Infiltrating the Metaverse: A Security Assessment of Multifaceted Voice Command Manipulation and Undermine*

Arpita Dinesh Sarang, Sang-Hoon Choi, and Ki-Woong Park†

Sejong University, Seoul 05006, South Korea
arpita.sarang@sju.ac.kr, csh0052@gmail.com, woongbak@sejong.ac.kr

Abstract

In the Metaverse, our virtual reality, new technologies are advancing every day. Combining these new technologies with the current systems compromises the security of the Metaverse modules. As a result, this gives attackers access to new potential attack vectors. Attacks against the Metaverse are inevitable due to the lack of thorough investigation and preparation for the mitigation of these potential target vectors. Voice command implementation is one of the significant threats to Metaverse users. Voice commands have the ability to retrieve, traverse, and interpret commands within a user's Metaverse environment, which contains vital user data. Misinterpreting and manipulating these vocal commands could negatively impact the user experience in the virtual world, potentially exposing user data to risk. In the future, the Metaverse will incorporate numerous voice-controlled applications and technologies. In order to investigate the impact of voice command manipulation and misinterpretation by attackers, we examine the voice command interpretation flow in Oculus Quest 2 utilizing Facebook Meta. We perform security analysis on two potential attack surfaces that have been discovered. We use Facebook Meta to analyze the voice command interpretation flow in Oculus Quest 2 in order to look at the effects of voice command manipulation and misinterpretation by attackers. We perform security analysis on the two identified potential attack surfaces. Finally, we arrive at an assessment of new potential attack vectors, and we extensively exploit them using external physical and internal noise.

Keywords: Security Analysis, Speech Recognition, Natural Language Processing, Metaverse, Cyber Security

1 Introduction

The world that people create to be more cozy, peaceful, and secure is known as the Metaverse or Virtual Reality(VR). In these virtual worlds, they engage in a variety of activities such as work, education, gaming, travel, and experiencing things that are not possible to obtain in the real world. Head Mounted Displays (HMD) and Handheld Augmented Reality (HAR) devices can be used to access these Metaverse or simulations of virtual reality [8]. Users of the Metaverse navigate the virtual environment via an avatar that is linked to their identity and data. They utilize handheld controllers, voice commands, and gestures to input commands in the Metaverse. Voice Commands among which have been determined to be the fastest way to navigate and communicate within the Metaverse. The microphones mounted on the devices

*Proceedings of the 8th International Conference on Mobile Internet Security (MobiSec'24), Article No. 79, December 17-19, 2024, Sapporo, Japan. © The copyright of this paper remains with the author(s).

†Corresponding author.

used for accessing the Metaverse detect them as inputs. In order to process and enable it to comprehend and execute the command in the Metaverse, the speech recognition and intent recognition modules in the voice command processing applications installed in the Metaverse environment are activated. The voice command processing applications require access to system privileges, other applications, user data associated with the avatar and Metaverse, and services that are operational in the Metaverse to interpret these commands smoothly. This improves accessibility, provides a more customized experience, enables maximal involvement regardless of user background, automates repetitive tasks, and fosters a deeper connection with the Metaverse environment for users.

Although there are advantages to employing this technology, there are also drawbacks, such as security threats, dependency on other variables, privacy concerns, accuracy issues, and functional restrictions based on different components of Metaverse infrastructure [11]. As a result of these voice-related privacy problems, the manufacturers of AR/VR headsets have restrictive voice access standards and require explicit consent before using microphones [12]. As a result, this paper outlines the meticulously examined voice command workflow in the VR device, as well as the Metaverse used for voice command processing and implementation. Using this workflow, we attempt to plot the known attack vectors for voice commands. Using threat analysis, we discover additional attack vectors in the same workflow. By consuming voice commands, we address the impact and potential threats to the Metaverse.

The remainder of this paper is as follows: Section 2, provides insight on speech enhancement to improve speech recognition. Section 3, describes Traditional Voice command-based attacks and their consequences. Section 4, conveys the Security Analysis based on voice commands in Metaverse. Section 5, concludes our learnings and future works.

2 Related Work

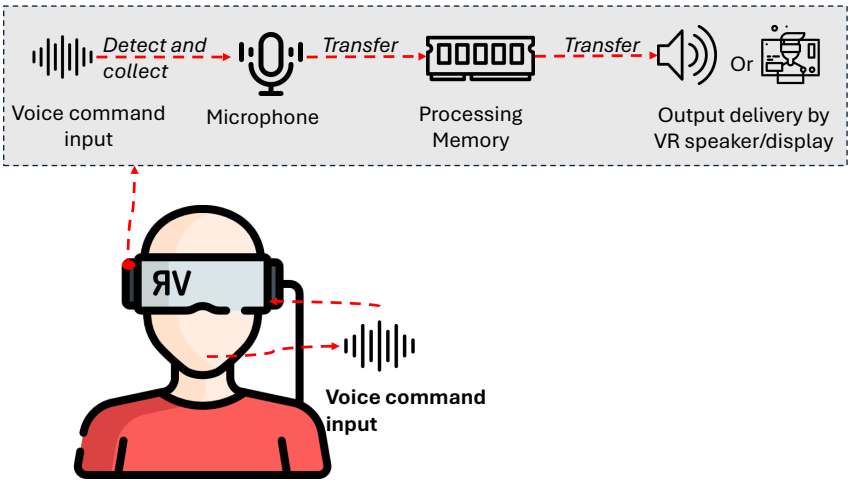

Figure 1: Hardware Voice command workflow

A multitude of parameters, some user-related and some hardware- and software-related, are necessary for high accuracy of voice command input to any device. Such as the pronunciation, the absence of distracting outside sounds, a superior sound card, a high-quality microphone,

and the required processing power [7]. This facilitates more effective voice input processing by the voice processing software. As an outcome, the way these parameters were incorporated for the VR headset devices advanced. For Microphones, Microelectromechanical systems (MEMS) have become more popular than traditional electret condenser microphones (ECM) in device hardware because of their superior electro-acoustical performance at smaller sizes and greater stability against environmental factors like temperature, humidity, and vibration [5]. When two microphones were combined, the primary microphone gathered the bearing acoustic signal, while the reference microphone recorded noise. For their effective functioning, their distinct relative spatial positions that adhere to the same Doppler distortion laws need to be considered [10].

To improve voice command detection, the focus of recent studies has switched from basic word recognition to lip-reading sentences or facial expression reading in surroundings. In order to explore an alternative schema and improve voice processing performance, phonemes as a categorization schema are utilized for lip-reading texts [3]. Speech enhancement is the technique of raising voice quality by reducing background noise consumed by Artificial Intelligence(AI) models to train. A review highlighted the shortcomings of AI algorithms for Automated Speech Recognition (ASR), stating that they are still incompetent to process audio or voice inputs with twisted accents, misinterpret children's speech, ignore different speaking styles and need to be developed more inclusive when it comes to human languages [1]. However, there are also methods like SEADNet which combines adversarial defenses with traditional speech enhancement at the model level to make traditional speech enhancement better [9]. These techniques need to be developed compatible with the traditional ASR since these methods are not yet fully matured. As a result, robust hardware components and speech recognition software are required for the HMD to recognize and process voice commands. In Figure 1, we depict the hardware voice commands workflow. The microphones in this diagram capture the voice command, which the speech recognition software then loads into its memory and processes. The software will execute the output command after receiving it.

Voice Command recognition software, which is installed on HMD, is regarded as a voice assistance tool that speeds up recurrent activities and navigation. In particular, a VR voice assistant is a luxury for its users. The development of these tools has not yet fully optimized them for the VR voice command experience, ensuring accuracy and security for consumers. There are distinct methods used to introduce voice assistance into the HMD. These include adapting the manufacturer's own application specifically designed for the HMD, utilizing a third-party application to process voice commands, or using a tool that is partially made by the manufacturer but uses third-party services to improve compatibility. Our research examines the voice command processing in the presence of noise for these applications manufactured for VR.

3 Voice Command Tool Workflow

Voice commands have made recurring tasks and navigation more convenient. In order to process more rapidly and accurately, the producers devise innovative strategies. These strategies include a third-party program to handle voice commands, utilising the manufacturer's own application, especially for the HMD, or employing a tool that is partially manufactured by the manufacturer but applies third-party services to increase compatibility.As a result, the use of voice commands has become increasingly adaptable. This is owing to Voice Command-based VR applications now having readily available web components, the novel approach. We can comprehend the primary components utilized in this application by having an explicit workflow outlined. Voice

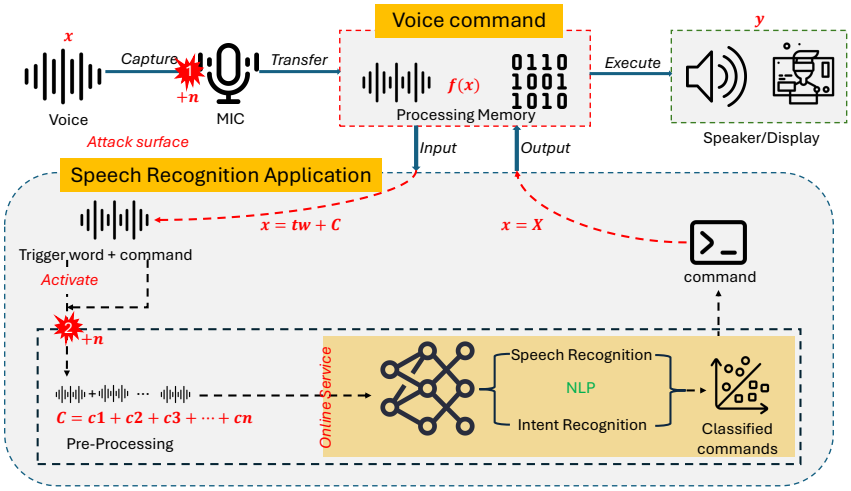

Figure 2: Voice command Processing in Head Mounted Display(HMD)

commands enable access to the VR environment that contains both sensitive and non-sensitive data. Security is a major concern. We could identify the attack surface in this infrastructure through a workflow for voice command-based programs.

Although these apps still lack accuracy, compatibility, and security, they are partially available to the VR user community. we utilised the existing Voice recognition program available on Oculus Quest 2 in order to determine the workflow for voice command-based application for VR. Figure 2, we draw and discuss in detail the workflow for voice command controls for Metaverse using HMD. We go into great detail about the application's voice command controls and pin-point the identified two attacks surfaces for virtual reality. In this workflow, the voice command(x), which is detected by the microphones mounted on the HMD, serves as the starting point for the infrastructure. This voice command($f(x)$) is recorded and then transferred to the processing memory to be processed further. When the voice's first word is identified as the default trigger word(tw), the Voice command is recorded. The trigger word and the command(C) are the two elements that make up the voice command. The Speech Recognition Application then preprocesses the command into time interval-wise chunks($C= c1+c2+c3+...+cn$). The voice command segments undergo noise reduction before being transferred to the appropriate NLP on remote server voice processing AI module. Based on the pre-trained data, this module performs intent and speech recognition and issues a classified command. The application receives the classified command, converts it to a binary command(X), and then enables the VR application or environment to execute it. The VR speaker or display provides confirmation of the command execution(y) if required.

We further investigate this workflow by inducing noise physically and internally, assuming that this is the direction voice commands control-based applications would take in the future. We accomplish this out of concern for user data privacy in Metaverse.

4 Security Analysis for Voice Commands in Metaverse

The user's experience in the metaverse can be seriously threatened by the execution of incorrect or not user-verified commands, which can also result in unauthorized access or the publication of

sensitive data. Given that the voice command can access and delete files from storage, conduct payments online, read sensitive data, and modify user and device configurations, we endeavored to conduct an investigation into the security of this infrastructure using the comprehensive workflow for the voice command processing application in Figure 2. Since the NLP framework is an external component of this infrastructure, voice command-based application utilization is possible when it is connected to the internet. The user's Metaverse environment is at risk since processing each command requires sending it over the network and back within the HMD device. As a result, we identify two attack surfaces in Voice Command processing in the metaverse inducing noise, which can be abused by executing different attack vectors.

4.1 Attack Surface (1)

The Attack Surface (1) is represented by the highlighted number 1 in Figure 2. During the process of voice command capture via microphone, noise(n) is physically induced on the attack surface. The Speech recognition application on the HMD's processing memory retrieves the digital voice commands(x+n) that were initially detected by the microphones as analogue voice commands.Traditional voice command exploitation, such as the 2017 Dolphin attack [18], adverserial audio attacks [17], voice spoofing [4], command injection ([13], [16], [15]), and physical proximity attacks ([6], [2], are possible because this infrastructure lacks reliable authorised user voice recognition that has to be mandated.

We conducted an experiment where we used a voice command HMD while inducing noise. For instance, when we attempted to say "Open facebook.com" with a voice command, the speech recognition software converted the audio to text as "Open facebook dot com," processed it, and then launched facebook.com in the browser. In order for the VR to recognize the voice more precisely, we repeated the exercise by inducing a noise that was identified when the speaker paused between sentences. Therefore, "Open Instagram, Facebook, dot com" was the command input. The remaining words were spoken by a female speaker, while the word "Instagram" was spoken by a male voice that was captured from external physical noise. In Figure 3, shows the outcome of our experiment that assessed the speaker's maximum pause time between words and the potential attack within a particular range of distance from the microphones. Both a male and a female voice were recorded as the speaker's voice at high and low frequencies. This concluded that the Attack Surface(1) is likely to get attacked when the distance is shorter and the speech frequency is high. Also, the pause time between words should be in between 0 to 2 seconds, irrespective of the speaker's gender. As a result, before being sent to the NLP module, the voice command was compromised and opened unexpected websites, applications, and access to HMD settings.

4.2 Attack Surface (2)

When other applications with hidden spy modules overhear the digital signals inserted into processing memory and influence user decisions by displaying contents or may abuse user information for other attacks.

The Attack Surface (2) is indicated in Figure 2 by the highlighted number 2. It indicates the moment the speech recognition application receives a voice command. Pre-installing a spy program on the HMD enables assessing this surface by listening to voice commands and inducing noises while it is being identified and recorded. This attack surface can be demonstrated by utilizing the attack strategy in [19]. They presented the side-channel attack through a spy application to read voice commands with an accuracy of more than 90%. With this possibility, we assume that such spy applications can induce noise(n) in voice commands. This leads to

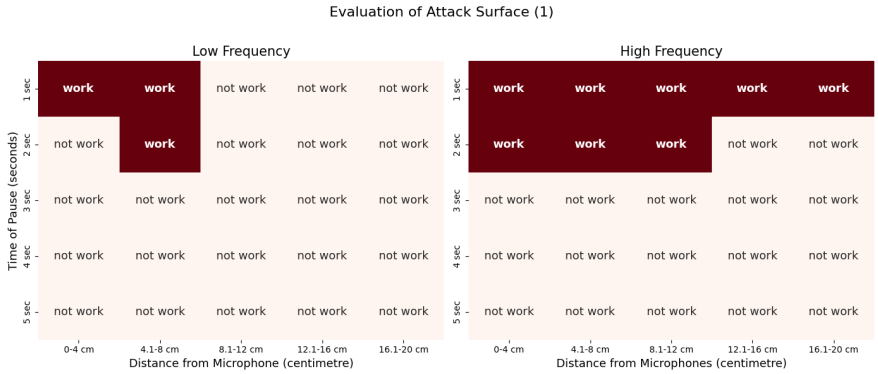

Figure 3: Potential Range of Attack for Attack Surface (1)

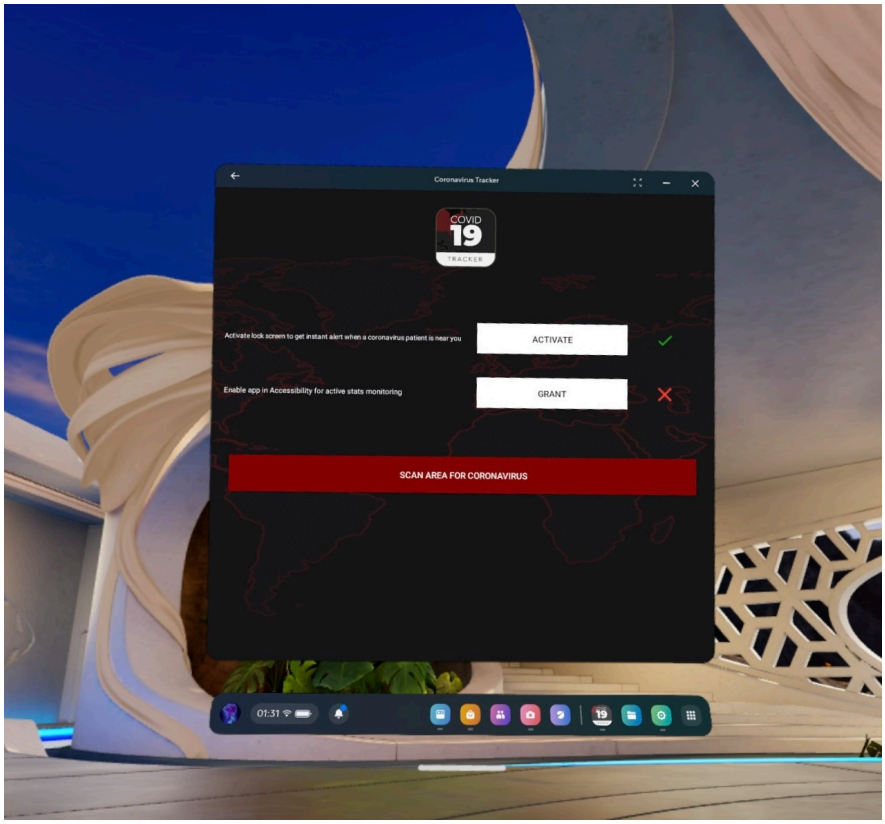

Figure 4: Potential attack through Attack Surface (2)

inducing noise internally when loaded and recorded in the storage before being processed by the speech recognition application. Therefore, to verify the possibility of sidechannel attack via installing a spy application on the OVR, we tried to install it. Based on the recent ransomware attack on Apple Vision Pro, we attempted to installed a Android Operating System(OS) based

ransomware onto our OVR [14]. Figure 4, states that the Coronavirus Tracker Ransomware APK(Android Application Package) successfully infiltrated into our OVR though all security setting were enabled and running. As a result, sensitive or malicious commands can be injected into voice commands and executed if there is a lack of rigorous scrutiny.

Attackers have an exceptional probability of taking advantage of these two attack surfaces. Attack surface (1)'s and attack surface (2)'s input are not validated prior to execution in the metaverse. Sensitive data and the Metaverse user experience are at stake. Inadequate verification may result in unapproved purchases, data breaches, and disruptions to services. The voice command processing's vulnerability to attack could cause financial loss, psychological anguish, and damage to the reputation of the VR device's manufacturer.

The developer of the voice command recognition application might employ a few techniques to lessen the impact of the aforementioned attack surface. In order for the ASR to accept only authenticated user voices, the developer should think about incorporating a user voice authentication module. The application must employ audio filtering and be able to differentiate between male and female voices. It should have contextual awareness for the command sequence. Blocking sensitive voice commands, restricting the scope of command could make the application less usable. Incorporating web-based NLP is revolutionizing voice command recognition applications. However, manufacturing their own NLP and managing the NLP data via secure network channels will be more secure. This lowers the likelihood of an attack via these attack surfaces.

5 Conclusion

This research elucidates the processing of voice commands within the metaverse. We delineated the workflow for processing voice commands utilizing speech recognition applicatins that employ the online NLP module. This strategy entails the future implementation of voice command processing in the Metaverse and the development of reliable voice command processing applications. In the workflow, we analyze the two potential attack surfaces and the methods by which attackers may exploit them using external physical noise and internal noise. This emphasized the cyber security gaps examined for voice commands in virtual reality. This emphasized the cyber security gaps examined for voice commands in virtual reality. In forthcoming research, we want to manipulate voice commands to exploit these vulnerabilities and devise effective mitigation techniques.

6 Acknowledgments

This work was supported by the Ministry of Science and ICT grant through the Information Technology Research Center (ITRC) Program (Project No. RS-2023-00228996, 50%), and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Ministry of Science and ICT (Project No. 2021-0-01816, 30%; RS-2024-00438551, 20%)

References

[1] Sneha Basak, Himanshi Agrawal, Shreya Jena, Shilpa Gite, Mrinal Bachute, Biswajeet Pradhan, and Mazen Assiri. Challenges and limitations in speech recognition technology: A critical review of

speech signal processing algorithms, tools and systems. https://cdn.techscience.cn/ueditor/files/cmes/135-2/TSP_CMES_21755/TSP_CMES_21755.pdf,, 2023.

[2] Alberto Compagno, Mauro Conti, Daniele Lain, and Gene Tsudik. Don't skype & type! acoustic eavesdropping in voice-over-ip. https://dl.acm.org/doi/abs/10.1145/3052973.3053005, 2017.

[3] Randa El-Bialy, Daqing Chen, Souheil Fenghour, Walid Hussein, Perry Xiao, Omar H Karam, and Bo Li. Developing phoneme-based lip-reading sentences system for silent speech recognition. https://ietresearch.onlinelibrary.wiley.com/doi/epdf/10.1049/cit2.12131, 2023.

[4] Serife Kucur Ergünay, Elie Khoury, Alexandros Lazaridis, and Sébastien Marcel. On the vulnerability of speaker verification to realistic voice spoofing. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7358783, 2015.

[5] Marc Fueldner. Microphones. https://www.sciencedirect.com/science/article/abs/pii/B9780128177860000487, 2020.

[6] Neil Zhenqiang Gong, Altay Ozen, Yu Wu, Xiaoyu Cao, Richard Shin, Dawn Song, Hongxia Jin, and Xuan Bao. Piano: Proximity-based user authentication on voice-powered internet-of-things devices. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7980172, 2017.

[7] MD Jon Wahrenberger. Microphone Selection Guide — speechrecsolutions.com. https://www.speechrecsolutions.com/MicGuide.htm. [Accessed 02-08-2024].

[8] Mikko Korkiakoski, Paula Alavesa, and Panos Kostakos. Preference in voice commands and gesture controls with hands-free augmented reality with novel users. [online], 2024. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10458875.

[9] Yihao Li, Xiongwei Zhang, and Meng Sun. A unified speech enhancement approach to mitigate both background noises and adversarial perturbations. https://www.sciencedirect.com/science/article/pii/S1566253523000799, 2023.

[10] Fang Liu, Xinhang Zhao, Zihao Zhu, Zhongping Zhai, and Yongbin Liu. Dual-microphone active noise cancellation paved with doppler assimilation for tads. https://www.sciencedirect.com/science/article/pii/S0888327022007993, 2023.

[11] Arpita Dinesh Sarang, Mohsen Ali Alawami, and Ki-Woong Park. Mv-honeypot: Security threat analysis by deploying avatar as a honeypot in cots metaverse platforms. https://cdn.techscience.cn/files/CMES/2024/online/CMES0801/TSP_CMES_53434/TSP_CMES_53434.pdf, 2024.

[12] Cong Shi, Xiangyu Xu, Tianfang Zhang, Payton Walker, Yi Wu, Jian Liu, Nitesh Saxena, Yingying Chen, and Jiadi Yu. Face-mic: inferring live speech and speaker identity via subtle facial dynamics captured by ar/vr motion sensors. [online], 2021. https://dl.acm.org/doi/pdf/10.1145/3447993.3483272.

[13] Takeshi Sugawara, Benjamin Cyr, Sara Rampazzi, Daniel Genkin, and Kevin Fu. Light commands:{Laser-Based} audio injection attacks on {Voice-Controllable} systems. https://www.usenix.org/conference/usenixsecurity20/presentation/sugawara, 2020.

[14] Kevin Townsend. Meta's virtual reality headset vulnerable to ransomware attacks: Researcher. https://www.securityweek.com/metas-virtual-reality-headset-vulnerable-to-ransomware-attacks-researcher/, 2024.

[15] Payton Walker, Tianfang Zhang, Cong Shi, Nitesh Saxena, and Yingying Chen. Barrierbypass: Out-of-sight clean voice command injection attacks through physical barriers. https://dl.acm.org/doi/pdf/10.1145/3558482.3581772, 2023.

[16] Chen Yan, Guoming Zhang, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. The feasibility of injecting inaudible voice commands to voice assistants. https://ieeexplore.ieee.org/document/8669818, 2019.

[17] Xuejing Yuan, Yuxuan Chen, Yue Zhao, Yunhui Long, Xiaokang Liu, Kai Chen, Shengzhi Zhang, Heqing Huang, Xiaofeng Wang, and Carl A Gunter. {CommanderSong}: a systematic approach for practical adversarial voice recognition. https://www.usenix.org/system/files/conference/usenixsecurity18/sec18-yuan.pdf, 2018.

[18] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. Dolphinattack: Inaudible voice commands. https://dl.acm.org/doi/pdf/10.1145/3133956.3134052, 2017.

[19] Yicheng Zhang, Carter Slocum, Jiasi Chen, and Nael Abu-Ghazaleh. It's all in your head (set): Side-channel attacks on {AR/VR} systems. https://www.usenix.org/conference/usenixsecurity23/presentation/zhang-yicheng, 2023.