemes and compare them with the same visemes within speech contexts, such as /əʊ/ (V_O) in "go" and /t/ ($V_{D,T,S}$) in "take". We ask a volunteer to pronounce these two viseme-word pairs and show the spectrograms in Figure 4. We can observe that visemes maintain their



(a) Facial vibrations of isolated viseme vs. viseme in a word



(b) Voice data of isolated phoneme vs. phoneme in a word

Figure 4: Comparisons between visemes and phonemes in isolated units and different words.

facial vibration patterns even when articulated within the words, showing their consistency under phonetic variations. In contrast, the vibration patterns of phonemes like /əʊ/ and /t/ are significantly influenced by adjacent phonemes due to co-articulation effects. The two levels of consistency in viseme-associated facial vibrations benefit SAFARI in realizing text-independent speech authentication.

4.2 Challenges

Significant Distortions Caused by Body Motions. While using SAFARI, users may interact with AR/VR devices with body gestures (e.g., looking around, interacting with virtual objects). These artifacts generated by body motions can significantly distort the viseme patterns in motion sensor readings. A simple band-pass filter cannot effectively remove these artifacts given that the responses overlap with the frequency of facial muscle movements (e.g., 0~100Hz). Therefore, it is essential to recover the viseme patterns from such significant motion artifacts to enable reliable authentication.

Difficulty in Prominent Viseme Segmentation. Prominent visemes do not exhibit clear and consistent boundaries. The transitions between the consecutive visemes can be smooth, thus posing challenges for pinpointing their exact starting or ending points. Developing an accurate prominent viseme segmentation approach is a critical step for extracting phonetic-invariant components.

Unclear User Biometrics Related to Visemes. The relationships between visemes and speech-induced facial muscle vibrations are not clear. To realize text-independent authentication with low enrollment costs, it is essential to extract representative viseme-associated features that remain consistent across different speech contents from a limited set of commands. Moreover, these extracted features should carry distinct and unique biometrics, thus ensuring clear discrimination between legitimate users and attackers.

4.3 System Overview

To address the aforementioned challenges, we design a suite of techniques. The overview of SAFARI is illustrated in Figure 5.

Viseme Pattern Reconstruction. To mitigate the effects of motion artifacts on facial vibrations, we develop a generative diffusion model, which is designed to restore facial vibrations to their original state, unimpacted by body motions. The model training comprises two distinct sub-processes: forward diffusion and reverse diffusion. In the forward diffusion process, we simulate the motion artifacts

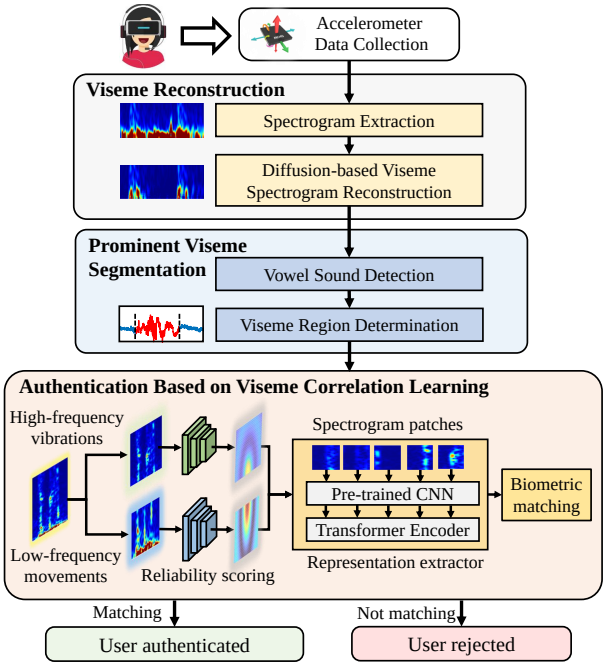


Figure 5: System overview of SAFARI.

by integrating them with clean facial vibration data. These artifacts are synthesized based on typical body movements encountered in AR/VR environments, such as head rotations and arm movements. Subsequently, the reverse diffusion process focuses on training the model to accurately reconstruct the clean vibration data. This involves gradually removing the synthesized motion artifacts to recover the original, undistorted facial vibration patterns.

Prominent Viseme Segmentation. We develop a two-step segmentation scheme to detect the temporal regions of prominent visemes that contain a complete mouth open-and-close cycle. In the first step, our method identifies vowel sounds by examining the high-frequency facial muscle vibrations (e.g., $\geq$ 100Hz), which is exclusive to vowel sound production. This ensures that the segmented region contains viseme rather than arbitrary facial expressions (e.g., smiling, yawning). In the second step, the scheme determines the starting and ending positions of the viseme by analyzing low-frequency facial movements below 100Hz, which effectively captures the opening and closing gestures of the mouth.

Speech authentication Based on Correlation Learning. Leveraging prominent viseme segments, SAFARI employs a transformer-based correlation learning strategy to extract phonetic-invariant facial biometrics. The learning strategy involves contrasting low-frequency facial movements with high-frequency muscle vibrations, providing a dual-perspective analysis of each viseme. To exploit the correlation between two types of facial vibrations, we have developed a reliability scoring model. This model dynamically assigns weights to more stable components of the spectrogram, thereby ensuring that the biometric representations remain consistent across varying speech contents. A transformer-based model then processes these weighted spectrograms to extract phonetic-invariant biometric representations. For each enrolled user, SAFARI constructs a binary classifier that distinguishes the user's biometric

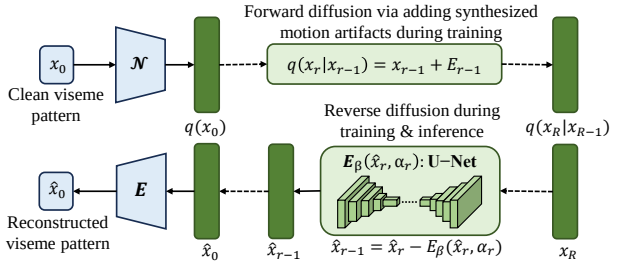
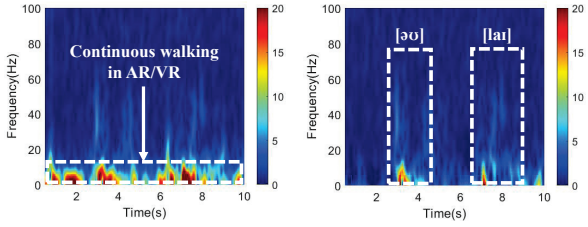


Figure 6: Illustration of the forward diffusion and reverse diffusion process for viseme pattern reconstruction.



(a) Visemes under continuous walking (b) Reconstructed visemes

Figure 7: Viseme pattern reconstruction for open ([əʊ]) and library ([laɪ]) under continuous walking with the headset.

signature from others, including specific anchor users. When voice commands normally contain multiple prominent visemes, SAFARI aggregates the authentication results from all prominent visemes using a max-vote strategy. This aggregation further enhances the robustness of the speech authentication. Additionally, SAFARI can accommodate multi-user enrollment, such as family settings where a single headset is shared. This is achieved by creating separate user profiles, each represented by a unique binary classifier, allowing individualized authentication for each legitimate user.

5 Viseme Reconstruction via Generative Deep Learning Model

In practical AR/VR scenarios, users may interact with the headsets through various types of body movements (e.g., rotating heads, manipulating controllers) while issuing voice commands. These motion artifacts can be mixed with viseme-associated facial vibrations, distorting the viseme patterns. To ensure reliable viseme biometric extraction, we design a reconstruction scheme that reverses the motion sensor data to a state that only contains facial vibrations.

Diffusion-based Viseme Reconstruction. Diffusion-based generative models have achieved state-of-the-art performance compared to other deep-learning-based generative techniques, such as Generative Adversarial Network (GAN) and Variational Autoencoders (VAE), in text [52], image [24], and audio generation [40] tasks. Motivated by its effectiveness, we develop a diffusion-based generative model to reconstruct visemes from body motions. The viseme reconstruction can be formulated as a parameterized stochastic process with variational noises for training and denoised viseme samples as outputs. We separate this process into two subprocesses, forward diffusion and reverse diffusion as illustrated in Figure 6. Specifically, we model the forward diffusion $q(x_R|x_0)$ as a Markov chain that gradually adds randomly-generated motion artifact E_r to the clean viseme x_0, which can be formulated as:

$$q(x_R|x_0) = \prod_{r=0}^{R} q(x_r|x_{r-1}), \tag{1}$$

where $q(x_r|x_{r-1}) = x_{r-1} + E_{r-1}$ denotes the diffusion function that generates the noise viseme x_r by mixing motion artifact E_{r-1} and noisy viseme x_{r-1} generated at step $r-1$. $E_r = \mathcal{N}(E_{r-1}, h, \tau, \alpha_r)$ refers using a pre-defined sampling function $\mathcal{N}$ to generate the motion artifact E_r. h refers to the motion artifact with length L that is randomly selected from a pre-collected motion dataset H and the motion artifact segment E_r is sampled at the starting time of $\tau \in [0, L-l]$ from h. Then the clean viseme x_0 can be reconstructed by gradually removing the body motion artifact E_r from the noisy viseme x_r via a reverse procedure, which can be described as:

$$x_{r-1} = x_r - E_\beta(x_r, \alpha_r), r \in [0, ..., R], \tag{2}$$

where $E_\beta(\cdot,\cdot)$ denotes the motion artifact prediction model that extracts body motion artifact E_r from the noisy viseme x_r and its magnitude ratio α_r. The model for body motion artifact prediction can be optimized through minimizing the following object:

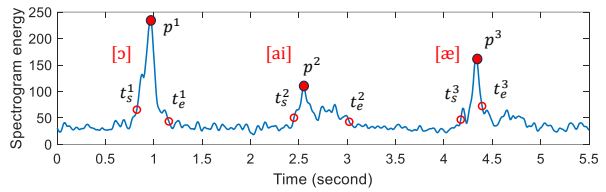
$$\arg\min_{\beta} \sum_{i=1}^{N} \sum_{r=1}^{R} \left\| E_\beta(x_{i,r}, \alpha_r) - E_{i,r} \right\|, \tag{3}$$

where $x_{i,r}$ and $E_{i,r}$ represent the i^{th} noisy viseme sample from step r and its corresponding motion artifact from the training set $D = \{(x_{i,r}, E_{i,r}), i = 1...., N\}$. N refers to the total number of samples in the training set D. β denotes the trainable parameters of the motion artifact prediction model and $E_\beta(\cdot,\cdot)$ can be utilized to expose the body motion artifact in the noisy viseme pattern x_r after optimization. The reconstructed viseme can be then obtained by subtracting the estimated motion artifact E_r from the noisy viseme data x_r. To realize the diffusion-based viseme pattern reconstruction, we build a model based on the structure of U-Net [39], which employs a multi-layer perception architecture to embed the features associated with the noise magnitude and a convolutional layer to encode the viseme pattern. Then these embedded features are fed into a structure that contains five down-sampling and up-sampling layers, and finally outputs the predicted body motion artifacts.

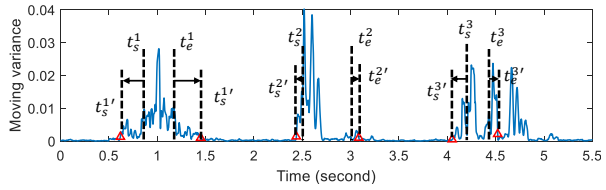
Training Procedure. To build the diffusion model, a set of clean viseme samples are collected. We also collect the motion sensor readings associated with a set of common body movements in AR/VR scenarios, including head rotation, walking around, swing controller, squatting, and turning around to create the body motion dataset H. E_r is generated using the random sampling fuction $\mathcal{N}$ on the motion artifact dataset H, and the training set D is constructed by mixing $E_{i,r}$ with corresponding clean viseme $x_{i,0}$. We set the length of motion artifact segment l to 3 seconds and the ratio α_r as $\alpha_r = 0.01 \times r$. The step number R of the diffusion and reverse procedure is fixed as 100. An example viseme spectrogram reconstructed from continuous walking is illustrated in Figure 7, where the motion artifacts caused by continuous walking can be effectively removed, and the viseme patterns are restored.

6 Prominent Viseme Segmentation

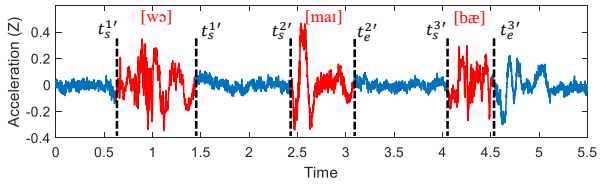
SAFARI first locates the prominent visemes that are embedded with strong and consistent visemes within facial vibrations. In particular, we define prominent visemes as either the first vowel viseme (V_O in "open") or the first viseme combination that involves a consonant



(a) Detecting the vowel sounds from facial muscle vibrations



(b) Examining facial muscle movements to detect mouth opening/closing



(c) Segmenting accelerometer data of prominent visemes

Figure 8: Illustration of our prominent viseme segmentation scheme that first detects the regions of vowel sounds (i.e., t_s^k and t_e^k) and then searches mouth opening and closing (i.e., $t_s^{k\prime}$ and $t_e^{k\prime}$) of the visemes from "What's my battery".

and a vowel, as they contain a complete open-and-close cycle of human mouth. While pronouncing visemes, users typically bring their mouth close, then open the mouth, move the lips forward (especially during speaking the vowel sounds), and finally close their mouth again [8]. These distinctive cycles of mouth movements make these visemes distinguishable to be separated from motion sensor readings. To precisely detect and separate prominent visemes, we design a two-step viseme segmentation scheme, including *vowel sound detection* and *viseme region determination*. Firstly, our scheme detects the pronunciation of vowel sounds by examining the existence of high-frequency facial muscle vibrations (e.g., $\geq$ 100Hz), which are distinguishable from the viseme spectrograms. Secondly, based on the identified region of the vowel sounds, our scheme searches for the starting and ending points of mouth closing by analyzing low-frequency facial muscle movements (e.g., $\leq$ 100Hz). With this approach, we can effectively prevent the incorrect detection of speech-irrelevant facial expressions in AR/VR scenarios (e.g., smiling, yawning) as prominent visemes.

Vowel Sound Detection. To detect the vowel sounds, our scheme first applies element-wise summation on the spectrograms from the z-axis readings of accelerometer. Based on the summed spectrograms, we accumulate the spectrogram energy across frequencies above 100Hz to measure the energy distribution of facial muscle vibrations. An example is illustrated in Figure 8(a), which demonstrates that the energy peak always locates within the vowel sound region. To precisely detect vowel sounds, a peak selection algorithm is developed to find the points (e.g., p^k) with prominent energy compared with other adjacent peaks [51]. Our scheme then searches for a pair of closest points where the mean and variance of energy show abrupt changes (e.g., t_s^1 and t_e^1) [29] larger than a pre-defined threshold. Examples of detected peaks and change points of the vowel sounds [ɔ], [ai], and [æ] are shown in Figure 8(a).

Viseme Region Determination. Given that users open their mouths to enlarge the vocal tract in preparation for sound productions (e.g., [əʊ], [ɔ], [æ]), the low-frequency facial muscle movements are produced prior to the associated facial muscle vibrations. Moreover, the facial movements will maintain for a duration after users complete the sound production. Therefore, the regions of prominent visemes should be larger and completely cover the detected vowel sounds. To determine the regions of prominent visemes, we apply moving variance upon motion sensor readings, where a large moving variance indicates the existence of a significant facial movement. An example of using moving variance for mouth opening and closing detection is shown in Figure 8(b). Similar with vowel sound detection, we detect the points with abrupt changes of the spectrogram energy to determine the starting time t_s^k and ending time t_e^k of the viseme corresponding to the k^{th} detected vowel. We then search for the second change point $t_s^{k\prime}$ closest to t_s^k with $t_s^{k\prime} < t_s^k$ and $t_e^{k\prime}$ closest to t_e^k with $t_e^{k\prime} > t_e^k$. The examples of detected prominent visemes and associated accelerometer readings (e.g., z-axis) are elaborated in Figure 8(b) and Figure 8(c).

7 Authentication Framework Based on Viseme Correlation Learning

7.1 Model Overview

To perform reliable user authentication, we design a correlation learning framework to derive viseme-associated biometrics. The idea is to contrast facial movements and vibrations within each prominent viseme, which highlights the phonetic-invariant components (e.g., face shape, bone and muscle properties) shared among these two types of dynamics. It thus reliably links the viseme patterns with users' unique biometrics. The framework takes facial muscle movement spectrogram x_m and vibration spectrogram x_v as inputs. The two spectrograms are fed into two scoring models, $G_m(\cdot)$ and $G_v(\cdot)$, which dynamically adjust their weights to highlight the phonetic-invariant parts of spectrograms. The two weighted spectrograms, f_m and f_v, are then concatenated, which is referred as $f = [f_m, f_v]$, and fed into a transformer-based encoder $D(\cdot)$ to extract biometric representations. For user authentication, we build a binary classifer for each legitimate user (e.g., $U^{(j)}(\cdot)$ for user j), which determines whether the representations belong to the legitimate user (e.g., user j) or not.

7.2 Reliability Scoring Model for Facial Muscle Movement and Vibration

To extract the emphasized spectrogram of prominent visemes, we develop two reliability score models based on Convolutional Neural Networks (CNNs). In particular, the scoring models take facial movement spectrogram $x_m \in \mathbb{R}^{C\times T\times F_m}$ and facial vibration spectrogram $x_v \in \mathbb{R}^{C\times T\times F_v}$ as inputs and generate two sets of reliability scores, which are denoted as $M_m \in \mathbb{R}^{C\times T\times F_m}$ and $M_v \in \mathbb{R}^{C\times T\times F_v}$. C is referred as the input channels associated with the 3-axis motion sensor readings. T and F_m/F_v denote the numbers of points in the temporal (e.g., ~0.2s) and frequency dimensions (e.g., $\leq$ 100Hz and

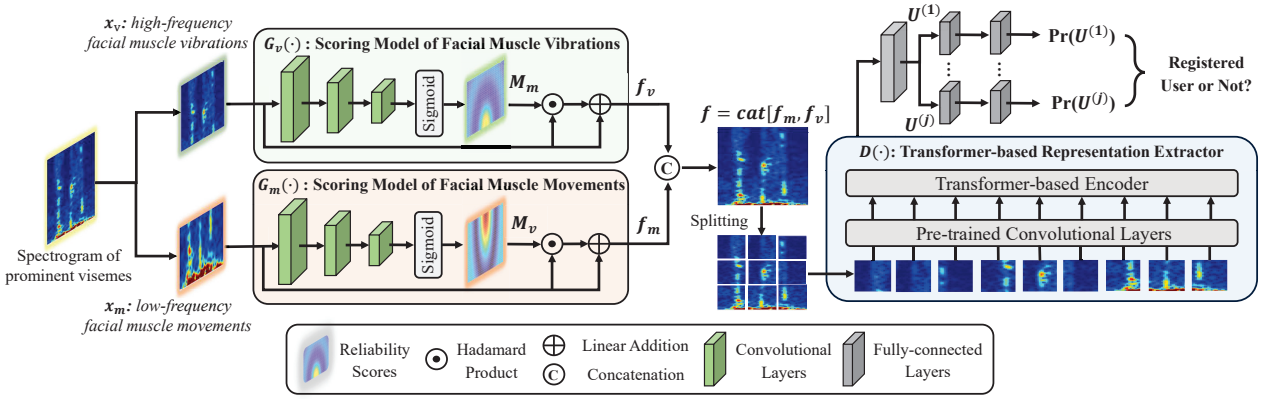


Figure 9: Model overview of the viseme correlation learning framework.

≥ 100Hz). High reliability scores highlight the phonetic-consistent part of prominent visemes that may facilitate user biometric derivation and differentiation, which can be described as:

$$f_m = x_m + M_m \odot x_m, f_v = x_v + M_v \odot x_v, \tag{4}$$

where f_m and f_v denotes the emphasized spectrograms of facial movements and vibrations, which are then concatenated along the frequency dimension to generate the combined emphasized viseme $f = [f_m, f_v] \in \mathbb{R}^{C\times T\times F}$, with $F = F_m + F_v$. Through the reliability scoring model, we highlight the phonetic-invariant features from prominent visemes, which are further utilized to extract facial representations for effective user authentication.

7.3 Facial Representation Extraction Based on Spectrogram Transformer

Transformer-based deep learning models, such as Audio Spectrogram Transformer (AST) [20], have outperformed traditional models (e.g., ResNet [23] and LSTM [25]) on speech and speaker recognition. Particularly, the multi-head self-attention mechanism of the transformer enables itself to focus on different segments of input sequences, which also facilitates capturing spatial and temporal features from human visemes. Inspired by the design of AST, we develop a transformer-based viseme representation extractor to derive viseme-associated facial biometrics from individual users. Specifically, the transformer-based representation extractor takes the emphasized viseme spectrogram f as input and splits it into several spectrogram patches. Multiple pre-trained convolutional layers are employed to embed these separated patches for a transformer-based encoder, which includes 7 heads with self-attention layers to derive users' distinctive facial representations.

7.4 Training Procedure for Representation Extractor and User Authentication Model

Representation Learning. We optimize the trainable parameters of the reliability scoring models $G_m(\cdot)$ and $G_v(\cdot)$ and the transformer-based representation extractor $D(\cdot)$ to derive facial representations. To validate that the extracted embeddings are effective in differentiating users, we build a user identifier $P(\cdot)$ with two fully-connected layers. During the training phase, we apply cross-entropy loss to optimize the scoring models $G_m(\cdot)$ and $G_v(\cdot)$, the representation extractor $D(\cdot)$, and the user classifier $P(\cdot)$. The loss function L_R used for optimization is formulated as:

$$L_R = \frac{1}{N}\sum_{i=1}^{N} y_i \cdot log\Big(P\big(D([G_m(x_{i,m}), G_v(x_{i,v})])\big)\Big), \tag{5}$$

where $y_i \in [1, ..., M]$ denotes the label of facial muscle vibration spectrogram $x_{i,m}$ and facial muscle vibration spectrogram $x_{i,v}$. N refers to the total number of viseme samples involved during training. Note that the user identifier $P(\cdot)$ is only involved in the training process to facilitate the extraction of representative user embeddings, and it will not be employed in the user authentication phase.

User Authentication Model. During the training phase of the user authentication model, we fix the parameters of the reliability scoring models $G_m(\cdot)$ and $G_v(\cdot)$ and the transformer-based extractor $D(\cdot)$. For each registered user j, we build a binary classifier $U^{(j)}(\cdot)$ with two fully-connected layers to determine whether the input representation d is from user j or not. The learnable parameters of the classifier $U^{(j)}(\cdot)$ are updated via the loss function $L_U^{(j)}$ corresponding to user j, which can be described as:

$$L_U^{(j)} = \frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(U^{(j)}(d_i)), \tag{6}$$

where y_i represents the user label of the viseme representation d_i extracted by the transformer-based extractor $D(\cdot)$. To improve accuracy, we perform authentication based on individual prominent visemes of speech and fuse their predictions with max-vote. This aggregation can further improve the robustness of SAFARI.

8 User Authentication Performance

8.1 Experimental Methodologies

AR/VR Headsets. We evaluate the user authentication performance of SAFARI on two widely-used standalone AR/VR headsets: Meta Quest and Meta Quest 2. Both of them are equipped with independent motion sensor modules for continuous motion tracking. For Meta Quest, it uses a motion sensor board with the series number of 330-00193-03, which is originally developed by Meta. Meta Quest 2 is equipped with a motion sensor board with the series 330-00829-04. Both Meta Quest and Meta Quest 2 operate on an Android-based system. Under this platform, we utilize Meta Mobile SDK [36] to develop an application and collect the accelerometer readings from the headsets. During the experiments, we set the sampling rates of motion sensors as 1000Hz on both devices.

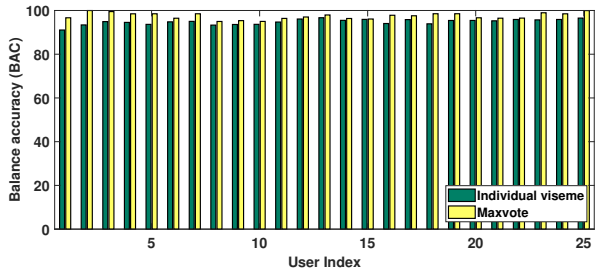


Figure 10: Overall authentication performance of SAFARI.

Voice Commands. We select a set of voice commands and collect the visemes from AR/VR accelerometers while users pronounce these commands. In particular, we choose 30 voice commands that are commonly used in AR/VR scenarios (e.g., "Open Beat Saber"), which have encompassed all 11 different visemes. The accelerometer readings of visemes are then divided into several prominent viseme segments lasting for 1.6 seconds. The average word count of the commands is 3.1, with the shortest and longest commands including 1 word and 7 words, respectively. Our selected 30 representative voice commands are illustrated in Table 4 in Appendix A.

Participants & Data Collection. We collect the visemes from a total of 35 participants with ages ranging from 18 to 37, including native and non-native English speakers. For Meta Quest, we involve 25 participants, with 22 males and 3 females aged from 20 to 33. For Meta Quest 2, we recruit 10 participants, including 7 males and 3 females with ages ranging from 17 to 37. Each participant is asked to wear the headset and pronounce the 30 voice commands for 10 repeats each. During data collection, we place a sound level meter 30cm away from the participants' mouths to measure the sound pressure levels (SPLs) during command pronunciation. In the experiments, we constrain the SPLs within 65dB to 75dB and set no specific restrictions on users' movements. In total, we collect 95, 570 and 38, 228 viseme samples from the accelerometers of Meta Quest and Meta Quest 2. The data collection procedures have been approved by our university's Institutional Review Boards (IRB).

Evaluation Metrics. We employ the following evaluation metrics to evaluate the user authentication performance of SAFARI. *(1) True Positive Rate (TPR):* The percentage of legitimate users who are correctly verified as such. *(2) True Negative Rate (TNR):* The percentage of unauthorized users who are correctly verified as such. *(3) Balanced Accuracy (BAC):* An evaluation metric that combines TPR and TNR with an equal weight. It is also important to note that the Receiver Operating Characteristic (ROC) curve is not utilized to evaluate the performance of SAFARI. The reason is that the classification boundary is determined by the deep-learning-based user authentication model in our design.

8.2 Overall User Authentication Performance

Setup: We utilize the viseme dataset collected from Meta Quest to evaluate the overall user authentication performance of SAFARI. Specifically, we take turns selecting each of the 25 participants as legitimate user and the remainings as unauthorized users. During the training phase, we randomly select 20 different commands and use the extracted prominent visemes as the training set. The visemes from the remaining 10 commands are employed as the testing set. Note that the voice commands of the testing set are entirely different from the training set to evaluate the performance of text-independent user authentication. The BAC of each user is calculated to evaluate the user authentication performance.

Table 2: The authentication performance of SAFARI with and without Viseme Pattern Reconstruction (VPR).

	Head Rotation		Walking Around	
	Without VPR	With VPR	Without VPR	With VPR
TPR	15.04%	96.44%	20.28%	95.55%
TNR	94.73%	99.87%	96.76%	99.22%
BAC	54.88%	98.15%	58.52%	97.38%

Results: The BACs of SAFARI corresponding to different users are illustrated in Figure 10. The results show that SAFARI achieves the BAC of more than 91.28% for most participants while only one individual viseme is utilized for authentication. After employing max vote across multiple visemes within one command, the BAC of SAFARI can be further improved, with more than 95.71% for most participants. High authentication accuracy demonstrates that SAFARI realizes effectively user authenticate by establishing correlations between visemes and users. The results also validate the effectiveness of SAFARI's text-independent authentication given the entirely different commands in training and testing sets.

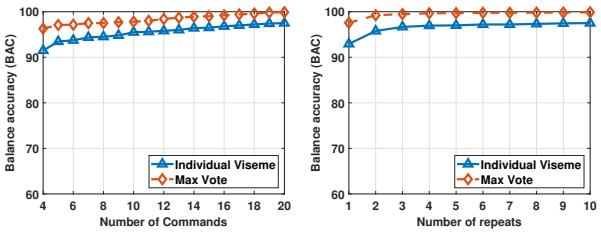
8.3 Impacts of Motion Artifact

Setup: To evaluate SAFARI's performance against motions, we conduct experiments on Meta Quest by recruiting 10 users and instructing them to pronounce the aforementioned commands while engaging in two pre-defined body motions. *(1) Head rotation.* The participants randomly rotate their heads in different directions. *(2) Walking around.* The participants randomly walk around within the boundary of the AR/VR virtual environment. In total, two viseme datasets corresponding to these two body motions are collected from the AR/VR accelerometer. To construct the authentication model, we utilize the extracted prominent visemes from the first 20 commands as the training set and the remaining 10 commands as the testing set. The average TPR, TNR, and BAC of all 10 participants are summarized to evaluate the user authentication performance.

Results: The average TPR, TNR, and BAC of SAFARI without and with motion artifact removal based on viseme reconstruction are illustrated in Table 2. The results show that the user authentication performance of SAFARI is compromised by the body motions in AR/VR scenarios, with TPR, TNR and BAC below 15.04%, 94.73% and 54.88%, respectively. After utilizing viseme reconstruction for mitigating motion artifacts, the TPR and TNR under head rotation and walking around is significantly improved, which achieve more than 96.44% and 99.87%. The BAC of SAFARI is also significantly enhanced, with more than 98.15% and 97.38% under head rotation and walking around scenarios. In summary, the substantial improvements in user authentication performance demonstrate the effectiveness of our designed motion artifact removal scheme via viseme reconstruction and SAFARI's robustness against body motions in practical AR/VR usage scenarios.

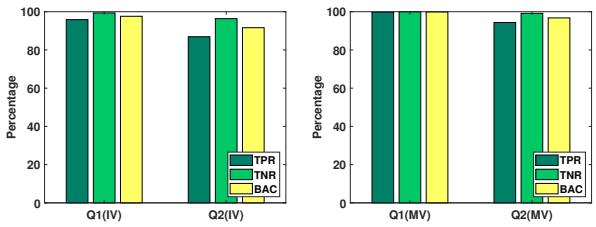
8.4 Impacts of Training Size

Setup: Although increasing the number of commands or repeats could facilitate authentication accuracy, it brings additional costs



(a) BAC with different command numbers (b) BAC with different repeats

Figure 11: Impact of the number of training voice commands and repeats for each command.



(a) Performance with individual viseme (b) Performance with Max-vote

Figure 12: Authentication results of Meta Quest (Q1) and Meta Quest 2 (Q2) with Individual Viseme (IV) and Max Vote (MV).

for user enrollment. An efficient and effective authentication system should maintain high accuracy while minimizing the number of commands required for enrollment. Based on this, we measure the BAC associated with different numbers of commands or repeats to explore the authentication performance with different training sizes. Specifically, we involve 10 participants and collect their visemes on Meta Quest. During training, we build the authentication model with 4∼20 different commands or 1∼10 repeats from the voice command datasets. In the testing phase, the visemes extracted from the remaining 10 voice commands are employed for evaluation.

Results: We summarize the BAC with varying numbers of voice commands for training in Figure 11(a). In particular, SAFARI attains a BAC of more than 96.29% while only 4 different commands are collected for user enrollment, which demonstrates that SAFARI accurately authenticates users with low data collection and training costs. The BAC of SAFARI with different repetitions for each command during model construction is illustrated in Figure 11(b). The results show that SAFARI achieves a BAC of more than 97.56% while only one repetition of each voice command is collected from the user. High authentication accuracy validates that SAFARI successfully authenticates users with limited numbers of voice samples for building user profiles. In summary, the high authentication accuracy with limited commands or repetitions for training demonstrates SAFARI's low costs on model construction and user enrollment.

8.5 Impacts of Headset Models

Setup: To explore SAFARI's effectiveness on different AR/VR devices, we evaluate its authentication performance on Meta Quest and Meta Quest 2. Compared with Meta Quest, Meta Quest 2 is built with lightweight materials and a more advanced motion sensor board (details in Section 8.1). Specifically, we collect two viseme datasets from the same 10 users on both Meta Quest and Meta Quest 2. The visemes extracted from 20 different commands and the remaining 10 commands are employed for training and testing. We summarize the average TPR, TNR, and BAC to evaluate SAFARI's authentication performance on two different headset models.

Results: The TPR, TNR, and BAC of SAFARI on Meta Quest and Meta Quest 2 are shown in Figure 12. For Meta Quest, SAFARI achieves TPR, TNR, and BAC of more than 95.85%, 99.37%, and 97.61% with one individual viseme. After involving max-vote for authentication, the TPR, TNR, and BAC are improved to 99.78%, 99.98%, and 99.88%. For Meta Quest 2, SAFARI reaches TPR, TNR, and BAC of more than 86.90%, 96.37%, and 91.64% if only one individual viseme is used. After applying max-vote, the TPR, TNR, and BAC can achieve more than 94.36%, 99.16%, and 96.76%, respectively. An explanation for Meta Quest's better performance could be attributed to its heavier head-mounted display. This characteristic makes the user's face in closer contact with Meta Quest, thus facilitating effective viseme capturing. Nevertheless, consistently high accuracy indicates that SAFARI can accurately authenticate users while deployed on different devices.

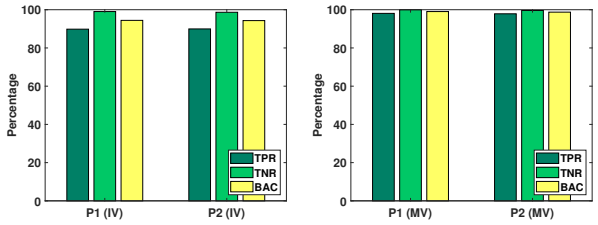
8.6 Impacts of Different Headset Placements on Human Face

Setup: To explore SAFARI's robustness against different headset placements on human faces, we conduct experiments on Meta Quest with 10 participants. We instruct the participants to wear the headset and maintain a consistent position in the virtual environment while collecting visemes associated with 30 different commands, which is referred to as Placement 1 (P1). The participants then take off the headset and wear it after a while to collect another group of visemes associated with the same 30 commands, which is defined as Placement 2 (P2). We summarize the average TPR, TNR, and BAC with (1) visemes of 20 commands in P1 for training and the other 10 commands in P2 for testing, and (2) visemes of 20 commands in P2 for training and the other 10 commands in P1 for testing.

Results: The TPR, TNR, and BAC of SAFARI under different headset placements on humans' faces are illustrated in Figure 13. In particular, SAFARI achieves TPR, TNR, and BAC of more than 89.94%, 98.66%, and 94.30% with P2 for training and P1 for testing with one individual viseme for authentication. After involving max-vote on multiple visemes, the TPR, TNR, and BAC are improved to 97.86%, 99.60%, and 98.73%. With P1 for training and P2 for testing, SAFARI realizes TPR, TNR, and BAC of more than 89.82%, 99.03%, and 94.43% with one individual viseme. After applying max vote, the TPR, TNR, and BAC achieve more than 98.08%, 99.98%, and 99.03%. Consistently high accuracy indicates that SAFARI has realized effective user authentication under different headset placements on user faces during practical AR/VR usage.

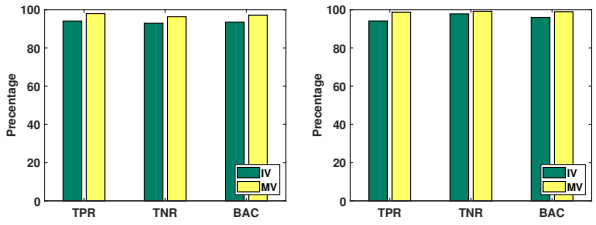
8.7 Evaluation of Computational Delay

Setup: While realizing user authentication schemes in practical scenarios, short inference time is crucial for achieving real-time user authentication and better user experience. To validate that SAFARI can be deployed to authenticate users in practical AR/VR scenarios, we evaluate the average computational time of different modules in SAFARI. In particular, we conduct experiments using an NVIDIA 4090 GPU and Intel-13900K CPU with a batch size of



(a) Performance with individual viseme (b) Performance with Max-vote

Figure 13: Authentication results of two headset placements (P1 and P2) using Individual Viseme (IV) and Max Vote (MV).



(a) Defending blind attack (b) Defending observe-and-mimic attack

Figure 14: Authentication performance on defending blind attack and observe-and-mimic attack.

64 for 1000 visemes to measure the average computational time of viseme pattern reconstruction, reliability scoring, transformer-based representation extractor, and user verification.

Results: We summarize the average computational time corresponding to different modules of SAFARI. Compared with other modules, the reliability scoring model has the longest computational time, with the average of 115ms. For viseme pattern construction, transformer-based representation extractor, and user verification, the average computational time are 56ms, 14ms, and 75ms, which indicates that SAFARI takes approximate 260ms in average to process a single viseme input and authenticate user. Short computational time cost validates that SAFARI can be deployed in practical AR/VR scenarios for realizing real-time user authentication.

9 Robustness to Spoofing Attack

9.1 Robustness to Blind Attack

Setup: During the blind attack, adversaries attempt to bypass SAFARI using their own visemes without any prior knowledge on the legitimate users' visemes. To simulate this attack, we collect visemes associated with 30 different commands from 10 participants on Meta Quest. Each user takes turns serving as the legitimate user, and the training and testing set includes the visemes of 20 commands and the remaining 10 commands. We then randomly select the other 10 participants as adversaries, and collect the visemes of the remaining 10 commands in the testing set. For evaluation, we combine the visemes from both legitimate users and adversaries and then summarize the average TNR, TPR, and BAC.

Results: The authentication performance of SAFARI against blind attack is illustrated in Figure 14(a). In particular, SAFARI achieves TPR and TNR of more than 94.04% and 92.94% with one individual viseme for user authentication. After employing max vote, SAFARI can realize TPR and TNR of more than 97.95% and 96.33%. For BACs, SAFARI remains high accuracy with more than 93.49% and 97.14% using one individual viseme and max vote for authentication. The results show that SAFARI successfully resists blind attack while maintaining effective user authentication, which can be attributed to the robust extraction of distinctive viseme representations related to the shape and structure of human faces.

9.2 Robustness to Vibration Replay Attack

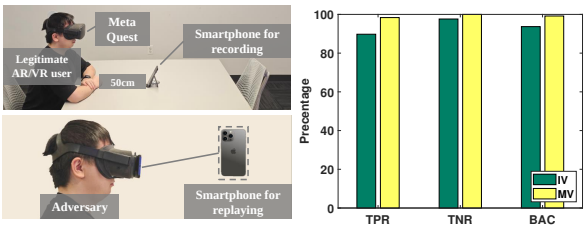
Setup: In the vibration replay attack, the adversaries attempt to bypass SAFARI by replaying commands with a loudspeaker, which is placed in direct contact with the headset to facilitate vibration capturing of the accelerometer. To evaluate SAFARI's robustness against vibration replay attack, we utilize an iPhone 13 smartphone, which is placed 50cm away from legitimate users to record their command pronunciation. During the attack, the adversary places the smartphone in direct contact with the headset and replays the recorded commands to generate sound vibrations. The experimental setup is illustrated in Figure 15(a). To build the authentication model, we collect visemes from 10 participants on Meta Quest, with each taking turns as the legitimate user. The collected visemes corresponding to 20 different commands are served as the training set. For the testing data, we collect the viseme samples (i.e., legitimate users) and the audio recordings (i.e., adversaries) of the remaining 10 commands. The average TNR, TPR, and BAC are measured to evaluate SAFARI's robustness against vibration replay attack.

Results: The authentication performance of SAFARI under vibration replay attack is summarized in Figure 15(b). While using one individual viseme for authentication, SAFARI achieves TPR, TNR, and BAC of more than 89.75%, 97.62%, and 93.68%. After incorporating max-vote, the TPR, TNR, and BAC against vibration replay attack are further improved, with more than 98.35%, 100.00%, and 99.17%. The results demonstrate that SAFARI maintain effective and robust against vibration replay attack, which can be attributed to the difference between the air-propagated sound vibrations and visemes as described in Section 2.

9.3 Robustness to Observe-and-mimic Attack

Setup: During the observe-and-mimic attack, the adversaries aim to bypass SAFARI by observing the speech articulation and mimicking the facial muscle vibrations of the legitimate users. To simulate observe-and-mimic attack, the viseme data associated with 20 different commands is first collected from 10 participants using Meta Quest for training the authentication model. We then randomly select the other 10 participants as the adversaries and instruct them to observe the facial deformations of the 10 legitimate users during voice command pronunciation. During the testing phase, we collect the viseme samples corresponding to the remaining 10 voice commands from both the adversaries with the mimicked facial deformations and the legitimate users, and combine them as the set for testing. The average TPR, TNR, and BAC are measured to evaluate SAFARI's robustness against observe-and-mimic attack.

Results: The average TPR, TNR, and BAC against observe-and-mimic attack are shown in Figure 14(b). Specifically, SAFARI achieves TPR, TNR, and BAC of more than 94.08%, 97.81%, and



(a) Experimental setup (b) Defending vibration replay attack

Figure 15: Experimental setup and authentication performance on defending vibration replay attack.

95.95% using only one viseme for user authentication. While applying max vote for authenticating users, SAFARI achieves TPR, TNR, and BAC with more than 98.69%, 99.14%, and 98.92%. To ensure that the adversary accurately mimics the users' facial movements, we also conduct experiments on attacks where the adversary can learn users' facial movements through a pre-recorded video of users issuing voice commands. It can be considered the best-case scenario as the adversary can repeatedly watch the video and practice before launching the attack. Moreover, we ask a third person to watch the adversary to confirm the proper mimicry of facial expressions. Specifically, 3 adversaries are asked to watch the videos of a user issuing 30 voice commands and try to replicate the user's facial expression. Each adversary repeats this process multiple times to ensure a high degree of mimicry. Even under such actions, the adversaries are not able to bypass SAFARI with a TNR of more than 98%. The results demonstrate that SAFARI exhibits strong resilience against observe-and-mimic attack and our designed viseme-associated biometrics are validated to be challenging for attackers to replicate through observing facial deformations.

10 Related Work

User Authentication on AR/VR. Traditional password-based authentication methods (e.g., passwords [19], PINs [53], and lock patterns [38]) originally designed for computers and smartphones have been adapted for AR/VR platforms. However, these methods are ill-suited to AR/VR's novel input interfaces via gestures. Unlike physical keyboards and touchscreens, AR/VR users are required to enter credentials using controllers or hand gestures, a process that can be both time-consuming and inconvenient. While two-factor authentication systems, such as those involving barcodes or smartphone messages [32], enhance security, they can disrupt the immersive AR/VR experience. Recent research has explored behavior biometrics of gestures, including gestures [13, 26, 37, 49, 55]. Other approaches have investigated user authentication via head-conducted vibrations [31] and sound signals [48]. These methods, requiring additional challenge signals like vibrations or sound chirps to extract biometrics, which can be intrusive and often impractical without hardware modifications. Compared with these existing AR/VR authentication, SAFARI is transparent as it does not requires additional actions or active involvement of users. It is also compatible with mainstream AR/VR headsets.

Voice Authentication and Liveness Detection. Prior research has explored voice authentication using acoustic features, such as Filter Banks [44] and Mel-Frequency Cepstral Coefficients (MFCC) [46, 56, 60]. While these methods have shown potential, they typically require extensive training data to develop a robust voiceprint for each user and are susceptible to spoofing attacks. In contrast, SAFARI accomplishes text-independent authentication with only 15 to 20 short voice commands (each under 5 seconds), enhancing efficiency and user-friendliness. Furthermore, these voice authentication methods are often vulnerable to spoofing attacks, including speech synthesis and replay attacks. Differently, SAFARI leverages facial vibrations that are confined to the human body and offers resilience against biometric leakage through the audio channel and subsequent attacks. To counteract voice spoofing, liveness detection techniques can be integrated with voice authentication to improve security. These techniques aim to distinguish between live human speech and machine-generated sounds by exploiting features inherent to either the human vocal tract structure [10], the magnetic field from loudspeakers [14], the time difference-of-arrival (TDoA) from two microphones [59], sound-field characteristics [54], or vibrations induced by the human body or speech [18, 42]. However, these liveness detection methods often necessitate the integration of specialized microphones or additional sensors, which introduce additional overhead for AR/VR users. SAFARI, on the other hand, leverages the built-in motion sensors readily available in most commercial AR/VR headsets, providing a more seamless and integrated solution. A comparative comparison of SAFARI with existing voice authentication and defense systems is presented in Table 3.

Table 3: SAFARI in comparison with existing authentication schemes on feature type, text independence, built-in device, spoofing resilience, and gesture requirement.

Authentication System	Feature Extraction	Text-Indep.	Built-in Device	Spoofing Resilience	Gesture Free
Variani et al. [46]	MFCC	✓	✓	×	✓
Snyder et al. [44]	Filter banks	✓	✓	×	✓
VoiceLive [59]	Phoneme	×	✓	×	×
CaField [54]	Sound field	✓	×	✓	✓
Blue et al. [10]	Vocal tract	✓	×	✓	✓
VoiceGesture [58]	Acoustic features	×	×	✓	×
WiVo [33]		×	×	✓	×
VAuth [18]	Speech vibrations	✓	×	✓	×
WearID [42]		✓	×	✓	×
SAFARI (ours)	**Viseme**	✓	✓	✓	✓

Speech Sensing Based on Motion Sensors. Existing studies have been utilizing MEMS motion sensors for speech sensing on smartphones [5, 6, 34, 57]. For example, Accelword [57] designs a benign application to sense speech content using the smartphone's accelerometer. AccelEve [7] and Speechless [5]investigate the potential for speech privacy leaks through accelerometers and gyroscopes in smartphones. A more recent study, Face-Mic [43], delved into the possibility of inferring sensitive user information, such as gender, identity, and speech content, through AR/VR motion sensor data. Particularly, Face-Mic captures the speech-related facial dynamics of headset wear and utilizes a deep learning model for the privacy attack. However, due to the reliance on speech patterns, the performance of Face-Mic is still susceptible to variations of speech content. In contrast, SAFARI takes a different approach by focusing on the extraction of phonetic-invariant biometrics from facial vibrations. It detects and segments prominent visemes from the facial vibrations to realize text-independent speech authentication.

11 Limitation

Limited Sample Size. To validate the effectiveness and robustness of SAFARI, we collect viseme samples from a group of 35 participants, including 29 males and 6 females. Based on these samples, we first explore the feasibility of leveraging viseme-associated biometrics to realize text-independent and spoofing-resistant user authentication for AR/VR voice interfaces. While the sample size of 35 participants could be limited, we believe that SAFARI can be generalized to authenticate more users. The viseme-associated biometrics capture the user's face shape, bone properties, and muscle characteristics, which are distinctive across large populations. To further demonstrate that SAFARI can realize general authentication on large groups of users, we plan to involve more participants and collect a more extensive set of viseme samples in our future work.

Risk of Biometric Information Leakage. Similar to existing biometrics (e.g., fingerprints, iris, faces), viseme-associated biometrics may contain sensitive information (e.g., the user's facial properties and behaviors). A potential solution to protect these biometrics from leakage while ensuring the authentication performance is to incorporate on-device machine learning techniques [15, 35] while constructing the authentication model of SAFARI. By employing on-device learning approaches, users can create their profiles using viseme-associated biometrics that are stored locally on AR/VR devices. SAFARI will also safeguard this local data from potential leakage, thus ensuring secure authentication. During the model construction stage of SAFARI, multi-party computation mechanisms [9, 16] can also be utilized to protect viseme-associated biometrics from privacy leakage. In this case, the users' facial representations are jointly generated by multiple cloud servers. The adversaries cannot derive users' biometrics by analyzing the data leakage from only one or several servers.

12 Discussion

Impacts of Environmental Noise and Low Voice Volumes. Traditional voice applications in AR/VR scenarios usually rely on built-in microphones to pick up human sound signals. The performance of these applications could be significantly downgraded under noisy environments or voice inputs with low volumes. Compared with traditional voice applications using microphones for sound capturing, SAFARI utilizes motion sensors to derive human visemes. Validated by previous works [5, 6], the motion sensors can pick up conductive vibrations (e.g., viseme-associated facial vibrations) and are insensitive to air-conducted sound vibrations. Therefore, SAFARI is inherently robust to airborne environmental noises. Additionally, SAFARI leverages speech-induced facial movements to realize user authentication, which does not directly rely on voice sound. With low-volume voice inputs, SAFARI will also maintain robust performance in authenticating users.

Robustness to Brute-force Attacks. To bypass SAFARI, adversaries can repeat the utterance until SAFARI fails to reject the adversaries' voice input. For instance, considering the SAFARI's performance in defending the blind attack in Section 9.1, SAFARI may accept another adversary as the legitimate user in 3 of 100 attempts given the TNR of 97%. In practical usage scenarios, SAFARI can also integrate system lockouts following consecutive unsuccessful attempts to defend against brute force attacks (e.g., repeating an utterance until a false acceptance occurs). For instance, the possibility of the adversary being continuously rejected by SAFARI for 3 attempts will be more than 91.2% and the system will be locked if the adversary still cannot bypass SAFARI after 3 attempts. With this design, SAFARI can successfully defend against brute-force attacks while maintaining effective user authentication.

Attacks With a Dummy Robot Head. To bypass SAFARI, a possible attack is to deceive the system by involving a dummy robot head, which mimics the facial patterns of the victims. Meanwhile, the adversary should employ a strategy to capture high-quality video recordings while the victims are speaking, which enables the robot to replicate corresponding facial deformations. However, crafting a dummy head with materials that exactly replicate the composition of facial tissue and head structure remains challenging. Furthermore, capturing and precisely reconstructing unique facial deformation poses another challenge for potential adversaries. Consequently, such an imaginary attack is demonstrated as highly intricate, which could be effectively thwarted with SAFARI by leveraging unique viseme representations of different users.

Attack by Eavesdropping Motion Sensors. Adversaries may potentially deploy another attack against SAFARI through malicious applications, which eavesdrop and record the facial muscle vibrations from AR/VR motion sensors for impersonation attacks. However, such attacks require social engineering skills from adversaries to fool users into installing the malicious applications, which could inadvertently reveal the malicious intent. Additionally, it is still challenging for adversaries to physically "replay" the visemes even if the motion sensor readings are available since this process requires accurate reconstruction of facial deformations and a well-controlled dummy robot head as we discuss previously.

13 Conclusion

In this paper, we present SAFARI, the first spoofing-resistant and text-independent speech authentication system for AR/VR headsets. SAFARI stands out by its ability to extract unique facial biometrics of users through sensing viseme-associated facial vibrations via the built-in accelerometer. Particularly, our system adeptly identifies and segments prominent visemes that contain significant facial deformations and rich biometric content for speech authentication. To mitigate the impacts of motion artifacts, we design a generative diffusion model. This model effectively reconstructs viseme patterns to their original state that are unaffected by body motions. Furthermore, We design a two-step scheme to segment the temporal regions containing prominent visemes. Based on the prominent viseme segments, a transformer-based correlation learning strategy is designed to contrast the facial muscle movements and vibrations to elicit phonetic-invariant facial biometrics for speech authentication. Extensive experiments show that SAFARI can authenticate users with over 96% true positive rates. Moreover, SAFARI can successfully defend against various spoofing attacks, including blind attacks, vibration replay attacks, and observe-and-mimic attacks.

Acknowledgment

This work was partially supported by the National Science Foundation Grants CNS2114220, CNS2120276, CNS2145389, CNS2201465, CNS2154507, CCF2211163, IIS2311596, IIS2311597, and OAC2139358.

References

[1] 2023. Apple Vision Pro. https://www.apple.com/apple-vision-pro/. (2023).
[2] 2023. Meta Quest 3. https://www.meta.com/quest/quest-3/. (2023).
[3] 2023. Microsoft HoloLens 2. https://www.microsoft.com/en-us/hololens/. (2023).
[4] Zakaria Aldeneh, Anushree Prasanna Kumar, Barry-John Theobald, Erik Marchi, Sachin Kajarekar, Devang Naik, and Ahmed Hussen Abdelaziz. 2021. On the role of visual cues in audiovisual speech enhancement. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8423–8427.
[5] S. A. Anand and N. Saxena. 2018. Speechless: Analyzing the Threat to Speech Privacy from Smartphone Motion Sensors. In *Proceedings of IEEE Symposium on Security and Privacy (SP)*. 1000–1017.
[6] S Abhishek Anand, Chen Wang, Jian Liu, Nitesh Saxena, and Yingying Chen. 2019. Spearphone: A speech privacy exploit via accelerometer-sensed reverberations from smartphone loudspeakers. *arXiv preprint arXiv:1907.05972* (2019).
[7] Zhongjie Ba, Tianhang Zheng, Xinyu Zhang, Zhan Qin, Baochun Li, Xue Liu, and Kui Ren. 2020. Learning-based practical smartphone eavesdropping with built-in accelerometer. In *Proceedings of the Network and Distributed Systems Security Symposium (NDSS)*. 23–26.
[8] Helen L Bear and Richard Harvey. 2017. Phoneme-to-viseme mappings: the good, the bad, and the ugly. *Speech Communication* 95 (2017), 40–67.
[9] Assaf Ben-David, Noam Nisan, and Benny Pinkas. 2008. FairplayMP: a system for secure multi-party computation. In *Proceedings of the 15th ACM Conference on Computer and Communications Security (CCS '08)*. Association for Computing Machinery, New York, NY, USA, 257–266. https://doi.org/10.1145/1455770.1455804
[10] Logan Blue, Kevin Warren, Hadi Abdullah, Cassidy Gibson, Luis Vargas, J. O'Dell, Kevin R. B. Butler, and Patrick Traynor. 2022. Who Are You (I Really Wanna Know)? Detecting Audio DeepFakes Through Vocal Tract Reconstruction. In *USENIX Security Symposium*. https://api.semanticscholar.org/CorpusID:249059207
[11] Luca Cappelletta and Naomi Harte. 2012. Phoneme-to-viseme mapping for visual speech recognition. In *International Conference on Pattern Recognition Applications and Methods*, Vol. 2. SCITEPRESS, 322–329.
[12] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. 2016. Hidden Voice Commands. In *USENIX Security Symposium*. 513–530.
[13] Jagmohan Chauhan, Hassan Jameel Asghar, Anirban Mahanti, and Mohamed Ali Kaafar. 2016. Gesture-based continuous authentication for wearable devices: The smart glasses use case. In *Applied Cryptography and Network Security: 14th International Conference, ACNS 2016, Guildford, UK, June 19-22, 2016. Proceedings 14*. Springer, 648–665.
[14] Si Chen, Kui Ren, Sixu Piao, Cong Wang, Qian Wang, Jian Weng, Lu Su, and Aziz Mohaisen. 2017. You can hear but you cannot steal: Defending against voice impersonation attacks on smartphones. In *Distributed Computing Systems (ICDCS), 2017 IEEE 37th International Conference on*. IEEE, 183–195.
[15] Sauptik Dhar, Junyao Guo, Jiayi (Jason) Liu, Samarth Tripathi, Unmesh Kurup, and Mohak Shah. 2021. A Survey of On-Device Machine Learning: An Algorithms and Learning Theory Perspective. *ACM Trans. Internet Things* 2, 3, Article 15 (jul 2021), 49 pages. https://doi.org/10.1145/3450494
[16] Wenliang Du and Mikhail J. Atallah. 2001. Secure multi-party computation problems and their applications: a review and open problems. In *Proceedings of the 2001 Workshop on New Security Paradigms (NSPW '01)*. Association for Computing Machinery, New York, NY, USA, 13–22. https://doi.org/10.1145/508171.508174
[17] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. 2016. Jali: an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on graphics (TOG)* 35, 4 (2016), 1–11.
[18] Huan Feng, Kassem Fawaz, and Kang G Shin. 2017. Continuous authentication for voice assistants. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*. ACM, 343–355.
[19] Ceenu George, M. Khamis, Emanuel von Zezschwitz, Marinus Burger, Henri Schmidt, Florian Alt, and Heinrich Hussmann. 2017. Seamless and Secure VR: Adapting and Evaluating Established Authentication Systems for Virtual Reality. https://api.semanticscholar.org/CorpusID:6671814
[20] Yuan Gong, Yu-An Chung, and James Glass. 2021. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778* (2021).
[21] Bill Hardcastle and Kris Tjaden. 2008. Coarticulation and speech impairment. *The handbook of clinical linguistics* (2008), 506–524.
[22] Rosa González Hautamäki, Tomi Kinnunen, Ville Hautamäki, Timo Leino, and Anne-Maria Laukkanen. 2013. I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry. In *Interspeech*. Citeseer, 930–934.
[23] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 770–778. https://api.semanticscholar.org/CorpusID:206594692
[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.
[25] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9 (1997), 1735–1780. https://api.semanticscholar.org/CorpusID:1915014
[26] Yi-Ta Hsieh, Antti Jylhä, Valeria Orso, Luciano Gamberini, and Giulio Jacucci. 2016. Designing a willing-to-use-in-public hand gestural interaction technique for smart glasses. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 4203–4215.
[27] Tomi Kinnunen, Md Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee. 2017. The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. (2017).
[28] Oscar Koller, Hermann Ney, and Richard Bowden. 2015. Deep Learning of Mouth Shapes for Sign Language. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*. 477–483. https://doi.org/10.1109/ICCVW.2015.69
[29] Marc Lavielle. 2005. Using penalized contrasts for the change-point problem. *Signal processing* 85, 8 (2005), 1501–1510.
[30] Soonkyu Lee and Dongsuk Yook. 2002. Audio-to-Visual Conversion Using Hidden Markov Models. In *Pacific Rim International Conference on Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:32064363
[31] Feng Li, Jiayi Zhao, Huan Yang, Dongxiao Yu, Yuanfeng Zhou, and Yiran Shen. 2023. VibHead: An Authentication Scheme for Smart Headsets through Vibration. *arXiv preprint arXiv:2306.17002* (2023).
[32] Florian Mathis, Hassan Ismail Fawaz, and M. Khamis. 2020. Knowledge-driven Biometric Authentication in Virtual Reality. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (2020). https://api.semanticscholar.org/CorpusID:212997365
[33] Yan Meng, Zichang Wang, Wei Zhang, Peilin Wu, Haojin Zhu, Xiaohui Liang, and Yao Liu. 2018. WiVo: Enhancing the Security of Voice Control System via Wireless Signal in IoT Environment. *Proceedings of the Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing* (2018). https://api.semanticscholar.org/CorpusID:49354919
[34] Yan Michalevsky, Dan Boneh, and Gabi Nakibly. 2014. Gyrophone: Recognizing Speech from Gyroscope Signals. In *Proceedings of USENIX Security Symposium*. 1053–1067.
[35] M. G. Sarwar Murshed, Christopher Murphy, Daqing Hou, Nazar Khan, Ganesh Ananthanarayanan, and Faraz Hussain. 2021. Machine Learning at the Network Edge: A Survey. *ACM Comput. Surv.* 54, 8, Article 170 (oct 2021), 37 pages. https://doi.org/10.1145/3469029
[36] Oculus. 2023. Oculus PC SDK v23. (2023). https://developer.oculus.com/downloads/ package/oculus-sdk-for-windows/.
[37] Ilesanmi Olade, Charles Fleming, and Hai-Ning Liang. 2020. BioMove: Biometric User Identification from Human Kinesiological Movements for Virtual Reality Systems. *Sensors (Basel, Switzerland)* 20 (2020). https://api.semanticscholar.org/CorpusID:218908243
[38] Ilesanmi Olade, Hai-Ning Liang, Charles Fleming, and Christopher Champion. 2020. Exploring the Vulnerabilities and Advantages of SWIPE or Pattern authentication in Virtual Reality (VR). *Proceedings of the 2020 4th International Conference on Virtual and Augmented Reality Simulations* (2020). https://api.semanticscholar.org/CorpusID:218830937
[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *ArXiv* abs/1505.04597 (2015). https://api.semanticscholar.org/CorpusID:3719281
[40] Flavio Schneider. 2023. Archisound: Audio generation with diffusion. *arXiv preprint arXiv:2301.13267* (2023).
[41] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 4779–4783.
[42] Cong Shi, Yan Wang, Yingying Chen, Nitesh Saxena, and Chen Wang*. 2020. WearID: Low-Effort Wearable-Assisted Authentication of Voice Commands via Cross-Domain Comparison without Training. In *Annual Computer Security Applications Conference (ACSAC)*. 829–842.
[43] Cong Shi, Xiangyu Xu, Tianfang Zhang, Payton Walker, Yi Wu, Jian Liu, Nitesh Saxena, Yingying Chen, and Jiadi Yu. 2021. 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*. 478–490.
[44] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur. 2017. Deep Neural Network Embeddings for Text-Independent Speaker Verification. In *Interspeech*. 999–1003.
[45] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 5329–5333.
[46] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez. 2014. Deep neural networks for small footprint text-dependent speaker verification. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4052–4056.

[47] Ethan Waisberg, Joshua Ong, Mouayad Masalkhi, Nasif Zaman, Prithul Sarker, Andrew G Lee, and Alireza Tavakkoli. 2023. The future of ophthalmology and vision science with the Apple Vision Pro. *Eye* (2023), 1–2.

[48] Ruxin Wang, Long Huang, and Chen Wang. 2023. Low-effort VR Headset User Authentication Using Head-reverberated Sounds with Replay Resistance. *2023 IEEE Symposium on Security and Privacy (SP)* (2023), 3450–3465. https://api.semanticscholar.org/CorpusID:260003730

[49] Xue Wang and Yang Zhang. 2021. 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*. 1–7.

[50] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135* (2017).

[51] Wikipedia. 2023. Topographic Prominence. (2023). https://en.wikipedia.org/wiki/Topographic_prominence.

[52] Tong Wu, Zhihao Fan, Xiao Liu, Yeyun Gong, Yelong Shen, Jian Jiao, Hai-Tao Zheng, Juntao Li, Zhongyu Wei, Jian Guo, et al. 2023. AR-Diffusion: Auto-Regressive Diffusion Model for Text Generation. *arXiv preprint arXiv:2305.09515* (2023).

[53] Dhruv Kumar Yadav, Beatrice Ionascu, Sai Vamsi Krishna Ongole, Aditi Roy, and Nasir D. Memon. 2015. Design and Analysis of Shoulder Surfing Resistant PIN Based Authentication Mechanisms on Google Glass. In *Financial Cryptography Workshops*. https://api.semanticscholar.org/CorpusID:16027399

[54] Chen Yan, Yan Long, Xiaoyu Ji, and Wenyuan Xu. 2019. The Catcher in the Field: A Fieldprint based Spoofing Detection for Text-Independent Speaker Verification. *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security* (2019). https://api.semanticscholar.org/CorpusID:207944110

[55] Shanhe Yi, Zhengrui Qin, Ed Novak, Yafeng Yin, and Qun Li. 2016. Glassgesture: Exploring head gesture interface of smart glasses. In *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*. IEEE, 1–9.

[56] Chunlei Zhang, Kazuhito Koishida, and John H. L. Hansen. 2018. Text-Independent Speaker Verification Based on Triplet Convolutional Neural Network Embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26, 9 (2018), 1633–1644. https://doi.org/10.1109/TASLP.2018.2831456

[57] Li Zhang, Parth H Pathak, Muchen Wu, Yixin Zhao, and Prasant Mohapatra. 2015. Accelword: Energy efficient hotword detection through accelerometer. In *Proceedings of the Annual International Conference on Mobile Systems, Applications, and Services (ACM MobiSys)*. ACM, 301–315.

[58] Linghan Zhang, Sheng Tan, and Jie Yang. 2017. Hearing Your Voice is Not Enough: An Articulatory Gesture Based Liveness Detection for Voice Authentication. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 57–71.

[59] Linghan Zhang, Sheng Tan, Jie Yang, and Yingying Chen. 2016. Voicelive: A phoneme localization based liveness detection for voice authentication on smartphones. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 1080–1091.

[60] Yingke Zhu and Brian Mak. 2023. Bayesian Self-Attentive Speaker Embeddings for Text-Independent Speaker Verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), 1000–1012. https://doi.org/10.1109/TASLP.2023.3244502

A appendix

Table 4: Voice commands involved in viseme collection.

Index	Voice command	Index	Voice command
1	Open Facebook.com.	16	Find some racing games.
2	What's my battery?	17	Show my favorite videos.
3	Open Beat Saber.	18	Show me events.
4	How do I change my profile picture?	19	Go to Photoshop.
5	Show me my packages.	20	Use with gaze.
6	Send a message.	21	Hide menu.
7	Call my mother.	22	Thank you.
8	Set the volume to full.	23	Hey, Facebook.
9	Turn off Bluetooth.	24	Open library.
10	Teleport.	25	Lower the volume to three.
11	Take a picture.	26	Reset view.
12	Shut down.	27	Reset guardian.
13	Disable airplane mode.	28	Turn on airLink
14	Turn on airplane mode.	29	Restart.
15	What's the weather next week?	30	Show me my alarms.