Face-Mic: Inferring Live Speech and Speaker Identity via Subtle Facial Dynamics Captured by AR/VR Motion Sensors

Cong Shi
WINLAB, Rutgers
cs1421@scarletmail.rutgers.edu

Xiangyu Xu
Shanghai Jiao Tong University
chillex@sjtu.edu.cn

Tianfang Zhang
WINLAB, Rutgers
tz203@scarletmail.rutgers.edu

Payton Walker
Texas A&M University
prw0007@tamu.edu

Yi Wu
University of Tennessee, Knoxville
ywu83@vols.utk.edu

Jian Liu
University of Tennessee, Knoxville
jliu@utk.edu

Nitesh Saxena
Texas A&M University
nsaxena@tamu.edu

Yingying Chen*
WINLAB, Rutgers
yingche@scarletmail.rutgers.edu

Jiadi Yu
Shanghai Jiao Tong University
jiadiyu@sjtu.edu.cn

ABSTRACT

Augmented reality/virtual reality (AR/VR) has extended beyond 3D immersive gaming to a broader array of applications, such as shopping, tourism, education. And recently there has been a large shift from handheld-controller dominated interactions to headset-dominated interactions via voice interfaces. In this work, we show a serious privacy risk of using voice interfaces while the user is wearing the face-mounted AR/VR devices. Specifically, we design an eavesdropping attack, *Face-Mic*, which leverages speech-associated subtle facial dynamics captured by *zero-permission* motion sensors in AR/VR headsets to infer highly sensitive information from *live human speech*, including speaker gender, identity, and speech content. *Face-Mic* is grounded on a key insight that AR/VR headsets are closely mounted on the user's face, allowing a potentially malicious app on the headset to capture underlying facial dynamics as the wearer speaks, including movements of facial muscles and bone-borne vibrations, which encode private biometrics and speech characteristics. To mitigate the impacts of body movements, we develop a signal source separation technique to identify and separate the speech-associated facial dynamics from other types of body movements. We further extract representative features with respect to the two types of facial dynamics. We successfully demonstrate the privacy leakage through AR/VR headsets by deriving the user's gender/identity and extracting speech information via the development of a deep learning-based framework. Extensive experiments using four mainstream VR headsets validate the generalizability, effectiveness, and high accuracy of *Face-Mic*.

CCS CONCEPTS

• **Security and privacy → Hardware attacks and countermeasures**.

KEYWORDS

Facial dynamics; AR/VR headsets; Speech and speaker privacy

ACM Reference Format:
Cong Shi, Xiangyu Xu, Tianfang Zhang, Payton Walker, Yi Wu, Jian Liu, Nitesh Saxena, Yingying Chen, and Jiadi Yu. 2022. Face-Mic: Inferring Live Speech and Speaker Identity via Subtle Facial Dynamics Captured by AR/VR Motion Sensors. In *The 27th Annual International Conference on Mobile Computing and Networking (ACM MobiCom '21), January 31-February 4, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3447993.3483272

1 INTRODUCTION

With the capability of creating 3D virtual worlds, which users can immerse in and interact with, augmented reality/virtual reality (AR/VR) devices have attracted millions of users. The market size is expanding drastically and is expected to reach 12.3 billion dollars by 2023 [36]. The rapid expansion of face-mounted devices (e.g., VR headsets) facilitates a broad array of AR/VR applications, including immersive multi-people gaming [40], virtual shopping [5], and banking [9]. As the AR/VR domain extends beyond 3D immersive gaming to a broader array of applications, the control logic of AR/VR devices has been largely shifting from controller-dominated interactions (which are mainly designed for gaming) towards headset-dominated interactions via voice user interfaces. For example, Oculus Quest supports voice dictation for entering web addresses [19], controlling the headset, and exploring commercial products [6, 26]. However, the frequent emerging usage of voice interface in AR/VR scenarios could result in severe privacy leakage if malicious actors can listen onto this communication medium. For instance, an adversary can snoop on sensitive information during AR/VR voice communications, such as credit card numbers and private healthcare/bank transaction information. Moreover, the personally identifiable information of headset wearers, such as gender and identities, could be leaked to the adversary, which may be leveraged for targeted advertising and fraud.

Due to these voice-related privacy concerns, the vendors of AR/VR headsets have rigorous policies on voice access and require explicit permission to use microphones. Given the privacy policies on smartphone-based Operating Systems (e.g., Android and iOS), low-cost cardboard headsets naturally require the highest level of permission to access microphones [2]. Similar policies are used in the operating systems of high-end standalone headsets [23, 28].

[*] Yingying Chen is the corresponding author.

ACM MobiCom '21, January 31-February 4, 2022, New Orleans, LA, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-8342-4/22/01…$15.00
https://doi.org/10.1145/3447993.3483272

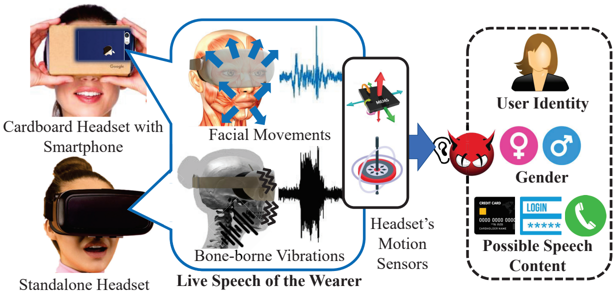


Figure 1: Illustration of the proposed eavesdropping attack leveraging facial dynamics (i.e., speech-associated facial movements and bone-borne vibrations) captured by AR/VR headsets' built-in motion sensors.

Therefore, it is not easy for a misbehaving app running on these headsets to gain access to the microphone and listen onto voice communications. In contrast, accessing the built-in motion sensors in VR headsets usually does not require user permission [27], since almost all AR/VR applications need to utilize the motion sensors to track the user's head motions and simulate the corresponding movements in the virtual environment. *Can these zero-permission motion sensors be used by an adversary to infer the live speech and speaker privacy of the headset user?*

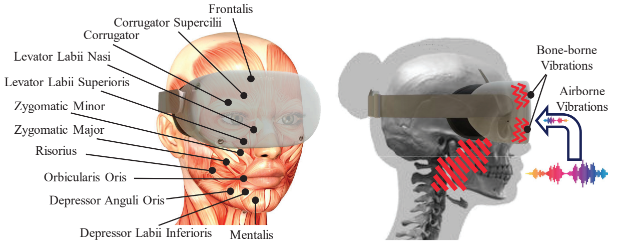
In this paper, we explore this question and demonstrate a new eavesdropping attack targeting AR/VR devices, *Face-Mic*. Our key insight is that the headset is closely mounted on the user's head and presses different parts of the face as shown in Figure 1. This unique and fundamental characteristic exists in both low-cost cardboard headsets (e.g., Google Cardboard [12]) and high-end standalone headsets (e.g., Oculus Quest, HTC Vive), and it makes the headset susceptible to the dynamics of the underlying facial muscles, which can reflect speech content as well as the wearer's unique private biometrics (e.g., behaviors of speaking and tissue properties). Furthermore, during the pronunciation of speech, the conductive vibrations (i.e., bone-borne vibrations) produced by the vocal cords can propagate through cranial bones, thereby vibrating the AR/VR headsets. By analyzing the captured facial dynamics, we show that for both cardboard headsets and standalone headsets, the adversary can easily infer the sensitive speech and speaker information, which raises extreme privacy concerns.

Fundamental Differences from Existing Attacks. *Face-Mic* exhibits several crucial differences compared to prior attacks [1, 4, 22]. First, it is the first motion sensor-based speech eavesdropping attack that targets AR/VR headsets, which represents a threat against a new emerging user interaction paradigm gaining rapid momentum in the real-world. Second, *Face-Mic* is designed to capture the *live human speech* of the device's wearer while prior attacks can only capture the speech, via smartphone motion sensors, that has been *replayed* by: (1) external loudspeakers [22], whose associated vibrations reach the smartphone through a shared surface propagation [1], or (2) in-built speakers of the smartphone [4] that create reverberations through the body of the smartphone. Smartphone motion sensors do not generally get impacted by the device user's live speech [1], which prevents these prior attacks from eavesdropping on the aerial speech of the wearer. Third, *Face-Mic* extracts the speech and speaker information via the subtle facial dynamics produced as the headset's wearer speaks, which is far more challenging due to the significant interference introduced by the user's body movements in immersive AR/VR scenarios (a challenge that we are able to overcome).

Challenges Addressed in Eliciting Speech via Facial Dynamics. To realize such an eavesdropping attack relying on built-in motion sensors, we face several challenges in practice: *1) Significant Impact Caused by Body Motion Artifacts:* In AR/VR scenarios, the headset wearer usually interacts with the virtual worlds through large-scale body movements. Therefore, *Face-Mic* needs to eliminate these motion artifacts to enable reliable facial dynamic extraction. *2) Unclear Response to Speech and Speaker Characteristics:* The relationship between the facial dynamics and the speaker/speech characteristics is not clear, so we need to explore the relationship between facial movements/bone-borne vibrations and speech. *3) Low-sampling Rate of Motion Sensors:* The built-in motion sensors in AR/VR headsets have limited sampling frequencies, which renders detecting live speech vibrations and its harmonics across $85Hz \sim 20kHz$ [24] highly challenging.

Proposed Face-Mic via Facial Dynamics Captured by AR/VR Motion Sensors. Based on the collected motion sensor data, *Face-Mic* first removes the artifacts of human body movements with a signal source separation technique, which utilizes time-frequency analysis to disentangle speech-associated facial movements among other types of body movements. Our attack system then separates the facial muscle movements and bone-borne vibrations based on their unique frequency bands. Through studying the characteristics of facial muscle movements and bone-borne vibrations, we extract two sets of features from the headsets' 3D acceleration, speed, and displacement, which capture the victim's unique private biometrics and sensitive speech content. Given the extracted features, Face-Mic performs gender detection, user identification, and speech recognition by developing a deep-learning based framework. Our main contributions are summarized as follows:

- To the best of our knowledge, *Face-Mic* is the first attack that infers private and sensitive information leveraging the facial dynamics associated with live human speech while using face-mounted AR/VR devices. By using zero-permission built-in motion sensors, *Face-Mic* can disclose the headset wearer's gender/identity and extract speech information.
- We thoroughly study the relationships between the speaker and speech characteristics and three types of vibrations captured by AR/VR headsets' motion sensors, including speech-associated facial movements, bone-borne vibrations, and airborne vibrations. We find that the speech effects exhibited in the motion sensor readings are dominated by the facial movements and bone-borne vibrations.
- We design a series of techniques to infer the headset wearer's gender, identity and simple speech, such as body motion artifact removal algorithm, feature extraction based on facial dynamics, and deep-learning-based sensitive information derivation.
- We validate the proof-of-concept attack by conducting extensive experiments with 4 mainstream VR headsets and 45 volunteers. The results show that *Face-Mic* can derive the headset wearer's gender, identity, and simple speech information.



(a) Speech-associated muscles (b) Bone-borne and airborne vibrations

Figure 2: Illustration of facial dynamics involved in human speech production.

2 PRELIMINARIES

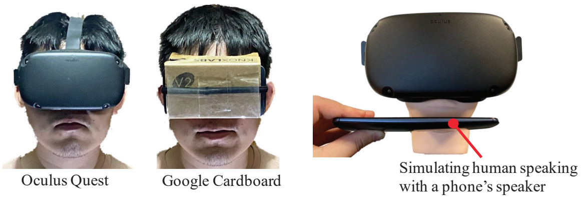
2.1 Speech-related Facial Dynamics

There are three types of speech-associated facial dynamics that could be captured by AR/VR motion sensors: speech-associated facial movements, bone-borne vibrations, and airborne vibrations.

Speech-associated Facial Movements. Human's facial muscles contract and relax regularly during speech production, which encodes both speech information (e.g., phoneme, tempo, loudness) and biometric characteristics (e.g., behaviors of speaking, face shapes, muscle, and tissue properties). Specifically, as shown in Figure 2(a), a subset of 12 pairs of muscles are involved in human speech, which can be categorized into two groups: upper face muscles and perioral muscles. Upper face muscles are the muscles that surround the eye socket, including corrugator, corrugator supercilli, etc. When wearing AR/VR headsets, these muscles are in direct contact with the device, and thus the muscles' contractions/relaxation while the user is speaking can directly move and rotate the headset in the 3D space, which can be captured by the built-in motion sensors. On the other hand, perioral muscles are the group of muscles that encircle the mouth, including depressor anguli oris, zygomatic major, etc., which are usually not in direct contact with the AR/VR headset. Yet during speech production, the strong contractions/relaxation of these muscles around the mouth could propagate to facial tissues that are in contact with the headset, thereby influencing AR/VR motion sensor readings in an indirect way.

Bone-borne Vibrations. Bone-conduction vibrations are the acoustic vibrations generated by human vocal folds and then propagate through cranial bones [21]. As a key organ in creating sounds, when humans speak, the vocal folds modulate the flow of air being expelled from the lungs during phonation. The vibrations are then filtered and modulated by the vocal tract, rendering human recognizable speech. Part of the vibration signal propagates through the cranial bones, and thus the vibrations can be measured by the built-in motion sensors in AR/VR headset that are closely mounted on the user's head as shown in Figure 2(b). Since the vibration signals are directly produced by the human sound production system, they are highly correlated with the human recognizable speech signals. Additionally, the bone-borne vibrations can also capture unique biometrics in users' sound production systems.

Airborne Vibrations. The airborne vibrations are the acoustic vibrations propagating over the air. Existing studies (e.g., [34, 41]) have shown that the accelerometer in smartphones and smartwatches can respond to airborne human voice at a close distance (e.g., 30cm for smartwatches [34]). Therefore, it is very likely that the AR/VR motion sensors can also capture such minute airborne vibrations given the short physical distance between the user's mouth and the AR/VR headset.



(a) Setup to capture live human speech (b) Setup to study airborne vibrations

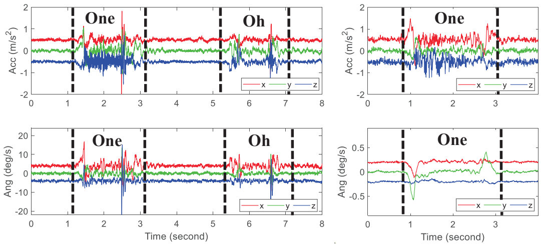
Figure 3: Experimental setup to study the impacts of speech-associated facial movements, bone-borne and airborne vibrations captured by the motion sensors.

2.2 Capturing Facial Dynamics via Motion Sensors in AR/VR Headsets

Most AR/VR devices are equipped with motion sensors, typically including a three-axis accelerometer and a three-axis gyroscope. Besides measuring the acceleration/angular velocity of the devices, these motion sensors also pick up conductive [1] and aerial vibrations [34], making the AR/VR headsets capable of capturing the three types of aforementioned facial dynamics. To demonstrate the feasibility of using built-in motion sensors to eavesdrop on live human speech, we conduct preliminary experiments by examining the speech effects on two representative AR/VR headsets: a cardboard headset (Google Cardboard with Nexus 6) and a standalone headset (Oculus Quest) with sampling rates of $227Hz$ and $1000Hz$ for their motion sensors, respectively.

Capturing Live Human Speech via AR/VR Headsets. To examine the effects of live human speech, we ask a volunteer to wear the two headsets, as shown in Figure 3 (a), and speak a couple of words (i.e., "one", "oh"). The raw accelerometer and gyroscope readings of Oculus Quest and Google Cardboard are shown in Figure 4 (a) and (b), respectively. We can find that with the headset mounted on the user's face, the built-in accelerometer and gyroscope can respond to the subject's speech, showing significant signal fluctuations. Such signal fluctuations can be observed across all three axes of the two sensors, showing their high sensitivity to speech associated facial dynamics. We then analyze the speech in the time-frequency domain by applying short-time Fourier transform to the motion sensor readings and obtain the spectrogram as shown in Figure 5(a). For Oculus Quest, we find that besides the strong responses in the low-frequency range (e.g., <100Hz), the spectrogram also shows high energy at high frequencies (i.e., $100 \sim 500Hz$). Meanwhile, Google Cardboard can only capture responses below $114Hz$ due to the low sampling rate.

Response Verification for Facial Dynamics. To further identify what types of facial dynamics are captured in the motion sensor readings, we conduct an experiment by asking the subject to perform facial movements of words "one" and "oh" without pronouncing, so that only facial movements are involved. We show the corresponding spectrograms of Oculus Quest and Google Cardboard in Figure 5 (b). An interesting finding is that compared to the spectrograms in Figure 5 (a), only the low-frequency responses remain, while the high-frequency responses (>100Hz) disappear.



(a) Oculus Quest (b) Google Cardboard

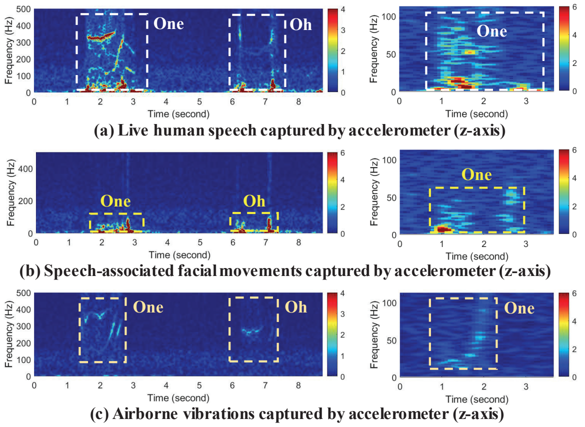
Figure 4: The response of accelerometer and gyroscope in two types of AR/VR headsets to live human speech.

Therefore, we can safely attribute the low-frequency responses of motion sensors to the facial movements. Furthermore, to study the influence of airborne vibrations, we replay the speech (i.e., "one", "oh") towards the headset mounted on a mannequin head as shown in Figure 3 (b), without any physical contact between the playback device (i.e., a smartphone's speaker) and the headsets. As shown in Figure 5 (c), only weak energy at the high-frequency band of 100 ~ 500Hz can be observed. Compared to the spectrograms in Figure 5 (a) with all three types of facial dynamics, and also Figure 5 (b) where only facial movements are involved, we could find that bone-borne and airborne vibrations have overlapping responses, while the response of bone-borne vibrations is much stronger than the corresponding airborne counterparts.

With all the above observations, we conclude that the speech-associated facial movements mainly influence the low-frequency (<100Hz) motion sensor readings, while the bone-borne vibrations strongly impact the sensor readings at high frequencies (e.g., >100Hz). Note that although vocal folds of male speakers can produce sound as low as 85Hz, the human skull only responds to sound vibrations at much higher frequencies, usually over 250Hz [13]. Therefore, bone-borne vibrations do not have strong energy below 100Hz. Since airborne vibrations share similar physical characteristics and time-frequency patterns with bone-borne vibrations, but the responses are much weaker, we consider these two vibrations together as *bone-borne vibrations*. Thus, in the rest of the paper, we exploit the facial movements and bone-borne vibrations (include the airborne vibrations) to realize *Face-Mic*.

3 ATTACK OVERVIEW AND THREAT MODEL

Privacy Leakage. *Face-Mic* can reveal private information associated with the user identity, such as the user's favorite AR/VR games, AR/VR travel histories, and watching/shopping preferences, which can be lucrative for advertising companies [25]. To derive the identity information, the adversary can eavesdrop on speech in various AR/VR scenarios, such as conversations during multi-player gaming and AR/VR meeting. At the same time, the adversary can also detect the gender of the victim, which can be used for advertising gender-specific products or analyzing gender-specific behaviors during AR/VR shopping [5], Internet surfing, or AR/VR social media usage [29], without the user's permission. More importantly, *Face-Mic* can derive simple speech content, i.e., digits and words. These two types of speech content can be used to infer a broad array of sensitive information, such as social security numbers, phone numbers, passwords, transactions, and healthcare information. Exposing such information could lead to identity theft, credit card fraud, confidential and healthcare information leakage, which puts the user's security and privacy at high risk. To obtain such sensitive speech information, the adversary can launch the attack when the victim is using voice dictation or chatting with other people during virtual communication.



(a) Live human speech captured by accelerometer (z-axis)

(b) Speech-associated facial movements captured by accelerometer (z-axis)

(c) Airborne vibrations captured by accelerometer (z-axis)

Figure 5: The frequency responses of accelerometer (z-axis) of the standalone headset Oculus Quest (on the left) and the cardboard headset (on the right). The facial movements, bone-borne vibrations, and airborne vibrations can be captured in the accelerometer readings of both headsets.

Adversary's Capability. We assume the adversary has a malicious app installed on the victim's AR/VR device, either through fooling the victim to install the app or posting the app on app stores. The malicious app collects motion sensor data in the background and sends the data to the remote adversary for gender/speaker detection and speech recognition. Since accessing the motion sensor does not require any permission, the app can be disguised as any AR/VR app (e.g., AR/VR games, web browsers). Such a malicious-app-based threat model has shown its effectiveness in smartphones [4, 22], and it can be directly applied to the cardboard headsets using smartphones as the central processing units. Our study on two mainstream AR/VR programming platforms (Oculus [27], OpenVR [37]) also confirms that such a threat model is feasible for standalone headsets. We built an AR/VR app based on Oculus SDK (v23) and successfully used the function $ovr\_GetTrackingState()$ to record Oculus Quest's accelerometer/gyroscope data in the background without user permission. We also confirm that such an app can be easily programmed with OpenVR, which supports a broader range of headsets (e.g., manufactured by HTC, Valve, and most Windows Mixed Reality headset manufacturers). In OpenVR, the app uses $GetRawTrackedDevicePoses()$ to collect the motion sensor data, also without user permission.

Attack Scenarios. We study the following three representative scenarios that could happen in practical environments:

Scenario-1: Attack with prior victim data. The adversary has opportunities to get access to the victim's motion sensor data (e.g., via the malicious app) and labels ahead of time. For gender detection and speaker identification, the labels are the victim's gender and identity, respectively. For speech recognition, the labels are the audio data of the victim. In practice, these labels can be obtained in several ways. For instance, the adversary is a friend of the victim

Table 1: Three attack scenarios studied in this work.

	Require the victim's data and corresponding labels (e.g., gender, identity, or audio data) before launching attacks?	Leverage the motion sensor data collected during attack phase to adapt the model to the victim?	Possible inference tasks

Scenario-1	✔	×	Gender detection;
			Speaker identification;
			Speech recognition

Scenario-2	×	✔	Gender detection; Speech recognition
Scenario-3	×	×	Gender detection; Speech recognition

and he/she knows the victim's gender and identity. If the adversary has a chance to be in the same room when the victim uses the AR/VR headset, the adversary can collect the victim's motion sensor data via the malicious app and at the same time record the victim's audio data using a microphone. The adversary may also record the victim's speech remotely in some AR/VR scenarios, where the adversary is communicating on a shared audio channel with the victim, such as multi-player AR/VR gaming or virtual meeting. In this case, the adversary can collect audio data from the shared audio channel using a microphone. The adversary then correlates the motion sensor data and the labels to train *Face-Mic*'s deep learning model for gender detection, speaker identification, and speech recognition. We note that most security studies [1, 4, 22] in this line of research (motion-sensor-based privacy leakage) have been reported under this attack scenario. We also note that speaker identification can only be launched under this scenario as the adversary needs to know the victim's identity as a priori.

Scenario-2: Adaptive training during attack phase. The adversary will leverage the victim's motion sensor data collected during the attack phase to perform adaptive training. Particularly, the adversary has a pre-trained deep learning model, built on other people's data. During the attack phase, the collected motion sensor data will be utilized to update the parameters of the pre-trained model, making the model better fit the victim's facial features. Then, the adapted model will be utilized to perform gender detection and speech recognition.

Scenario-3: Real-time attack without prior victim data. This is the most challenging attack scenario. Once the victim's motion sensor data is collected, the adversary will directly use the pre-trained model built on other people's data to perform gender detection and speech recognition. This scenario does not use the collected motion sensor data to perform adaptive training as in Scenario-2 and thus the sensitive information can be inferred in real time.

We would like to point out that in most of such attacks, the adversary does not need to infer the sensitive information in real time. So the attack under Scenario-2 is a more acceptable attack under practical constraints (e.g., without the direct access to victim's labeled data). We summarize the three attack scenarios in Table 1.

4 ATTACK DESIGN

4.1 Challenges

Significant Impact Caused by Body Motion Artifacts. In AR/VR scenarios, the headset wearer usually interacts with the virtual worlds through large-scale body movements, such as moving within the play area, rotating the head, and moving the controllers. Such unpredictable movements produce a significant amount of artifacts in the motion sensors' readings. Thus, we need to eliminate these motion artifacts to enable reliable facial dynamic extraction.

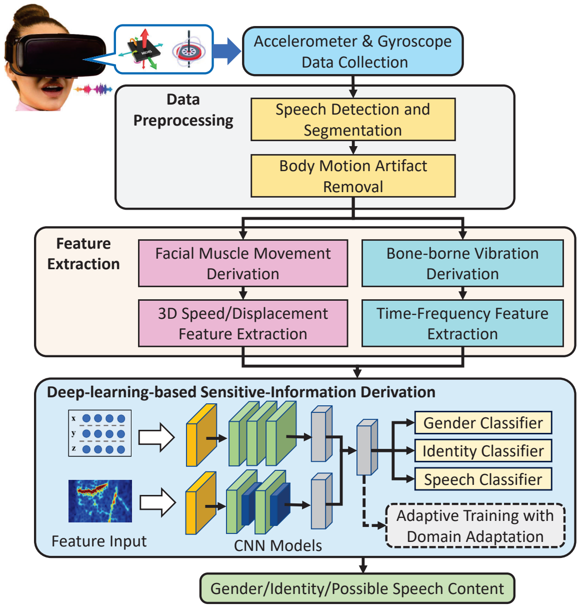

Accelerometer & Gyroscope Data Collection
Data Preprocessing
Speech Detection and Segmentation
Body Motion Artifact Removal
Feature Extraction
Facial Muscle Movement Derivation
Bone-borne Vibration Derivation
3D Speed/Displacement Feature Extraction
Time-Frequency Feature Extraction
Deep-learning-based Sensitive-Information Derivation
x
y
z
Gender Classifier
Identity Classifier
Speech Classifier
Adaptive Training with Domain Adaptation
Feature Input
CNN Models
Gender/Identity/Possible Speech Content

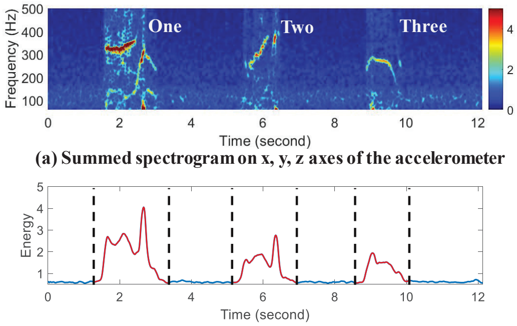
Figure 6: System overview of *Face-Mic*.

Deriving Speech and Speaker Characteristics from Facial Dynamics. The relationship between the facial dynamics and the speaker/speech characteristics remains unclear. We need to explore the relationship between facial movements/bone-borne vibrations and speech, and extract representative features that carry unique private biometrics and speech characteristics.

Low-sampling Rate of Motion Sensors. The built-in motion sensors in AR/VR headsets have only around $200Hz$ for Google Cardboard and $1000Hz$ for Oculus Quest, while the human voice and its harmonics span across $85Hz \sim 20kHz$ [24]. Such low sampling frequencies mean the motion sensor only captures low-fidelity speech characteristics. To realize an effective attack, *Face-Mic* needs to derive reliable measurements to best capture the embedded speech-associated facial dynamics.

4.2 System Overview

The basic idea of *Face-Mic* is to capture speech associated facial dynamics to reveal the encoded gender, identity, and speech information. As illustrated in Figure 6, the malicious app monitors the motion sensor in the background and detects human speech based on the high-frequency bone-borne vibrations, which are only present during speech pronunciation. Upon detecting human speech, the app segments the motion sensor data associated with speech and sends the segmented data to a remote adversary for data processing. To deal with motion artifacts (e.g., head rotations) that contaminate the sensors' readings, we then develop a signal source separation technique based on time-frequency analysis to disentangle the measurements associated with facial movements from the contaminated accelerometer/gyroscope data.



(a) Summed spectrogram on x, y, z axes of the accelerometer

(b) Accumulated spectrogram energy over 100Hz

Figure 7: Illustration of speech detection and segmentation (i.e., on Oculus Quest) based on the high-frequency bone-borne vibrations (i.e., over 100Hz).

Next, *Face-Mic* separates facial muscle movements and bone-borne vibrations from the pre-processed data and extract effective features from them, respectively. Since facial muscle movements and bone-borne vibrations reside in different frequency bands as discussed in Section 2, we use a filter to separate these two types of facial dynamics. Then for facial muscle movements, we first calculate 3D speed and displacement/rotation by applying numerical integration on the three-axis accelerometer/gyroscope readings, which characterize the headset's spatial movements involved in speaking. On top of that, a set of 11 time-domain features and 2 frequency-domain features are extracted. For bone-borne vibrations, we explore the time-frequency representations (i.e., spectrogram) as the feature map to capture high-frequency patterns.

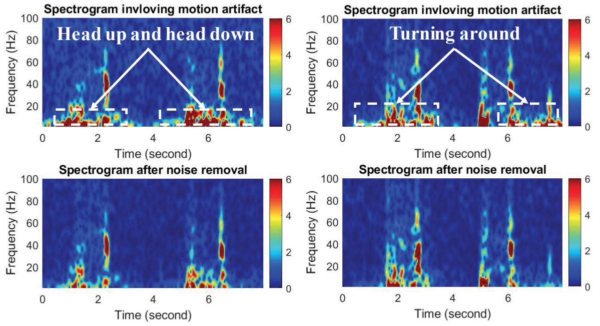
Based on the features of facial dynamics, we develop a deep-learning-based framework to derive speaker- and speech-related sensitive information. Two convolutional neural network (CNN) models are used to derive the feature representations of facial movements and bone-borne vibrations, and the derived feature representations are then concatenated and fed to a SoftMax layer for the three inference tasks. If the victim's motion sensor data and labels can be obtained ahead of time, *Face-Mic* correlates the motion sensor data and the labels to train the CNN model for gender detection, speaker identification, and speech recognition (i.e., *Scenario-1*). Otherwise, *Face-Mic* could utilize the victim's motion sensor data collected during the attack phase to adapt the pre-trained CNN model built on other people to the victim's feature space to improve the inference accuracy (i.e., *Scenario-2*). To realize such an attack, we design an adaptive training scheme based on domain adaptation to update the parameters of the pre-trained CNN model, making the knowledge learned from the pre-trained model to be transferred to the inference task targeting the victim. The adversary can also directly apply the pre-trained CNN model without adaptive training (i.e., *Scenario-3*), enabling the adversary to infer sensitive information in a real-time manner.

5 DATA PREPROCESSING

5.1 Speech Detection and Segmentation

To conduct a practical eavesdropping attack, we first need to detect the presence of human speech based on the motion sensor readings. In AR/VR scenarios, the built-in motion sensor of the headset can capture other body movements, thereby making it very difficult to perform speech detection. To circumvent this issue, we leverage the fact that most human body movements reside in low-frequency bands [8] and detect human speech through examining the presence of high-frequency bone-borne vibrations, which are only present during speech pronunciation as demonstrated in Section 2. We empirically find that the accelerometer is more sensitive to speech-associated vibrations than the gyroscope, which is also supported by existing studies [1, 34, 39], so we use it for speech detection and segmentation. Particularly, we calculate the spectrograms of x, y, z axes of the accelerometer by applying Short-Time Fast Fourier Transform (STFT) and conduct element-wise summation on their magnitude. Based on the summed spectrogram, we accumulate the energy across frequencies over $100Hz$ to detect speech.

We conduct an experiment by asking a subject to wear the Oculus Quest and speak three words (i.e., "one", "two", "three"), and the summed spectrogram and the accumulated spectrogram energy of the three words are shown in Figure 7 (a) and (b). We observe that the accumulated energy exhibits high values within the area of speech. We are thus inspired to utilize a threshold-based method to detect the starting point and the ending point of the speech. Figure 7 (b) shows that our method can correctly locate the starting and ending points of the speech, which confirms its effectiveness.



(a) Conducting head movements

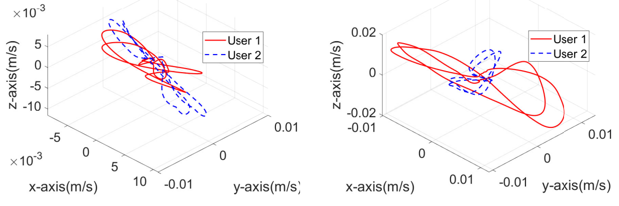
(b) Conducting body movements

Figure 8: Demonstration of removing the motion artifact entangled in the spectrogram of facial movements (i.e., associated with "one" and "oh").

5.2 Body Motion Artifact Removal

To achieve reliable facial dynamic extraction, we need to remove the motion artifacts contaminating the motion sensor data. Since human movements normally impact the motion sensor readings below $60Hz$ [8], a straight-forward approach is to employ a high-pass filter to remove the low-frequency artifacts. However, finding an optimal cut-off frequency is challenging, since the speech-associated facial movements are also captured in readings below $60Hz$. Therefore, we develop a body motion artifact removal (BMAR) approach based on signal source separation techniques [3], which were used to separate the mixed speech of multiple speakers in audio recordings, to extract the sensor readings of facial movements.

We model the problem of signal source separation as a regression problem (i.e., estimating the "clean" motion sensor data based on the noisy readings distorted by human movements), and develop a deep regression model. The regression model takes the spectrogram of the accelerometer/gyroscope readings as input. Particularly, we use



(a) "One" (b) "Two"

Figure 9: 3D headset speeds from two wearers are distinguishable when each wearer speaks "one" and "two".

the spectrogram across all available frequencies (e.g., $0 \sim 500Hz$ for Oculus Quest), so that the more robust bone-borne vibrations can help in separating signals of facial movements from noisy motion sensor readings. Given the spectrogram of motion sensor readings, $X(t, f)$, the objective of the deep regression model is to estimate a mask $\hat{M}_s(t, f)$ that reconstructs the spectrogram of speech:

$$\hat{X}(t, f) = \hat{M}_s(t, f) \circ X(t, f), \tag{1}$$

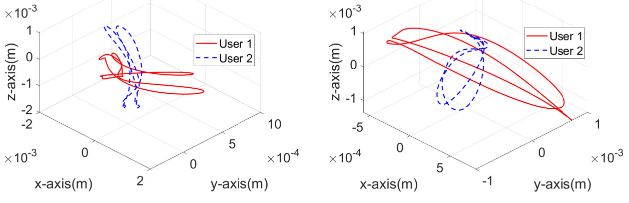
where $\circ$ denotes the element-wise product of the two operands. An inverse short time Fourier Transform can then be applied to reconstruct the "clean" motion sensor data. To realize such a regression model, we build a deep neural network consisting of two fully-connected layers for representation derivation and a regression layer for spectrogram mapping. A dropout layer is attached to each fully-connected layer to prevent over-fitting.

To train the deep regression model, we collect motion sensor data of speech (i.e., no body movement involved) and a set of representative body movements (e.g., head rotations, body movements) without speaking, separately, and then mix them to generate the training data. The sensor data of speech is used as the target regression variables. By using this approach, our attack can produce a huge amount of training data at a very low cost. Both the generated training data and the target regression variables are fed to the deep regression model for training. We use the mean square error as the loss for the optimization. Figure 8 (a) and (b) show the effects of noise removal for a head movement and a body movement. We can observe that the energy of these two movements is significantly reduced after passing our regression model, though minor energy remains. The results confirm the effectiveness of the proposed BMAR approach based on deep regression.

6 FEATURE EXTRACTION FOR FACIAL MOVEMENT AND BONE-BORNE VIBRATION

Given the denoised motion sensor data, we use a low-pass filter and a high-pass filter, with the same cut-off frequency of $100Hz$, to extract facial movements and bone-borne vibrations, respectively.

Feature Set for Facial Movements. Based on the accelerometer readings, we calculate the 3D speed and displacement of the VR/AR headset through the first- and second-order numeric integration, which depict the geometric kinematics model of facial muscle movements when a specific user speaks specific content. Since raw motion sensor readings involve substantial hardware noises that bring linearly increasing integration errors over time, we take the starting and ending point of a data segment as references and then calculate and subtract the average integration error from each data point in the segment. We show the calibrated speed and displacement of two digits (i.e., "one", "two") in Figure 9 and Figure 10, respectively. We observe for the same digit, both speed and displacement exhibit different 3D trajectories between the two users, showing that such measurements may embed personal characteristics. Furthermore, for the same user, the distinctive speed and displacement trajectories between the two digits are different, which also confirm the existence of content-related patterns.



(a) "One" (b) "Two"

Figure 10: 3D headset displacements from two wearers are distinguishable when each wearer speaks "one" and "two".

Face-Mic extracts 11 time-domain features and 2 frequency domain features from short frames of 3D acceleration, speed, displacement, as well as angular velocity. A sliding time window with a length of $128ms$ and a step size of $16ms$ is used to sample the short frames. The time-domain features include variance, maximum, minimum, range, mean, root mean square, median, interquartile range, mean crossing rate, skewness, and kurtosis. These statistic features encode the magnitude and speed of facial muscle movements and properties (e.g., size and strength of facial muscles). In addition, we extract frequency domain features by applying FFT on the accelerometer/gyroscope readings of each frame. The FFT coefficients are used to derive frequency-domain features, including mean and entropy of energy, which capture the periodic nature of speech/speaking behaviors. In total, we extract 234 time- and frequency-domain features from each frame.

Feature Set for Bone-borne Vibrations. Since bone-borne vibrations are only present in high-frequency ranges, we calculate the spectrogram of accelerometer and gyroscope readings and use them as features. We do not extract time-domain features from the bone-borne vibrations, since the high-pass filter used to extract the vibrations can significantly distort the time-domain characteristics (e.g., mean) in low-frequency sensor readings, rendering the time-domain features not stable. Given sensor readings of accelerometer and gyroscope, *Face-Mic* computes spectrogram based on the sensor readings of x, y, z axes, and removes the frequency components below $100Hz$. To provide fine-grained frequency representations, we compute the spectrogram by applying $1000 - point$ FFT in each $128ms$ Hanning window, shifting $16ms$ each time.

7 DEEP LEARNING-BASED SENSITIVE INFORMATION DERIVATION FRAMEWORK

Given the extracted features, we develop a deep-learning-based framework to perform sensitive information derivation. If the victim's motion sensor data and labels (gender, identity, or audio data) can be obtained ahead of time, *Face-Mic* correlates the motion sensor data and the labels to train a deep learning model based on CNN (i.e., Scenario-1). Otherwise, *Face-Mic* will only leverage the motion sensor data collected during the attack phase to adapt the pre-trained deep learning model built on other people's data, to the victim's feature space to improve the accuracy. We build an adaptive training scheme (Section 7.2) grounded on domain adaptation, which transfers the knowledge learned from the pre-trained model to the inference tasks targeted at the victim (i.e., Scenario-2). The adversary can also directly utilize the pre-trained model to perform sensitive information derivation, without utilizing the victim's motion sensor data for adaptive training (i.e., Scenario-3).

7.1 CNN-based Sensitive Information Derivation

Representation Extractor. Since the features of facial movements and bone-borne vibrations have very different properties and dimensions, we use two CNN models to process the features of these two types of facial dynamics as shown in Figure 11. For both CNNs, a batch normalization layer is applied to the input features to remove the mean and scale the features to unit variance, aiming to mitigate small-scale variations across data samples. To process the features of facial movements, we use a CNN consisting of 3 convolutional layers with 2D kernels to calculate feature maps. The x, y, z axes of the accelerometer/gyroscope are considered as 3 separated channels of the CNN. For the bone-borne vibration, due to the large size of the spectrograms, we attach a max-pooling layer to each convolutional layer for dimension reduction. The 2D feature maps of the two CNNs are then flattened and compressed with two fully-connected layers. The concatenated outputs of the two CNNs are used as feature representations to perform the attack. The activations of all layers are ReLU.

Sensitive Information Classifier. To derive sensitive information, we feed the feature representations to a classifier, which consists of two fully-connected layers and a SoftMax layer to map the feature representations into the probabilities over different classes (e.g., different speakers). During training, we use categorical cross-entropy as the loss function, which examines the differences between the model predictions and the labels. Based on the adversary's objective, the sensitive information classifier can be easily modified to perform the three inference tasks (i.e., gender detection, speaker identification, and speech recognition).

7.2 Unsupervised Domain Adaptation

The facial dynamics usually contain substantial information specific to the headset wearer, including unique biometric characteristics (e.g., face shapes, tissue properties) and behaviors of speaking. The adversary may not able to obtain a sufficiently large dataset to build a general model to suppress such individual variations. To enable an effective attack while circumventing the training requirement, we apply a domain adaptation technique that can effectively transfer the knowledge learned from the pre-trained model to a specific inference task targeting the victim. Specifically, we employ domain adversarial training [43] to remove the speaker-dependent characteristics embedded in the facial features, by leveraging the victim's motion sensor data collected during the attack phase.

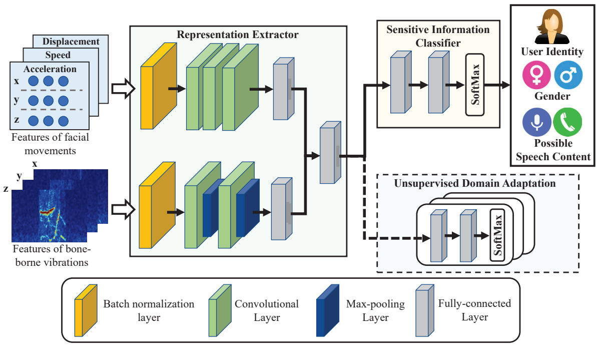


Figure 11: Deep learning-based framework used to derive, gender, identity, and speech content.

The key components of our scheme are a set of domain discriminators that adaptively update the parameters of the representation extractor as shown in Figure 11. We define a wearer (non-victim) as an individual involved in the training of the pre-trained model, and we build a domain discriminator (i.e., with 2 fully-connected layers and a SoftMax layer) for each wearer-victim pair, which adaptively transfers shared features from each wearer to the victim's feature space. By taking the representations from the representation extractor as input, each domain discriminator predicts the domain label (i.e., the victim or the corresponding wearer). Then, by applying a generative adversarial loss [11], we can use the domain discriminator to guide the representation extractor to learn speaker-independent representations. The idea is to apply a negative factor $-\lambda$ to the domain loss during the optimization, so that the representation extractor is trained to "confuse" each domain discriminator. Given K wearer-victim pairs (i.e., between each of the K wearers and the victim), the loss function to optimize the representation extractor is defined as:

$$L_f = log \sum_{k=1}^{K} exp(L_k^s - \lambda L_k^d), \tag{2}$$

where L_k^s denotes the loss of the classifier computed using the data of the k^{th} wearer, and L_k^d is the loss of the k^{th} domain discriminator. The log-sum-exp trick adaptively combines the losses of all k wearer-victim pairs and smooths the final loss for accelerating the convergence. The factor λ is used to balance the trade-off between the representations' transferability and the distinctiveness.

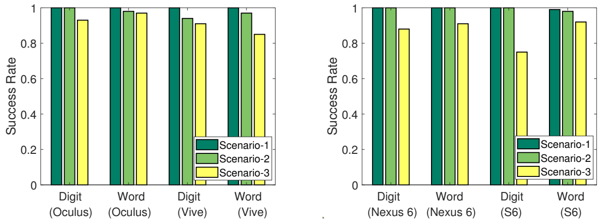
8 ATTACK EVALUATION

We validate *Face-Mic* on two standalone headsets (i.e., Oculus Quest and HTC Vive Pro) and two low-cost cardboard headsets (i.e., Cardboard headsets with Nexus 6 and Samsung Galaxy 6 smartphones).

8.1 Experimental Setup

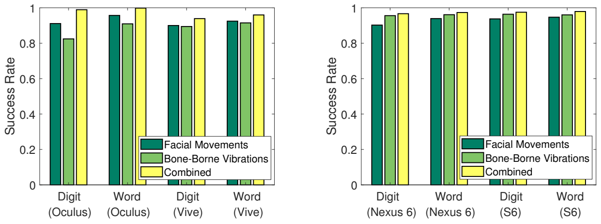
Face-mounted AR/VR headsets.

- *Operating Systems:* VR/AR headsets normally run on smartphone and computer operating systems. Particularly, HTC Vive Pro is connected to a desktop with i7-8700 CPU and GeForce RTX 2080 Graphics Card (8G), running on Windows 10. Similarly, the host computer of Oculus Quest runs on Window 10 installed with Oculus Platform. For the cardboard headsets, we use Nexus 6 and Samsung Galaxy 6 running on Android.



(a) Standalone headsets (b) Cardboard headsets

Figure 12: Performance of gender detection under *Scenario-1*, *Scenario-2*, and *Scenario-3*.



(a) Standalone devices (b) Cardboard headsets

Figure 13: Performance of speaker identification under different feature sets of facial dynamics under Scenario-1. Note that speaker identification needs the victim's identity during training, so speaker identification only applies to *Scenario-1*.

- *Motion Sensors:* HTC Vive Pro and Samsung Galaxy 6 are equipped with the same motion sensor module, i.e, Invensense MPU-6500 [14], with the sensitivities of accelerometer and gyroscope range 2048 /4096/8192/16384 LSB/g and 16.4/32.8/65.5/131 $LSB/°/sec$, respectively. Nexus 6 equips a Bosch Sensortec BMI160 motion sensor [7], with the same accelerometer/gyroscope sensitivity. Oculus Quest uses an original motion sensor board, 330-00193-03 1PASF8K, designed by Facebook, but the specifications was not published. Although the motion sensor chips support high sampling frequencies (e.g., around $8KHz$), the vendors constrain the sampling rates to ensure low power consumption, i.e., $227Hz$ for Nexus 6, $203Hz$ for Samsung Galaxy 6, and $1000Hz$ for both Oculus Quest and HTC Vive Pro. Note that we select the two low-end smartphones with the objective of demonstrating *Face-Mic*'s generalizability, since the motion sensors in most current smartphones have similar or even higher sampling rates.

Speech Datasets.

- *Digit Dataset:* Digits can be associated with a wide range of highly sensitive information (e.g., SSN, credit card number). To evaluate *Face-Mic*, we collect digit datasets consisting of 11 digits borrowed from the TIDigits corpus [31]. Besides digits 0 ∼ 9, the pronunciation "oh", a synonym of digit 0, is also collected.
- *PGP Word Dataset:* To evaluate the performance of our attack on inferring more generalized words, such as private and sensitive information in group voice chats, we apply a subset of the PGP words list [16]. Specifically, we select 20 frequently used words from the PGP word list with different length and syllables to evaluate *Face-Mic*'s generalizability to different words.

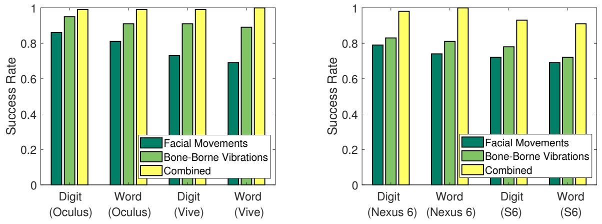
Participants & Data Collection. We collect digit/word datasets with the 4 aforementioned AR/VR headsets by involving 45 participants in total, aging from 24 to 36. Particularly, the experiments involve 15 participants for Oculus Quest (11 males and 4 females), 10 participants for HTC Vive Pro (8 males and 2 females), 10 participants for the cardboard headset with Nexus 6 (7 males and 3 females), and 20 participants for the cardboard headsets with Samsung Galaxy 6 (13 males and 7 females). Note that the 10 participants involved in the Cardboard (Nexus) dataset are also involved in the Oculus Quest dataset. Each participant is asked to wear the headset and speak the aforementioned digits and words 10 times. During the experiments, we measure the sound pressure levels (SPLs) when the participants speak using a sound level meter [35], which is placed at around 30cm to the participant's mouth, and the measured SPLs are around $67dB \sim 73dB$. The experiments are conducted in 6 different environments, including 3 university offices, 2 residential apartments, and an outdoor environment, which involve different acoustic noises (e.g., conversations, air-condition noises). We found that these acoustic noises have limited impacts on the motion sensor readings, which is consistent with the findings from prior work [1]. In all our experiments, we let participants wear headsets by themselves in a comfortable way, and we did not limit the movements of the participant. In total, we collect 4, 650 data segments for Oculus Quest, 3, 100 segments for HTC Vive Pro, 3, 100 segments for the cardboard headset with Nexus 6, and 6, 200 segments for the cardboard headset with Samsung Galaxy 6. The data collection procedures were approved by our university's IRB.

Evaluation Methodology. We examine gender detection and speaker identification by measuring attack success rate. For gender detection, the success rate is defined as the percentage of segments correctly detected as belonging to male or female, while for speaker identification, the attack success rate is the percentage of segments correctly identified as belonging to the corresponding victims. For more challenging tasks for speech recognition, we define attack success rate by employing the top-N accuracy. The top-N accuracy is defined as the probability that the actual digits/words are within the top N classes predicted by our deep learning model. We use top-1, top-2, and top-3 accuracies to quantify the attack's effectiveness.

To evaluate *Face-Mic* under *Scenario-1*, we partition a digit/word dataset (i.e., including both motion sensor data and labels) involving all users randomly into 10 subsets with equal size, with 9 subsets used for training and the remaining 1 subset for testing. To examine our attack's effectiveness under *Scenario-2*, which does not require the victim's labels for training, we take turns considering each participant as the victim, and leverage the labeled data of all remaining participants to pre-train a deep learning model. We then leverage the victim's motion sensor data to adapt the pre-trained model based on the proposed unsupervised domain adaptation method. For *Scenario-3*, we directly apply the pre-trained model on the victim's data, without applying domain adaptation.

8.2 Gender Detection

Attacks via High-end Standalone Headsets. The gender detection performance on the two high-end standalone headsets under the three attack scenarios is shown in Figure 12(a). We find that our attack has prominent performance on both headsets, with over 99% success rates under *Scenario-1*. Furthermore, we find that even



(a) Standalone headsets (b) Cardboard headsets

Figure 14: Performance of speech recognition under different feature sets of facial dynamics under *Scenario-1*.

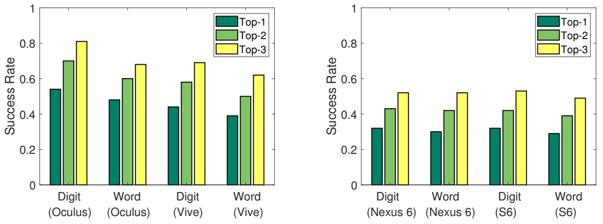
under more challenging scenarios without leveraging the victim's labels (i.e., gender) for training (i.e., *Scenario-2* and *Scenario-3*), *Face-Mic* can still achieve over 85% success rates on gender detection, which shows the effectiveness of the proposed attack.

Attacks via Low-cost Cardboard Headsets. Next, we examine the gender detection performance of *Face-Mic* on the two cardboard headsets with Nexus 6 and Galaxy S6 smartphones, with the performance of the three attack scenarios shown in Figure 12(b). We find that even with much lower sampling rates (around 200Hz), *Face-Mic* can still achieve remarkable gender detection accuracies, with over 99% attack success rates under *Scenario-1*. An encouraging finding is that the designed unsupervised domain adaptation scheme can enable high attack success rates under *Scenario-2*, with over 98% success rates, which are comparable with the results of *Scenario-1*. In addition, we observe that even without applying domain adaptation (i.e., *Scenario-3*), *Face-Mic* can still achieve over 75% gender detection performance. The results demonstrate the effectiveness of *Face-Mic* on detecting the gender of cardboard headset wearers.

8.3 Speaker Identification

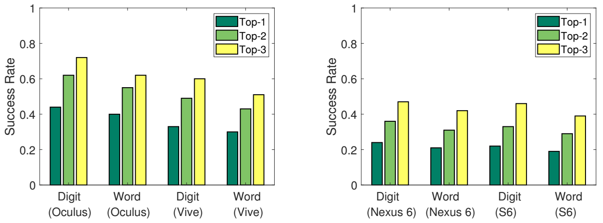
Attacks via High-end Standalone Headsets. We then evaluate *Face-Mic* on identifying speakers with Oculus Quest and HTC Vive Pro, with the results shown in Figure 13(a). We find that when leveraging both facial muscle movements (i.e., Combined), *Face-Mic* can achieve over 97% and 93% attack success rates on Oculus Quest and HTC Vive Pro, respectively. For both headsets, the attack success rates on the word datasets are generally higher than the success rates on the digit datasets. We believe this is due to the longer length and richer syllables of words, which encodes more biometric characteristics.

Attacks via Low-cost Cardboard Headsets. The speaker identification performance of *Face-Mic* on the two cardboard headsets with Nexus 6 and Galaxy S6 smartphones is shown in Figure 13(b). Although the motion sensors' sampling rates are much lower in the cardboard headsets (around 200Hz), *Face-Mic* can still achieve over 94% success rates for both headsets when using both facial muscle movements and bone-borne vibrations (i.e., Combined), which demonstrates the effectiveness of *Face-Mic* with low-cost cardboard headsets. Furthermore, compared to the high-end standalone headsets, we find that the cardboard headsets have higher success rates on bone-borne vibrations. Such results can be attributed to the lower vibration damping ratio of the cardboard headsets, as the stand-alone headsets are normally equipped with thick face covers pad for comfort, which greatly attenuate the bone-borne vibrations.



(a) Standalone headsets (b) Cardboard headsets

Figure 15: Performance of speech recognition with top-1, top-2 and top-3 accuracies under *Scenario-2*.



(a) Standalone headsets (b) Cardboard headsets

Figure 16: Performance of speech recognition with top-1, top-2 and top-3 accuracies under *Scenario-3*.

8.4 Speech Recognition

Performance under *Scenario-1*. We then evaluate *Face-Mic* on deriving simple speech content, including digits and PGP words. The attack success rates of Oculus Quest and HTC Vive Pro under *Scenario-1* are shown in Figure 14(a). It is encouraging that for both Oculus Quest and HTC Vive Pro, *Face-Mic* achieves over 99% top-1 accuracies when using both facial muscle movements and bone-borne vibrations (i.e., *Combined*). We also separately examine the impact of feature sets (i.e., facial muscle movements and bone-borne vibrations) on the attack, and the average top-1 accuracies are 89% and 96%, respectively. The bone-borne vibrations are directly generated by vocal folds, and thus they are encoded with rich phoneme characteristics, leading to higher speech recognition performance. We also evaluate the speech recognition performance under *Scenario-1* leveraging the two low-cost cardboard headsets, with the results shown in Figure 14(b). We find that even with much lower sampling rates, *Face-Mic* can achieve over 93% top-1 accuracies for speech recognition when using both feature sets. The results demonstrate the effectiveness of *Face-Mic* under *Scenario-1*.

Performance under *Scenario-2*. The attack performance of standalone headsets under *Scenario-2* is shown in Figure 15(a). We find that even the victim's labels (i.e., audio data) are not available, *Face-Mic* achieves on Oculus Quest with (54%, 48%) top-1, (70%, 60%) top-2 and (81%, 68%) top-3 accuracies on digit recognition and word recognition, respectively, on Oculus Quest. As a comparison, the random guess probabilities of recognizing these digits and words are only 9.1% and 5%, respectively. For HTC Vive Pro, the attack performance are (44%, 39%) top-1, (58%, 50%) top-2 and (69%, 62%) top-3 accuracies. These results indicate that the attack performance is comparable when using the standalone headsets under the adaptive retraining. As shown in Figure15(b), for the same attack scenario, both cardboard headsets achieve around (32%,30%) top-1, (43%,42%) top-2 and (53%,49%) top-3 accuracy on digit recognition and word recognition, respectively. The top-3 success rates of both headsets reach 50%, which can allow the adversary to extract some speech content. Although the accuracies are lower on the cardboard headsets, the digit and word recognition accuracies are still 2× and 3× over the corresponding random guess accuracies in top-1 and 5× and 9× over the random guess in top-3.

Performance under *Scenario-3*. For the most challenging scenario, as shown in Figure 16(a), *Face-Mic* can still obtain (44%,40%) top-1, (62%,55%) top-2 and (72%,62%) top-3 accuracies for digit and word recognition on Oculus Quest. The results are 4× (top-1) to 8× (top-3) and 7× (top-1) to 12× (top-3) over the random guess (i.e., 9.1% for digit recognition and 5% for word recognition).These results show that *Face-Mic* has the capability to extract speech information even without any data from the victim. We find HTC Vive Pro achieves a slightly lower performance, (33%,30%) top-1, (49%,43%) top-2 and (60%,51%) top-3 for digit and word recognition. Under the same scenario, as shown in Figure 16(a), the results of the two cardboard headsets are lower with (24%,21%) top-1, (36%,31%) top-2 and (46%,39%) top-3 accuracies for digit and word recognition.

8.5 Impacts of Body Movements

To evaluate the robustness of *Face-Mic* under motion interference, we conduct a case study by asking 5 participants (3 males and 2 females) to wear Oculus Quest and Google Cardboard with Nexus 6 smartphone, and conduct large-scale body movements while speaking digits (i.e., "Oh", 0 ∼ 9). Particularly, we collect digit datasets for two representative movements in real-world AR/VR scenarios: *1) Head Movements:* the participant moves his/her head upward or downward when speaking each digit, and *2) Body Movements:* the participant moves one step forward and one step backward when speaking each digit. We do not constrain the postures of the participants during the experiments. Table 2 shows the attack success rates on detecting gender, identifying users, and recognizing digits. We find that *Face-Mic* is much more effective in eavesdropping the sensitive information when using the proposed body motion artifact removal (BMAR) approach (introduced in Section 5.2). Specifically, given the large-scale body movements, BMAR improves the success rates for user identification and digit recognition by 8.3% and 13.3%, respectively. Even higher improvements can be observed for the cardboard headset, with 25.4% and 34.9% higher success rates for identifying users and recognizing digits. The gender detection model is less susceptible to body movements. In general, the BMAR approach can greatly improve the eavesdropping performance of *Face-Mic*. Such improvements are brought by the well-designed BMAR model that mitigates the impacts of motion artifacts.

8.6 Impact of Training Data Size

We consider the affect of training size deriving the headset wearer's gender, identity, and simple speech information, taking a least knowledge approach to evaluate attack success (using 1 ∼ 10 samples for training/testing) under *Scenario-1*. Figure 17 shows a line graph comparing the attack success rates of four available devices as the training size is varied. The x-axis represents the number of samples that is used for training. For both low-cost headsets, gender and speaker identification accuracy can reach 92% with only two training samples. For Samsung Galaxy S6, *Face-Mic* achieves almost 100% gender detection accuracy. For Oculus Quest and HTC Vive Pro, gender detection and speaker identification accuracies are over 96% using five training samples. For digits and word recognition, the success rates reach 83% with six samples for training. Overall, our results demonstrate the low training requirement of *Face-Mic*.

Table 2: Comparing the performance of *Face-Mic* with and without body motion artifact removal (BMAR).

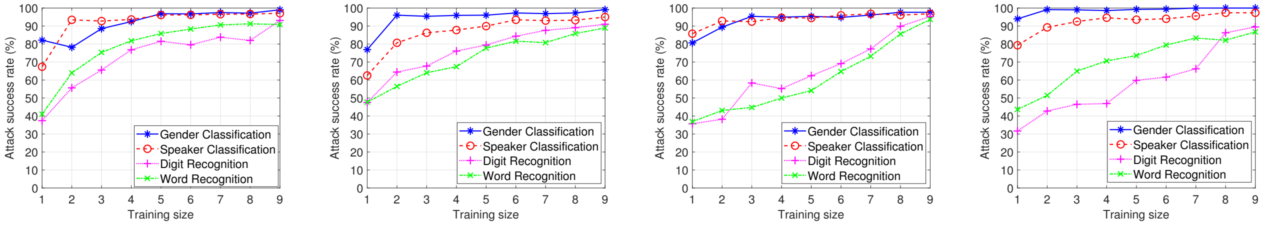
Gender Detection				
	Head Movements		Body Movements	
	Without BMAR	With BMAR	Without BMAR	With BMAR
Google Cardboard (Nexus 6)	89.37%	93.23%	85.13%	93.47%
Oculus Quest	96.81%	98.87%	96.24%	98.56%
Speaker Classification				
	Head Movements		Body Movements	
	Without BMAR	With BMAR	Without BMAR	With BMAR
Google Cardboard (Nexus 6)	75.74%	89.74%	63.47%	88.91%
Oculus Quest	90.12%	96.75%	88.87%	97.21%
Speech Recognition				
	Head Movements		Body Movements	
	Without BMAR	With BMAR	Without BMAR	With BMAR
Google Cardboard (Nexus 6)	37.49%	56.25%	18.72%	53.71%
Oculus Quest	79.37%	89.99%	74.12%	87.49%

9 RELATED WORK

AR/VR Headset Security: AR/VR headsets and their potential security vulnerabilities are gaining attention in academic research. Roesner et al. [32] identified different security and privacy challenges of current AR/VR technology and systematizes their knowledge across two factors: system scope and functionality. The authors described that the privacy risk is much greater in AR/VR devices, compared to conventional systems, because of specific AR/VR behaviors, such as continuous sensor monitoring (i.e., accelerometer, GPS, etc.) and unrestricted sensor access. Initial research in AR/VR privacy focused on user input security [15, 18, 33]. Specifically, they looked at the challenges that can arise from unrestricted sensor access (e.g., microphone, video, MEMS) and how it can be compromised to infer user keystrokes. Another work explored the potential risks of unregulated AR/VR visual output [17]. And a case study by Chen et al. [10] revealed potential security vulnerabilities for *face-mounted* VR devices, which is a key interest of our own work.

Motion Sensor Based Speech Eavesdropping: Academic research has been committed to MEMS motion sensor eavesdropping attacks on user speech. In AccelEve [4] and Gyrophone [22] the authors demonstrated attacks that use smartphone accelerometer and gyroscope data to compromise speech privacy. Similarly, Accelword [41] is a benign application designed to recognize the airborne speech of a user. In a broader study, Anand et al. [1] analyzed the speech privacy threat from MEMS motion sensors and found parameters such as same surface propagation medium (observed in [22]) and loud volume speech (observed in [41]) may be required to successfully compromise speech. In our work, we capture the *facial dynamics* of the speaker (combination of bone-borne vibrations and facial muscle movements), introducing a novel feature source for speech classification that has yet to be explored in academia.

Bone-Borne Vibrations and Defensive Applications: Recent work has emerged that explores *bone-borne vibrations* for speech recognition. An initial study by [20] used high-fidelity piezoelectrical disc mounted on the nose pad of a pair of glasses to capture vibrations and reconstruct speech. In a study by Zhang et al., the authors explored the use of bone-borne vibrations for authentication/defense purposes and presented a continuous liveness detection system called VibLive [42]. In our work, we explore the unique scenario when someone uses an AR/VR headset, and the headset



(a) Oculus Quest (b) HTC Vive Pro (c) Google Cardboard (Nexus 6) (d) Google Cardboard (Galaxy S6)

Figure 17: Performance of *Face-Mic* under different sizes of training data.

itself allows for direct vibration propagation between the user's face and the *inbuilt*, smartphone motion sensors; capturing both *bone-borne vibrations* and *facial muscle movements*.

10 DISCUSSION

10.1 Potential Defenses

An intuitive way to defend Face-Mic is to disable the background use of the motion sensors. However, motion sensors need to be used in almost all AR/VR apps to sense head movements. And even if an app requests the background permission of motion sensor during installation, most users will agree on the permission Moreover, it is also possible to perturb the correlation between the motion sensor data and the speech-associated facial dynamics to defend against this type of attack. This can be achieved through programmatically injecting noises to the motion sensor readings without interfering with AR/VR motion tracking. Along this direction, the privacy risks introduced by *Face-Mic* may be mitigated through designing sensory noises to obfuscate the reconstruction of facial movements and bone-borne vibrations. Alternatively, the manufacturers of VR headsets may add ductile materials in the foam replacement cover and the headband, which may attenuate the facial vibrations that would be captured by the built-in accelerometer/gyroscope. Compared to the Oculus Quest with a thin face cover, the HTC Vive is equipped with a thick mask exhibiting much lower success rates for eavesdropping digits and PGP words. Motivated by this observation, we suggest the vendor of AR/VR headsets to add some ductile materials between the headset and the user's face to weaken the facial vibrations. Another effective defense is to constrain the sampling rate of the accelerometer and gyroscope in AR/VR Operating Systems (e.g., Windows for HTC Vive). However, limiting the sampling rate may also influence the functionality and usability of some benign apps. In addition, besides high-frequency bone-brone vibrations, Face-Mic also leverages facial muscle movements reside at low-frequency ranges (e.g., below 100Hz), making it still capable of deriving some sensitive information.

10.2 Potential Attack Improvement

As the first work in this line of research, we demonstrate that *Face-Mic* can classify speech content, including 11 digits and 20 PGP words with top-1 recognition accuracies of around 99%, 54%, and 44% under *Scenario-1*, *Scenario-2*, and *Scenario-3*, respectively, on the currently most popular commercial VR headset Oculus Quest. Such a capability already allows an adversary to derive a broad range of sensitive information, such as phone numbers, social security numbers, and passwords. We believe that if the dataset can be sufficiently extended (e.g., by involving data of more people), our attack performance will be further improved. Another challenge is that in some AR/VR scenarios (e.g., AR/VR virtual meeting), the victim's speech may involve a much larger set of words, making it difficult for a pre-trained model built on a limited vocabulary to reconstruct all speech content. A possible solution is to train a general deep learning model leveraging a huge dataset with an extensive vocabulary, but such a method will require a significant amount of manpower, making it infeasible in practice. Since both bone-borne vibrations and human speech are produced by vocal folds, it is possible to develop a speech reconstruction scheme that can map the motion sensor readings with bone-borne vibrations into audio signals, which resemble the microphone recordings of human speech. Such a mapping can be realized with a deep autoencoder network that consists of an encoder to convert bone-borne vibrations into hidden representations and a decoder to transform the representations into audio signals. Furthermore, we may use more sophisticated network architectures (e.g., WaveNet [30], Transformer [38]) to improve the speech reconstruction performance. In this way, we can directly apply an existing audio-based speech recognition model (e.g., Google Speech-to-Text) to infer the speech content of the victim. We leave this as our future work.

11 CONCLUSION

In this paper, we propose *Face-Mic*, a devastating attack on AR/VR devices that leverages the facial dynamics captured by zero-permission motion sensors to infer private speech and speaker information. We determined two types of facial vibrations that capture the speech effects, namely speech-associated facial movements and bone-borne vibrations. To render a practical attack, we developed a novel signal source separation technique based on deep regression to eliminate the impacts of human body movements. Based on the unique characteristics of facial movements and bone-borne vibrations, we extract two sets of features that capture unique private biometrics and speech characteristics. A deep-learning-based framework is developed leveraging the extracted features to derive the headset owner's gender/identity and possibly recover speech content. We validate *Face-Mic* via extensive experiments and demonstrate its generalizability and effectiveness. We believe that *Face-Mic* demonstrates a real threat to the users of AR/VR devices, which calls for additional research to develop defensive solutions in the future.

12 ACKNOWLEDGMENT

This work was partially supported by the National Science Foundation Grants CCF1909963, CCF2028876, CNS1526524, CNS1714807, CNS1820624, CNS2030501, CNS2120396, OAC1547350, and the Grant of National Natural Science Foundation of China No.62172277.

REFERENCES

[1] 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.

[2] Android. 2020. MediaRecorder overview. https://developer.android.com/guide/topics/media/mediarecorder.

[3] Barry Arons. 1992. A review of the cocktail party effect. *Journal of the American Voice I/O Society* 12, 7 (1992), 35–50.

[4] 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.

[5] BFW. 2020. VR in Advertising: Examples and Predictions for the Next Decade. https://www.gobfw.com/vr-advertising/vr-in-advertising-examples-predictions-for-2020s/.

[6] Bootcamp. 2020. Case study: Navigating shopping malls with augmented reality. https://medium.com/design-bootcamp/navigation-in-shopping-malls-through-augmented-reality-d8194f1a7a23.

[7] BOSCH. 2020. IMU: BMI160. https://www.bosch-sensortec.com/products/motion-sensors/imus/bmi160.html.

[8] Carlijn VC Bouten, Karel TM Koekkoek, Maarten Verduin, Rens Kodde, and Jan D Janssen. 1997. A triaxial accelerometer and portable data processing unit for the assessment of daily physical activity. *IEEE transactions on biomedical engineering* 44, 3 (1997), 136–147.

[9] The Financial Brand. 2017. 10 Ways Banks And Credit Unions Are Using Virtual Reality. https://thefinancialbrand.com/68593/banks-credit-unions-finances-virtual-reality/.

[10] S. Chen, Zupei Li, F. Dangelo, C. Gao, and X. Fu. 2018. A Case Study of Security and Privacy Threats from Augmented Reality (AR). *2018 International Conference on Computing, Networking and Communications (IEEE ICNC)* (2018), 442–446.

[11] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* 17, 1 (2016), 2096–2030.

[12] Google. 2020. Google Cardboard. https://arvr.google.com/cardboard/.

[13] Paula Henry and Tomasz R Letowski. 2007. *Bone conduction: Anatomy, physiology, and communication*. Technical Report. Army Research Lab Aberdeen Proving Ground Human Research and Engineering.

[14] InvenSense. 2020. MPU-6500 Six-Axis (Gyro + Accelerometer) MEMS MotionTracking Devices. https://invensense.tdk.com/products/motion-tracking/6-axis/mpu-6500/.

[15] Suman Jana, David Molnar, Alexander Moshchuk, Alan Dunn, Benjamin Livshits, Helen J Wang, and Eyal Ofek. 2013. Enabling fine-grained permissions for augmented reality applications with recognizers. In *Proceedings of USENIX Security Symposium*. 415–430.

[16] Patrick Juola and Philip Zimmermann. 1996. Whole-Word Phonetic Distances and the PGPfone Alphabet. In *The International Conference of Spoken Language Processing(ICSLP)*. 98–101.

[17] Kiron Lebeck, Tadayoshi Kohno, and Franziska Roesner. 2016. How to Safely Augment Reality: Challenges and Directions. *Workshop on Mobile Computing Systems and Applications (ACM HotMobile)*.

[18] Zhen Ling, Zupei Li, Chen Chen, Junzhou Luo, W. Yu, and X. Fu. 2019. I Know What You Enter on Gear VR. *IEEE Conference on Communications and Network Security (IEEE CNS)* (2019), 241–249.

[19] Google LLC. 2020. YouTube VR. https://play.google.com/store/apps/details?id=com.google.android.apps.youtube.vr&hl=en_US.

[20] Héctor A Cordourier Maruri, Paulo Lopez-Meyer, Jonathan Huang, Willem Marco Beltman, Lama Nachman, and Hong Lu. 2018. V-Speech: Noise-Robust Speech Capturing Glasses Using Vibration Sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–23.

[21] Maranda McBride, Phuong Tran, and Tomasz Letowski. 2008. Head mapping: Search for an optimum bone microphone placement. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 52. SAGE Publications Sage CA: Los Angeles, CA, 503–507.

[22] Yan Michalevsky, Dan Boneh, and Gabi Nakibly. 2014. Gyrophone: Recognizing Speech from Gyroscope Signals. In *Proceedings of USENIX Security Symposium*. 1053–1067.

[23] Microsoft. 2020. Use Speech in Windows Mixed Reality. https://support.microsoft.com/en-us/windows/use-speech-in-windows-mixed-reality-af24e0a9-7e17-b542-3720-203e278e588e.

[24] Brian B Monson, Eric J Hunter, Andrew J Lotto, and Brad H Story. 2014. The perceptual significance of high-frequency energy in the human voice. *Frontiers in Psychology* 5 (2014), 587.

[25] Nick Nikiforakis, Alexandros Kapravelos, Wouter Joosen, Christopher Kruegel, Frank Piessens, and Giovanni Vigna. 2013. Cookieless monster: Exploring the ecosystem of web-based device fingerprinting. In *IEEE Symposium on Security and Privacy (IEEE S&P)*. 541–555.

[26] Obsess. 2020. Virtual Reality Shopping Platform. https://obsessar.com/virtual-reality-shopping/.

[27] Oculus. 2020. Oculus PC SDK v23. https://developer.oculus.com/downloads/package/oculus-sdk-for-windows/.

[28] Oculus. 2020. Oculus Privacy Policy. https://www.oculus.com/legal/privacy-policy-for-oculus-account-users/.

[29] Oculus. 2020. VrApi. https://uploadvr.com/oculus-go-rooms-alternatives/.

[30] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).

[31] George Doddington R. Gary Leonard. 1993. TIDIGITS LDC93S10. Web Download. Philadelphia: Linguistic Data Consortium. https://catalog.ldc.upenn.edu/LDC93S10.

[32] Franziska Roesner, Tadayoshi Kohno, and David Molnar. 2014. Security and Privacy for Augmented Reality Systems. *Commun. ACM* (2014), 88–96.

[33] Franziska Roesner, David Molnar, Alexander Moshchuk, Tadayoshi Kohno, and Helen J Wang. 2014. World-driven access control for continuous sensing. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (ACM CCS)*. 1169–1181.

[34] 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.

[35] SkyPaw. 2021. Decibel X: dB Sound Level Meter. https://apps.apple.com/us/app/decibel-x-db-sound-level-meter/id448155923.

[36] Statista. 2020. Immersive technology consumer market revenue worldwide from 2018 to 2023. https://www.statista.com/statistics/936078/worldwide-consumer-immersive-technology-market-revenue/.

[37] STEAMWORKS. 2020. OpenVR. https://partner.steamgames.com/doc/features/steamvr/openvr.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*. 6000–6010.

[39] Chen Wang, S Abhishek Anand, Jian Liu, Payton Walker, Yingying Chen, and Nitesh Saxena. 2019. Defeating hidden audio channel attacks on voice assistants via audio-induced surface vibrations. In *Annual Computer Security Applications Conference (ACSAC)*. 42–56.

[40] Wikipedia. 2020. Virtual reality games. https://en.wikipedia.org/wiki/Category:Virtual_reality_games.

[41] 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.

[42] Linghan Zhang, S. Tan, Z. Wang, Yili Ren, and J. Yang. 2020. VibLive: A Continuous Liveness Detection for Secure Voice User Interface in IoT Environment. *Proceedings of Annual Computer Security Applications Conference (ACSAC)* (2020).

[43] Han Zhao, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira, and Geoffrey J Gordon. 2018. Adversarial multiple source domain adaptation. *Advances in Neural Information Processing Systems (NeurIPS)* 31 (2018), 8559–8570.