CHINESE JOURNAL OF ELECTRONICS

Hierarchical Security Survey of Voice Control Systems: Vulnerabilities, Countermeasures, and Future Horizons

Yangyang Gu, Haozhe Xu, Jing Chen, Jiahua Xu, Yebo Feng, Ju Jia, Cong Wu, Teng Li, Jian Shen, Jianfeng Ma

Citation: Yangyang Gu, Haozhe Xu, Jing Chen, Jiahua Xu, Yebo Feng, Ju Jia, Cong Wu, Teng Li, Jian Shen, Jianfeng Ma, Hierarchical Security Survey of Voice Control Systems: Vulnerabilities, Countermeasures, and Future Horizons, Chinese Journal of Electronics, 2026, 35(3), 1153–1179. doi: 10.23919/cje.2025.00.174.

View online: https://doi.org/10.23919/cje.2025.00.174

Chinese Journal of Electronics
vol. 35, no. 3, pp. 1153–1179, May 2026
https://doi.org/10.23919/cje.2025.00.174

Cybersecurity

REVIEW

Hierarchical Security Survey of Voice Control Systems: Vulnerabilities, Countermeasures, and Future Horizons

Yangyang Gu[1], Haozhe Xu[1], Jing Chen[1], Jiahua Xu[2,3], Yebo Feng[3,4], Ju Jia[5], Cong Wu[1,3,6], Teng Li[7], Jian Shen[8], and Jianfeng Ma[7]

1. *Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China*
2. *Computer Science Department, University College London, London WC1E 6BT, UK*
3. *Exponential Science, Grand Cayman KY1-1106, Cayman Islands*
4. *College of Computing and Data Science, Nanyang Technological University, Singapre 639798, Singapre*
5. *School of Cyber Science and Engineering, Southeast University, Nanjing 211189, China*
6. *Centre for Blockchain Technologies, University College London, London WC1E 6BT, UK*
7. *School of Cyber Engineering, Xidian University, Xi'an 710071, China*
8. *School of Information Science and Engineering, Zhejiang Sci-Tech University, Hangzhou 314423, China*

Corresponding author: Cong Wu, Email: cnacwu@whu.edu.cn
Manuscript Received May 8, 2025; Accepted October 9, 2025; Published Online November 18, 2025

Abstract — The widespread adoption of voice control systems (VCSs) in smart devices and their growing integration into daily life highlight the urgent need to address their security challenges. Recent studies have revealed a range of vulnerabilities in VCSs, which threaten user privacy and safety. Despite these findings, there remains a notable absence of a comprehensive and systematic review that thoroughly examines these vulnerabilities and the corresponding defense strategies. This gap hinders VCS designers from fully understanding and addressing the security risks inherent in these systems. To bridge this gap, our research introduces a hierarchical model for VCS, providing a structured and innovative framework for organizing and analyzing existing literature. We categorize attacks based on their underlying technical principles and conduct an in-depth analysis of various aspects, such as attack methodologies, targets, vectors, and behaviors. Furthermore, we synthesize and critically evaluate current defense mechanisms, offering practical recommendations to strengthen VCS security. Analyzing VCS security from a hierarchical perspective, this work provides designers with the tools to effectively identify and mitigate potential threats while establishing a foundation for future research and advancements in VCS security.

Keywords — Speaker recognition, Voice control systems, Replay attack, Jamming attack, Voice-synthesis attack.

Citation — Yangyang Gu, Haozhe Xu, Jing Chen, *et al.*, "Hierarchical security survey of voice control systems: Vulnerabilities, countermeasures, and future horizons," *Chinese Journal of Electronics*, vol. 35, no. 3, pp. 1153–1179, 2026. doi: 10.23919/cje.2025.00.174.

I. Introduction

Voice control systems (VCSs) have emerged as a cornerstone of modern smart technologies, revolutionizing user-interaction paradigms across diverse applications, ranging from mobile devices to sophisticated smart home ecosystems. The rapid evolution of VCS has driven their widespread adoption, with the market projected to reach $15.6 billion by 2025. This remarkable expansion is fueled by the versatile capabilities of VCS, enabling users to seamlessly manage smart home environments, conduct online transactions, and access a wide range of digital services through intuitive voice commands. However, this rapid proliferation has simultaneously unveiled a complex array of security challenges, particularly with regard to data privacy and the

Associate Editor: Prof. Youliang Tian, Guizhou University.

increasing sophistication of cyberattacks targeting VCS infrastructures.

The complex architecture of VCS, with both software and hardware components, inherently hides a spectrum of security vulnerabilities. These vulnerabilities serve as potential avenues for malicious exploitation, posing serious challenges for VCS developers tasked with preemptively identifying and mitigating these multifaceted threats. For example, adversaries have demonstrated the ability to exploit hardware vulnerabilities in microphones to inject unauthorized commands into VCSs employing unconventional methods such as laser [1], [2]. Furthermore, the advent of adversarial machine learning techniques has enabled attackers to engineer audio commands that are imperceptible to human ears yet are interpreted as malicious instructions by VCS algorithms. In the realm of third-party VCS services, the integration of malicious applications that can be inadvertently activated by unsuspecting users further compounds security risks [3], [4]. Despite the diversity and complexity of these threats, most defense mechanisms often adopt a myopic focus, addressing isolated attack vectors without considering the holistic implications of vulnerabilities across the entire VCS. This fragmented approach not only limits the effectiveness of existing countermeasures but also prevents a comprehensive understanding of the interconnected risks and attack methodologies that threaten VCS infrastructures. Consequently, there is a pressing imperative for a systematic and integrative analysis that not only analyzes contemporary attack strategies but also critically evaluates the robustness and limitations of current defense frameworks, thereby paving the way for more resilient and adaptive security solutions in the VCS domain.

Related survey and limitations Our survey distinguishes itself from prior research by categorizing the structure of VCS into distinct layers: the physical layer, the data layer, the model layer, and the service layer. This approach enables a comprehensive examination of security risks and defense mechanisms, considering each layer's specific integration perspective. Table 1 ([4]–[10]) provides a comparative summary of our work in relation to similar studies. For example, Zhang *et al.* [4] focused specifically on security threats from malicious services within VCS. Bai *et al.* [6] explored acoustic technologies and their applications in domestic and industrial settings. Abdullah *et al.* [7] and Chen *et al.* [8] provided an in-depth analysis of the risks associated with automatic speech recognition (ASR) and speaker verification (SV) systems, integral components of VCS. Zhang *et al.* [9] surveyed adversarial attacks against ASRs in detail. Gurowiec *et al.* [10] provided a detailed study on attacks and defense schemes in speech emotion recognition systems. These studies, while insightful, often focus on specific layers or subsets of layers within VCS, possibly overlooking the full view of security threats. For example, Zhang *et al.* [4] and Bai *et al.* [6] did not comprehensively address vulnerabilities in the service layer. Similarly, studies in [7]–[9] primarily analyze the model layer, neglecting other critical layers. In contrast, our survey aims to bridge these gaps by providing an all-encompassing analysis of VCS security across all layers. Although the

Table 1 A comparison of existing survey work in the smart voice control systems

Article	Year	Major contribution	Scope									
			Attack					Defense				
			SVS	NLPS	HS	SWSS	MDS	SVS	NLPS	HS	SWSS	MDS
[4]	2019	A comprehensive study about third party skills with threats in VA and how to avoid them	✘	✘	✘	✔	✔	✘	✘	✘	✔	✔
[5]	2020	A thorough survey over the security and privacy of smart home personal assistants	✔	✔	✔	✔	✘	✔	✔	✔	✔	✘
[6]	2020	A comprehensive survey on acoustic-based technologies and applications in life and industry	✔	✘	✔	✘	✔	✘	✘	✘	✘	✔
[7]	2021	A detailed study about security threats and defense schemes in ASR and SV	✔	✘	✘	✘	✔	✔	✘	✘	✘	✔
[8]	2022	A modularized and comprehensive survey about ASR security and comparison between ASR and image recognition system (IRS) security	✘	✔	✘	✘	✔	✘	✔	✘	✘	✔
[9]	2022	A detailed investigation of adversarial attacks against ASR	✘	✘	✘	✘	✔	✘	✘	✘	✘	✔
[10]	2024	A detailed study about attacks and defense schemes in speech emotion recognition systems	✔	✔	✔	✔	✘	✔	✔	✔	✔	✘
Ours	2025	A comprehensive overview of the challenges faced by VCS and its future direction	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔

Note: SVS: speaker verification security. NLPS: natural language processing security. HS: hardware security. SWSS: skill web service security. MDS: mobile device security.

work in [5] presents a comprehensive overview of VCS attacks, it mainly focuses on smart home applications, omitting extensive discussion on mobile device integration. The work in [11] offers an exhaustive review of attacks and countermeasures in voice assistants (VAs), but lacks a structured view of VCS.

Vision and scope In addition to current surveys lacking a comprehensive perspective on examining VCS security issues, researchers also urgently need to address the following questions:

- VCS systems are vulnerable to a variety of attacks, such as transduction attacks, voice synthesis attacks, and adversarial attacks. This susceptibility stems from two main factors: the advancement of technology providing attackers with more sophisticated tools and algorithms, and the inherent complexity of VCS as a system comprising multiple hardware and software components, each with unique security vulnerabilities. Understanding how these attacks operate and identifying the specific vulnerabilities of VCS components are crucial for designers.
- Beyond understanding attack mechanisms, it is vital that VCS designers are aware of effective defense solutions. Current defense strategies, such as adversarial training, are often tailored to specific attack types [12]–[18], posing a challenge in selecting appropriate defense combinations to robustly enhance the VCS's ability to withstand a wide range of attacks.
- The diverse nature of VCS, which encompasses both ASR and SV, often leads to the study of these components in isolation [7], [8]. However, many VCS systems incorporate both ASR and SV functionalities, and attacks against these systems often exhibit overlapping characteristics. Designers must therefore jointly consider the vulnerabilities of both ASR and SV when developing VCS.

In our paper, we conduct a comprehensive and systematic study of VCS security. To thoroughly discuss the various attacks facing VCS, we categorize these attacks based on the hierarchical model of VCS, which includes the physical layer, the data layer, the model layer, and the service layer, and then analyze and assess the corresponding defense schemes, tailored to each type of attack within these layers as shown in Figure 1. For each attack, we conduct a multidimensional evaluation, considering aspects such as attack methods, targets, vectors, and performance impact, and discuss potential directions for future research. In addition, we of-

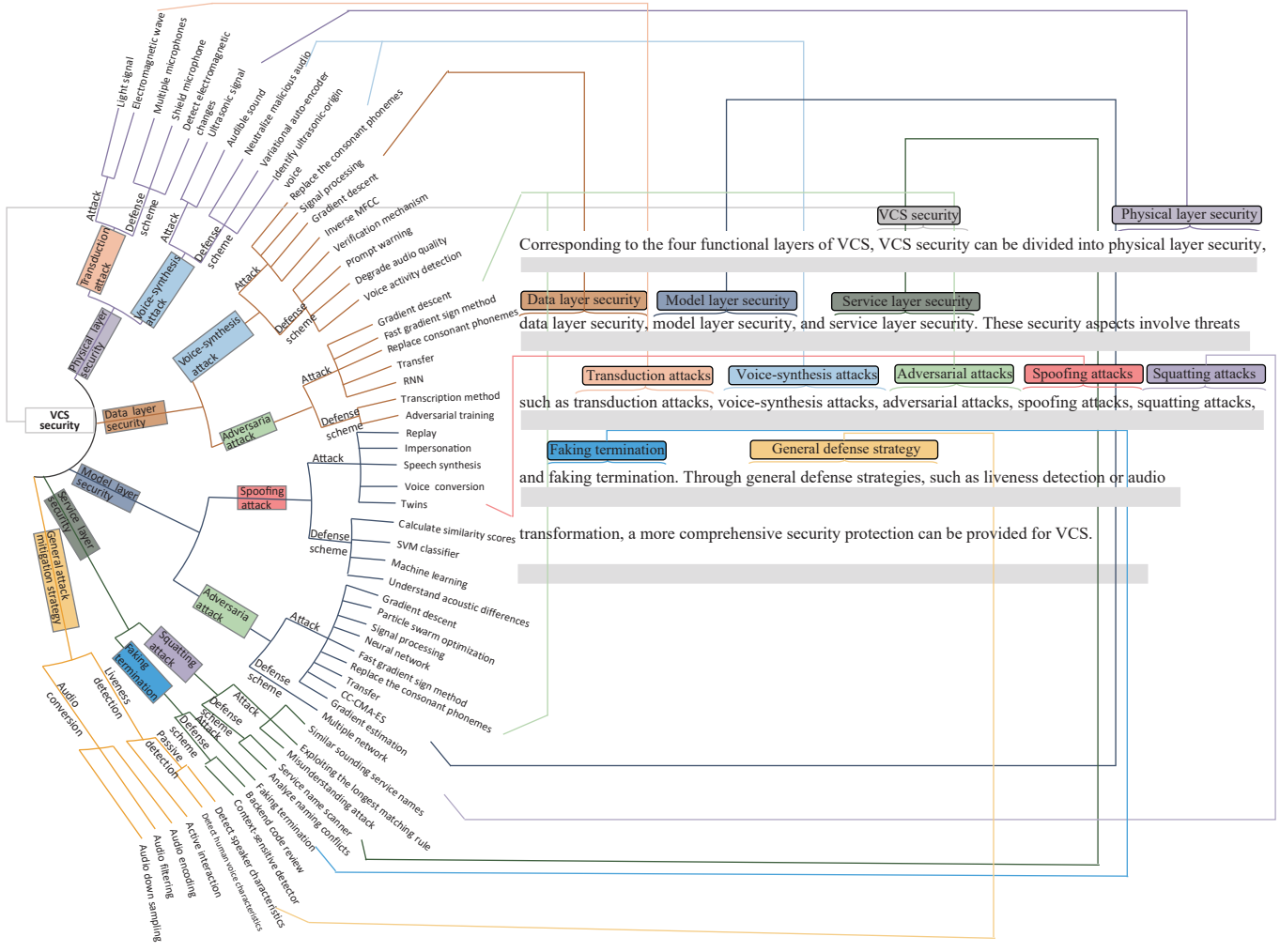


Figure 1 An illustrative overview of VCS security.

fer practical combinations of defense schemes, providing valuable insights for VCS researchers to effectively counter diverse attacks. An illustrative organizational structure for our survey article is presented in Figure 2.

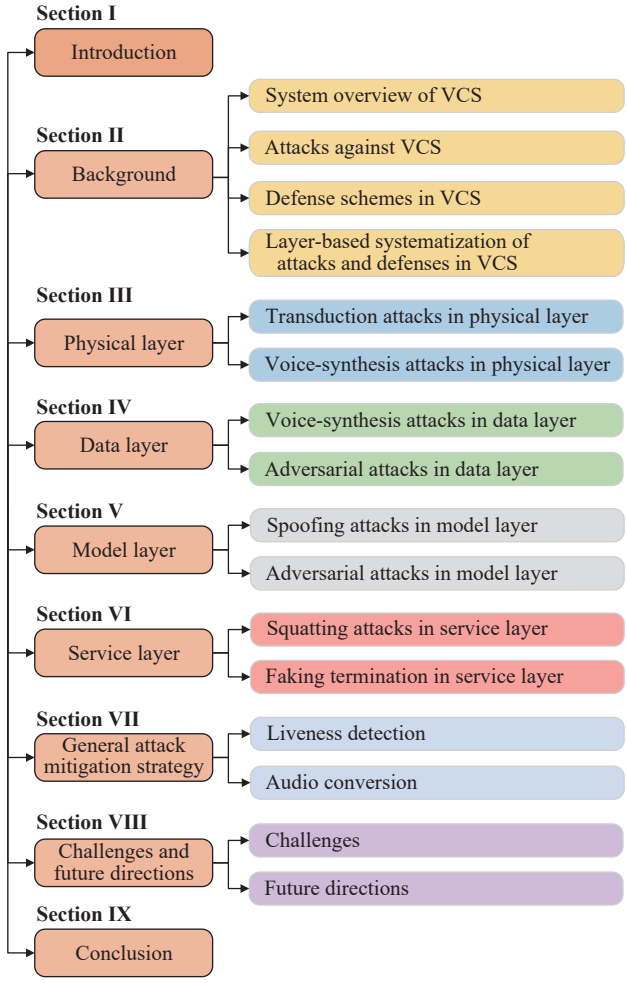


Figure 2 An illustrative overview of structure of our survey.

Our work distinguishes itself from existing surveys in several comprehensive aspects: i) We provide extensive coverage of a broad range of attack methods applicable to each layer within the hierarchical model of VCS. This approach offers a detailed understanding of the potential vulnerabilities at different levels of the system. ii) Our overview is expanded to include various devices that incorporate VCS, where we detail the unique security risks associated with each device type. This broadens the scope of our survey to encompass a wider array of real-world applications. iii) We present a detailed introduction to defense schemes corresponding to these attack methods, thereby enabling readers to not only understand these strategies but also to implement them in practical scenarios. This integrated approach to both attack and defense perspectives provides a well-rounded view of VCS security.

Contributions We summarize our contribution as follows:

- Our unique approach to examining VCS's security involves dividing VCS into four distinct layers: the physical layer, data layer, model layer, and service layer. This hierarchical model allows for a systematic and detailed analysis of both attacks and defense schemes at each layer, providing VCS researchers with a comprehensive framework to understand and evaluate associated risks accurately.
- Using our proposed hierarchical model, we have systematically integrated and assessed potential attacks on each VCS layer. By employing various metrics such as attack complexity and potential impact, we offer a feasibility analysis that enables users to accurately gauge the security threats posed to VCS by these attacks.
- We have developed and evaluated defense strategies tailored to each layer of VCS, considering practical effectiveness and usability from multiple dimensions. In addition, we introduce a generalized attack mitigation strategy to help designers construct a more robust and comprehensive defense system for VCS.
- Through an extensive analysis of existing literature, we provide targeted recommendations to enhance current security research and directions for future development in the field, addressing gaps and emerging trends in VCS security research.

II. Background

In this section, we brief the system overview of VCS, and then systematically describe layer-based attacks and defensive schemes.

1. System overview of VCS

As Siri achieved significant success on iPhones, VCS have seen widespread integration into daily life. Nowadays, smart devices including smartphones, smart home devices, and smart vehicles, are equipped with voice assistants. VCS enable users to perform various tasks such as unlocking devices, setting alarms, and opening doors, thereby enhancing interaction with smart technology. Popular VCSs including Google Home, Amazon Echo, and Apple HomePod, have expanded their capabilities through third-party services, providing features like voice translation and audio transcription.

With the rapid development and increasing prevalence of VCS, security concerns are likewise growing. To thoroughly analyze and effectively organize the potential risks and defensive strategies for voice assistants, we categorize the VCS into four functional layers: physical layer, data layer, model layer, and service layer. This layered approach allows for a detailed study of how a voice control system processes voice commands and provides services to users, as illustrated in Figure 3.

1) Physical layer

Transmission channels Acoustic signals, when transmitted over-the-air in VCS, are subject to various environmental influences that can degrade their quality [19]. The signal intensity diminishes with distance, which constrains the effective operational range be-

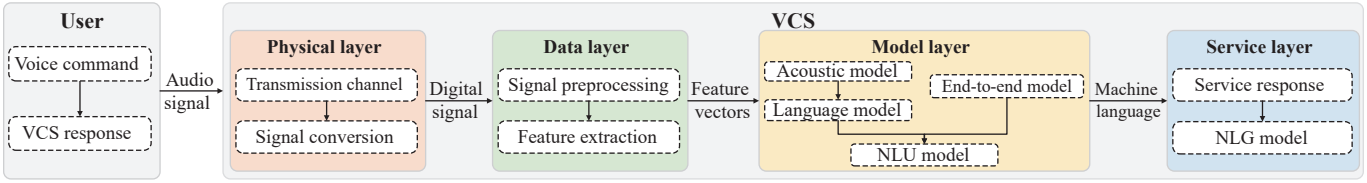


Figure 3 A hierarchical workflow diagram of VCS. After a user issues a voice command to the VCS, the user will receive a response from the VCS. Based on the response, the user can choose to issue another relevant command to the VCS.

tween the source and the target device. Furthermore, background noise and multipath effects caused by signal reflection result in the reception of mixed signals, complicating the interpretation of the original acoustic input. Like the channel vulnerability in WiFi sensing [20]–[24], this vulnerability in the transmission channel is particularly concerning in security contexts, where attackers could exploit these transmission challenges by deploying malicious audio. Such attacks might aim to compromise the VCS by leveraging signal interference, necessitating countermeasures. Attackers, aware of these transmission challenges, might enhance the robustness of their malicious audio or resort to generating electronic signals that mimic acoustic ones, to increase the likelihood of successful system penetration.

Acoustic signal conversion Acoustic signals must first be converted to digital signals for the VCS employing two critical stages. In the first stage, the system employs sensors, typically microphones, to transform acoustic signals into electrical signals. Among microphones, electrodynamics and capacitive types are widely used, with the latter being more prevalent in modern devices due to their sensitivity and reliability. Capacitive microphones operate by detecting changes in capacitance caused by the vibration of moving plates in response to sound waves, thus converting acoustic signals into electrical ones [25].

The second stage involves an analog-to-digital converter (ADC), which digitizes the electrical signals. The ADC employs a sample and holds circuit to capture and stabilize the signal value, followed by a quantizer that processes these samples, converting them into binary numbers representing the digital signal. This conversion is pivotal for effective VCS processing, but each stage also presents potential security vulnerabilities, such as susceptibility to signal manipulation or interference, that must be carefully managed to ensure system integrity.

2) Data layer

Signal preprocessing Signal preprocessing is vital in VCS, as it involves processing raw speech signals to extract clean speech for analysis. Band-pass filtering is a commonly used technique for VCS to isolate speech signals within the 300–3400 kHz frequency band. For devices with multiple microphones, techniques such as multichannel acoustic echo cancellation, reverberation control, and source separation, are essential for multichannel speech capture. These typical techniques ensure high-quality signal capture, crucial for accurate speech recognition. Recently, the development of the deep learning offers improved novel methods for noise mitigation and signal clarity [26], which is critical for robust VCS functioning, particularly in complex acoustic environments.

Acoustic feature extraction This module aims to extract distinctive features from speech signals, accounting for variations in speaker characteristics, environmental noise, and transmission channels. The process begins with segmenting the processed audio into frames using a sliding window technique, followed by feature extraction. Common algorithms for feature extraction include short-time Fourier transform (STFT), linear predictive coding (LPC) [27], Mel frequency cepstral coefficients (MFCC) [28], and perceptual linear prediction (PLP) [29]. Each algorithm plays a unique role in enhancing the robustness of VCS by efficiently capturing and representing different acoustic properties of speech.

3) Model layer

Acoustic model After the feature extraction in the data layer, the acoustic model establishes a statistical relationship between these features and acoustic units like phonemes using machine learning algorithms. Traditional models commonly use Gaussian mixture models (GMM) and hidden Markov models (HMM), while contemporary models increasingly employ neural networks such as convolutional neural networks (CNN) and recurrent neural networks (RNN) [30], enhancing the accuracy and efficiency of phoneme conversion.

Language model The language model predicts the likelihood of word sequences forming coherent sentences and distinguish homophones based on context. This model is pivotal in controlling VCS output using common words and grammatical rules. Models are generally divided into statistical models (e.g., N-gram, HMM) and neural network-based models, with the latter becoming increasingly prevalent in modern VCS.

Phonetic dictionary This component links the acoustic and language models by mapping words to phonemes, essential for decoding the user's intent in VCS. The phonetic dictionary facilitates the establishment of relationships between the models' units, thereby aiding in effective system decoding.

End-to-end model To streamline VCS, end-to-end models have been developed. These models directly convert acoustic features into word sequences, integrating the functions of both acoustic and language models. This simplification reduces potential attack vectors, en-

hancing system security. Examples include connectionist temporal classification (CTC) [31], Transducer [32], and Attention models [33].

Natural language processing (NLU) NLU model enables machines to understand and act upon natural language inputs. Its development has evolved from rule-based to statistical models, and now to neural network-based models like Word2vec [34] and BERT [35], reflecting the industry's shift towards more sophisticated, context-aware processing.

Speaker verification (SV) SV focuses on identifying the legal speaker rather than interpreting speech content. And it is also different from speaker identification (SI), with the latter confirming a speaker's identity. In the model layer, speaker verification differs from standard VCS in that it determines the likelihood that a speaker is who it claims to be, thus requiring specialized discussion regarding its security implications.

4) Service layer

After interpreting the user's intent in the model layer, the VCS responds in the service layer. Utilizing natural language generation (NLG), the system conveys the required information in text or audibly through text-to-speech (TTS) technology [36]. Technological advancements have expanded VCS capabilities to include third-party services, enabling more complex functionalities. However, this expansion introduces new security challenges, such as increased exposure to data breaches and exploitation of service vulnerabilities, necessitating robust security measures.

2. Attacks against VCS

As shown in Figure 4, current researches highlight the vulnerability of VCS to a range of attacks, stemming from its complex architecture which presents potential vulnerabilities at each layer. To equip system designers with a comprehensive understanding of these threats, this section delves into the nature of attacks against VCS, examining them from the perspectives of threat models, attack metrics, and attack objectives. We will explore how threat models help identify potential security weaknesses, discuss attack metrics to assess the impact and severity of attacks, and analyze the typical goals of attackers targeting VCS.

1) Threat model

We consider attack scenarios under these assumptions:

i) Attackers may have prior knowledge of the target VCS, including its brand and model. For common VCS, attackers might possess the same device model.

ii) The VCS operates on a trusted platform with

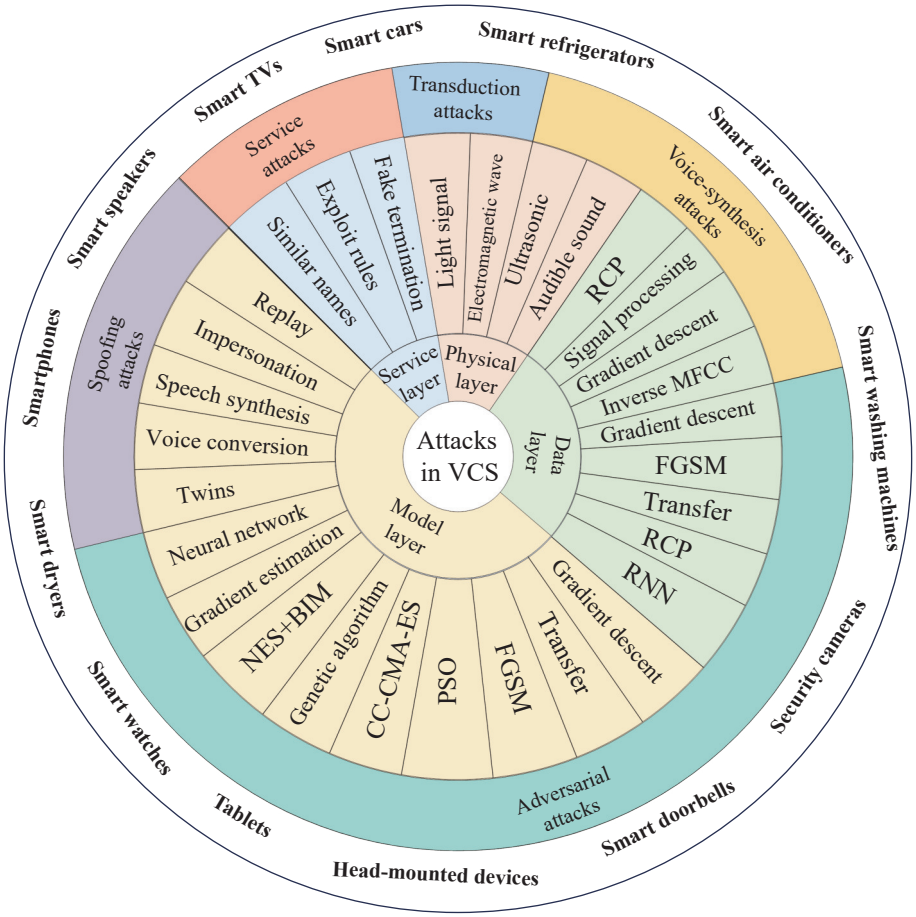


Figure 4 Attacks in the different layers of VCS. PSO: particle swarm optimization. ATN: adversarial transformation network. NES: natural evolution strategy. BIM: basic iterative method. CC-CMA-ES: cooperative-coevolution covariance-matrix-adaptation evolution-strategy. FGSM: fast gradient sign method. GN: generative network. GCA: convolutional auto-encoder. RCP: replacing the consonant phonemes.

encrypted communication, preventing attacks via the operating system, hardware, or communication interception.

iii) Attackers' knowledge varies: a) White/Grey-box: Full or partial model knowledge, applicable to open-source VCS. b) Black-box: No model knowledge, typical in commercial VCS.

2) Attack metrics

Based on the threat model, we assess VCS attacks as follows:

Attack method The method determines the targeted VCS layer. Some methods exploit non-audio signals to attack the physical layer, while others use audio-based methods to deceive the model layer.

Attack target Targets can vary from specific models (requiring detailed knowledge) to general hardware vulnerabilities that affect multiple VCS models [37].

Attack vector The vector refers to the carrier of the attack (e.g., the air, the power line, the solid surface) and must be robust against environmental factors and signal degradation.

Performance The effectiveness of an attack is measured by its success rate and the maximum effective distance, indicating its practicality in real scenarios.

3) Attack goal

Attackers primarily aim to disrupt VCS operations and cause command execution to be executed incorrectly. Untargeted attacks may cause incorrect audio decoding, while targeted attacks seek to execute specific malicious commands. Security impacts include unauthorized control of smart devices (e.g., unlocking doors), while privacy concerns involve eavesdropping or accessing personal information via malicious services.

3. Defense schemes in VCS

The range of defense strategies for VCS is relatively narrow compared to the diversity of attack methods. Current defenses primarily fall into two categories: i) Layer-specific solutions, such as adversarial training, which fortify particular layers like the model layer against specific types of attacks, yet they might not be effective against threats targeting other layers. This category also includes specialized encryption for communication security and hardware-focused protections. ii) Holistic strategies, like liveness detection, are designed to protect the entire VCS by verifying the authenticity of voice inputs, thereby securing all layers from a range of potential attacks. The VCS security often depends on the integration of both layer-specific and holistic defenses to fully address the multifaceted nature of potential threats.

4. Layer-based systematization of attacks and defenses in VCS

The approach for classifying attacks on VCS typically revolves around either human perception or specific VCS components. However, this method can inadvertently group distinct attack methodologies under the same category, despite their differing technical principles. For instance, both DolphinAttack [38] and hidden voice command [39] are categorized as voice-synthesis attacks. Nonetheless, they fundamentally differ in their mechanisms. DolphinAttack exploits microphone hardware vulnerabilities, while hidden voice command targets the data preprocessing algorithms.

Although the component-based classification method provides a systematic approach to categorize attacks, it often leads to an isolated analysis of SV and ASR components, as seen in existing literature [7], [8]. Our research suggests that despite the distinct outputs of these two components, they share similar technical principles and vulnerabilities that can be exploited by attackers. Thus, analyzing them separately may obscure the full understanding of the potential attack vector.

To address these classification challenges, our work proposes a novel categorization of attacks based on the VCS architecture, dividing them into attacks on the physical layer, data layer, model layer, and service layer, as outlined in Figure 3. This approach allows for a more precise identification of the vulnerabilities specific to each layer. We will also categorize defense solutions accordingly. Those designed to protect specific layers will be discussed in their respective sections, while generalized defense strategies applicable across all VCS layers will be comprehensively covered in Section VII. The following sections will provide a detailed analysis of the attack threats and corresponding countermeasures for each layer within the VCS.

III. Physical Layer

The physical layer of VCS plays a pivotal role in converting acoustic signals into digital formats. This conversion process involves the transmission of acoustic signals through channels to microphone devices, which are then digitized. However, this layer is vulnerable to a range of attacks due to the lack of signal verification in the transmission channel and the susceptibility of microphone devices to manipulation. Attackers can exploit these vulnerabilities by transmitting malicious audios or using specific signals to make microphones generate attacker-desired digital signals. This section explores various attack types targeting the physical layer, including transduction attacks that manipulate physical components to create false inputs and voice-synthesis attacks that craft misleading audio inputs, A comprehensive summary and comparison of these attacks are detailed in Table 2, presenting their methods, targets, and potential impacts on VCS.

1. Transduction attacks in physical layer

Transduction attacks exploit microphone vulnerabilities in VCS by using non-standard signals that microphones should typically not recognize. These attacks manipulate the physical properties of microphone equipment to induce unintended digital signal genera-

tion, leading to VCS compromise.

Light signal Microphones, designed to be sensitive only to sound signals, can unexpectedly respond to modulated light signals. Sugawara *et al.* [1] demonstrated that by modulating laser amplitude, attackers can transmit malicious commands to VCS from distances exceeding 100 m, exploiting this unintended sensitivity.

Electromagnetic wave (EW) Microphone circuits can inadvertently act as antennas, receiving EWs which are then demodulated into malicious commands. This vulnerability allows attackers to use electromagnetic coupling, as shown by Kasmi and Esteves [40], to control voice assistants through headphone cables. Further research indicates that specific EWs generated by wired [41] or wireless chargers [42] can also inject commands into VCS.

Defense schemes against transduction attacks For light-based attacks, sensor fusion technology [43] offers a software-level defense by exploiting the difference between the omnidirectional nature of sound and the directional nature of light, which typically affects only one microphone. This approach detects discrepancies between multiple microphone inputs. Hardware solutions include shielding microphones with opaque covers to block light, but this may attenuate sound signals [44]. Against EW attacks, sensors detecting electromagnetic changes can alert to potential intrusions. A comprehensive defense strategy involves adding authentication mechanisms to VCS, effectively mitigating various transduction attacks.

2. Voice-synthesis attacks in physical layer

Voice-synthesis attacks in the physical layer of VCS differ from transduction attacks as they primarily utilize ultrasonic signals, which are inaudible to humans but can be received by most microphones. These attacks exploit the fact that ultrasound can be demodulated in the system through nonlinear amplifiers after being received by the microphone. Attackers modulate malicious commands into these ultrasonic signals, leading to interference with normal functions of VCS. Researchers have demonstrated such attacks using solid surface transmission [45] and air transmission methods [38], [46]–[53]. Additionally, an innovative method has been proposed that leverages specific current fluctuations to remotely attack VCSs using acoustic waves generated by switching mode power supplies (SMPS) within the same power grid [54].

Defense schemes against voice-synthesis attacks To counteract these attacks, hardware-based defenses include using microphones that are incapable of receiving ultrasonic signals, similar to the iPhone 6 Plus microphone [46], or implementing a "guarding" mechanism or "voice cloaks" with an external signal generator to neutralize malicious audio [55], [56]. However, replacing microphones or adding signal generators in existing VCS products is impractical, making software-level defenses more viable. For air-transmitted voice-synthesis attacks, the demodulated attack signal in the 500–1000 Hz range differs significantly from the original signal, detectable via a variational auto-encoder [57]. Additionally, a microphone array can calculate the attenuation rate of sound waves at different microphones, identifying ultrasonic-origin voice commands [58]. Solid surface attacks lack this attenuation distinction, but their recovered attack signal introduces new frequency components from 10 kHz to 20 kHz, providing another basis for detection and differentiation from baseband signals.

IV. Data Layer

In VCS, the data layer is essential for refining the digital signals obtained from the physical layer. Its primary function is to filter and enhance audio signals, which often contain a mix of human voices and background

Table 2 Comparison of transduction attack and voice-synthesis attack in physical layer

Attack type	Article	Signal	Attack target	Attacker vector	Performance	
					Success rate (%)	Distance
Transduction attacks	Sugawara *et al.* [1]	Light	Microphone	Over-the-air	90	25 m
	Kasmi *et al.* [40]	EW	Headphone cable	Over-the-air	Not avaliable	Not avaliable
	GhostTalk [41]	EW	Charging Port	Power line	100	Not avaliable
	Dai *et al.* [42]	EW	Microphone	Over-the-air	91	5 cm
Voice-synthesis attacks	Surfingattack [45]	Ultrasonic	Microphone	Solid Surface	100	2.5 m+
	Dolphinattack [38]	Ultrasonic	Microphone	Over-the-air	100	165 cm
	Iijima *et al.* [47]	Ultrasonic	Microphone	Over-the-air	100	10 m
	Capspeaker [48]	Ultrasonic	Capacitor	Over-the-air	80	10.5 cm
	Roy *et al.* [49]	Ultrasonic	Microphone	Over-the-air	50	760 cm
	Song *et al.* [50]	Ultrasonic	Microphone	Over-the-air	80	3 m
	Yan *et al.*[51]	Ultrasonic	Microphone	Over-the-air	100	165 cm
	Nuit [46]	Ultrasonic	Microphone	Over-the-air	100	360 cm
	Yang *et al.* [54]	Audible sound	SMPS	Over-the-air	90	23 m

noise. Voice activity detection algorithms, like G.729 [59], are used to isolate segments containing human speech, followed by the application of low-pass filters to reduce noise and improve VCS's recognition capabilities. Advances in Deep Neural Networks (DNN) have further enhanced these preprocessing tasks [26], [60]. Besides noise reduction, this layer is responsible for extracting relevant features from audio signals using techniques such as delta frequency cepstral (DFC), mel-frequency cepstral coefficients (MFCC), linear predictive coding (LPC), mel-frequency cepstral smoothing (MFCS), and perceptual linear prediction (PLP), preparing them for subsequent phoneme, character, or word sequence recognition. However, vulnerabilities within these feature extraction processes have been identified, allowing attackers to execute voice-synthesis attacks and adversarial attacks. This section discusses these attacks, exploring how they exploit the data layer's weaknesses. Table 3 provides a comparative summary of these attacks, outlining their strategies and impact on VCS.

1. Voice-synthesis attacks in data layer

Voice-synthesis attacks in the data layer of VCS utilize technology to create specific speech patterns that can be misinterpreted as legitimate instructions. Due to the lossy nature of signal processing and feature extraction in this layer, attackers can craft voice segments that mimic target voice features. These segments are then mistakenly recognized as intended voice signals by the VCS after undergoing preprocessing. A key characteristic of these attacks is the significant difference between the synthesized and target audio signals, making detection challenging for users.

For instance, Bispham *et al.* [61] produced nonsensical voice commands by altering consonant phonemes in the target voice. Despite the apparent errors, the VCS's voice model could correct these anomalies, leading to system recognition of these commands. Abdullah *et al.* [62] employed various signal processing techniques like time domain inversion and high frequency addition to morph the target voice into an unintelligible audio signal. They leveraged the VCS's sensitivity to specific keywords to construct more effective malicious samples. In another approach, Carlini *et al.* [39] and Vaidya *et al.* [63] targeted the MFCC feature extraction, common in VCS. By calculating MFCC features of the target voice and then reversing this process, they generated synthetic audio that successfully deceived the VCS. While these attacks effectively disguise their intent, the unusual noises they produce may still raise user suspicion. Therefore, Ji *et al.* [64] proposed a new attack that purely speeds up or slows down original audio instead of adding perturbations to inject malicious voice commands.

Defense schemes against voice-synthesis attacks Defending against voice-synthesis attacks in the data layer involves two strategies: detection and prevention. Detection refers to the system's ability to recognize an attack and alert the user. One common approach is the use of verification mechanisms such as CAPTCHAs [65], where the system requests a verification response from the user before executing a voice command. This allows users to confirm whether the command aligns with their intentions. Additionally, prompt warnings can notify users when the VCS receives voice commands, providing an opportunity to recognize and counteract potential attacks, even if the user may overlook the unusual noise generated by the attack [39].

Prevention strategies focus on averting the damage from an attack. Voice activity detection (VAD) is proposed as a preventive measure [62]. As a speech processing algorithm, VAD can distinguish between noise and actual speech [66], [67], effectively identifying and disregarding synthesized audio attacks. Moreover, adjusting VCS filters to slightly degrade audio quality can disrupt synthetic audio, which typically lacks robustness, while still preserving the recognition of genuine human speech [39]. These preventive measures operate by exploiting the inherent weaknesses in synthesized audio, thereby reducing the probability of successful attacks.

2. Adversarial attacks in data layer

Originally emerging in the field of image recognition [68], the concept of adversarial attacks has gained significant attention in speech recognition. Contrary to voice-synthesis attacks, adversarial attacks craft voice commands that sound normal to human users but lead to incorrect predictions by the machine learning models in VCS. These attacks typically involve introducing subtle perturbations into benign audio samples, creating what are known as adversarial samples. These samples exploit the sensitivity of deep learning models to small input changes, posing serious security and privacy risks to audio-based systems [69], [70].

In audio applications, these perturbations can be added directly to the audio waveform or to its features. When perturbations are applied to audio features, attackers must reverse-engineer the audio waveform from these features to ensure the altered audio is still recognizable by the VCS. This subsection focuses on adversarial attacks that perturb audio features, as this method intricately involves the operational principles of the data layer in VCS. Such attacks highlight the inherent vulnerabilities in the feature extraction and processing mechanisms, underscoring the need for robust defense strategies in this layer of VCS.

1) Taxonomy of adversarial attacks in VCS

Adversarial attacks in VCS can be classified based on different criteria, including attack scenario, attack target, and target vector.

Attack scenario Adversarial attacks are divided into white-box, black-box, and gray-box attacks based on the attacker's knowledge of the target VCS model. White-box attacks assume complete access to the VCS model, including its internal parameters. In contrast, black-box attacks only allow attackers to access the in-

Table 3 Comparison of voice-synthesis attacks and adversarial attacks in data layer

Attack type	Article	Attack technology	Goal	Attack target (Category)	Model knowledge	Attack vector	Cost (Time)	Performance	
								Success rate (%)	Distance
Voice-synthesis attacks	Bispham *et al.* [61]	RCP	Targeted	Google Assist (ASR)	Black	WAA	N.A.	N.A.	N.A.
	Abdullah *et al.* [62]	SP	Targeted	10 models (ASR)	Black	WTA	N.A.	100	N.A.
						WAA	N.A.	80	1 foot
				2 models (SV)		WTA	N.A.	100	N.A.
	Carlini *et al.* [39]	GD	Targeted	CMU Sphinx (ASR)	White	WAA	32 h	82	0.5 m
		IMFCC		Google Now (ASR)	Black		N.A.	80	$\leq$ 3.5 m
	Vaidya *et al.* [63]	IMFCC	Targeted	Google Now (ASR)	Black	WAA	N.A.	N.A.	N.A.
Adversarial Attacks	Houdini [71]	GD	Untargeted	DeepSpeech-2 (ASR)	White	WTA	N.A.	N.A.	N.A.
				Google Voice (ASR)	Black				
	Iter *et al.* [14]	GD	Targeted	WaveNet (ASR)	White	WTA	N.A.	N.A.	N.A.
	Yakura *et al.* [15]	GD	Targeted	DeepSpeech (ASR)	White	WWA	N.A.	50–100	$\leq$ 0.5 m
						WAA		60–90	
	Chang *et al.* [72]	RNN	Targeted	KWS (ASR)	White	WAA	$\approx$ 0.096 s	84.30	$\leq$ 4 m
	Kreuk *et al.* [73]	FGSM	Untargeted	DNN End-to-end (SV)	White	WTA	N.A.	64–94 (FAR)	N.A.
					Black			46 (FAR)	
	Li *et al.* [74]	FGSM	Untargeted	GMM i-vector (SV)	White	WTA	N.A.	99.99 (FAR)	N.A.
					Black			< 70 (FAR)	
		Transfer		DNN x-vector (SV)	Black			< 60 (FAR)	
	Marras *et al.* [75]	GD	Targeted	VGGVox (SV)	White	WTA	N.A.	< 80	N.A.
	Wang *et al.* [76]	GD	Targeted	DNN x-vector (SV)	White	WTA	N.A.	< 98.5	N.A.
	Specpatch [77]	RCP	Targeted	DeepSpeech2 (ASR)	Black	WTA	N.A.	100	N.A.
						WAA		80	1 m
	Wang *et al.* [69]	GD	Targeted	6 models (ASR)	Black	WAA	N.A.	< 91.7	10 cm

Note: RCP: replacing the consonant phonemes. SP: signal processing. GD: gradient descent. IMFCC: inverse MFCC. FGSM: fast gradient sign method. FAR: false acceptance rate. N.A.: not avaliable.

puts and outputs of the VCS, without gaining access to internal model details. Gray-box attacks represent an intermediate scenario, where the attacker has partial knowledge of the model's parameters.

Attack target In terms of objectives, adversarial attacks can be categorized into targeted attacks and untargeted attacks. Targeted attacks manipulate the VCS to produce a specific, incorrect output defined by the attacker. In contrast, untargeted attacks aim to induce any error in the system's output without a predetermined result. For example, for SV, untargeted attacks might result in the identification of any incorrect speaker, while targeted attacks would misidentify a speaker as a specific individual chosen by the attacker. Similarly, in ASR, untargeted attacks cause any incorrect transcription, whereas targeted attacks aim for a specific false transcription.

Attack vector Based on the transmission medium, adversarial attacks in VCS can be categorized into wav-to-API (WTA) attacks and wav-air-API (WAA) attacks [78]. WTA attacks involve directly feeding an adversarial audio file into the VCS API, while WAA attacks transmit the adversarial sample through the air, capturing it with the system's microphone, and then process it. Additionally, there is a wav-wave-API (WWA) attack, where adversarial audio deceives VCS after transmission through the transmitter and radio play [15]. This classification aids in understanding how adversarial audio affects VCS, indicating the need for diversified defense strategies.

2) Basic idea of designing adversarial samples

Adversarial audio samples are essentially a fusion of benign audio with carefully crafted adversarial perturbations. The primary challenge in the industry is to ensure that these perturbations, when added to benign audio, remain imperceptible to human users while effectively leading the VCS to produce erroneous outputs [79]. This task can be conceptualized as an optimization problem where the attacker's goal is to minimally alter the audio to cause misclassification by the VCS.

The process of designing such attacks in the audio domain, similar to other domains, often involves training neural networks [11]. However, the methods used in the image recognition domain, the most prevalent area of adversarial attacks application, are not directly transferable to audio due to fundamental differences. In audio, perturbations are added in the temporal domain requiring higher precision compared to the spatial do-

main in images [80]–[84].

The optimization problem in this context is expressed as follows:

$$F(x,\delta,y)=\min_{\|\delta\|} l_m(f(x+\delta),y)+c\cdot\|\delta\| \quad (1)$$

where x represents the benign audio sample, δ is the added perturbation, and y is the label output by the model $f(\cdot)$. The loss function $l_m(\cdot)$ measures the discrepancy between the model's actual output $f(x+\delta)$ and the label y. The term $c\cdot\|\delta\|$ quantifies the perturbation's impact on the original audio, where c is a balancing coefficient between attack effectiveness and audio quality. Higher values of c indicate substantial degradation in audio quality, potentially compromising the integrity of the original content.

For untargeted attacks, the attacker seeks the smallest δ that drastically alters the output $f(x+\delta)$ from y (given $f(x)=y$). Conversely, for targeted attacks, the goal is to minimize δ such that $f(x+\delta)$ closely matches a specified target output y' (where $f(x)\neq y$).

To address the inherent optimization problem in designing adversarial attacks in the data layer, several methods, primarily centered around gradient descent and deep learning techniques, have been proposed in the industry.

Gradient descent Gradient descent (GD) is a primary method for creating adversarial audio, widely studied in current research [14], [15], [71], [69], [85]. GD optimizes adversarial samples by iteratively calculating the local minimum of a loss function using partial derivatives and the chain rule. Typically utilized in white-box attack scenarios, GD requires an understanding of the target loss function's specifics. Although increasing the number of iterations can enhance the success rate of the attack, it may also compromise the quality of adversarial audio [86].

Deep learning-based methods To overcome the time and computational resource constraints of iterative GD, deep learning models, such as RNNs [72], are employed. These pre-trained models simulate the gradient descent process, enabling real-time generation of adversarial samples, thus enhancing efficiency.

Fast gradient sign methods Fast gradient sign method (FGSM), introduced by Goodfellow *et al.* [68], offers a faster approach to generate adversarial samples. It involves adding perturbations in the gradient direction of the original audio using a sign function, requiring only a single gradient calculation and addition/subtraction operation. Proven effective in the audio domain, FGSM can efficiently produce adversarial samples [73], [74].

Transferability of adversarial samples Given the limited access to commercial VCS models, attackers often rely on gray-box or black-box approaches, relying on querying the model and interpreting outputs. Transferability of adversarial samples between models becomes crucial in such scenarios. Studies in computer vision have shown the effectiveness of transferring adversarial samples across different models [87]–[90]. In audio, the success of transfer-based attacks is linked to the similarity between the parameters and structures of different models [91]–[93]. Li *et al.* [74] demonstrated this by successfully executing a transfer-based adversarial attack between i-vector and x-vector models, exploiting their structural similarities. Yao *et al.* [93] proposed a spectral transformation attack based on a modified discrete cosine transform to improve the transferability of adversarial voices to black-box victim models.

3) Characteristics of adversarial attacks in data layer

A key characteristic of adversarial attacks in the data layer is the indirect addition of perturbations. Rather than directly altering the audio signal, attackers inject perturbations into neural network features, such as those obtained from STFT or MFCC. For STFT-based approaches, benign audio is first converted into a spectrogram, and adversarial perturbations are added at this stage, followed by inverse transformation to generate adversarial audio [71], [72]. In contrast, while perturbing MFCC features can mitigate some challenges, the lossy nature of the inverse MFCC transformation leads to a significant degradation in the quality of adversarial audio [14], [15], [94].

4) Defense schemes against adversarial attacks

Adversarial training, a method proven effective in image processing, involves integrating adversarial examples into training data to enhance model robustness [95]. However, this approach has limitations:

i) It can lead to decreased accuracy in predicting legitimate examples due to label leakage [12].

ii) The effectiveness of adversarial training depends on the availability of a diverse set of known adversarial examples, making it less effective against novel attacks [13].

iii) This method increases training iterations, thereby raising the model's training cost.

Given that these attacks often require in-depth knowledge of the VCS's preprocessing techniques, they are generally categorized as white-box attacks. A counter-strategy involves modifying the targeted data preprocessing methods. For example, if attacks primarily target MFCC processing, alternative feature extraction techniques (e.g., neural networks) may be employed [16], [17]. Du *et al.* [96] designed a unified preprocessing mechanism to disrupt the continuity and transferability of adversarial attacks and an adaptive ASR transcription method to further bolster the defense strategy. Park *et al.* [97] demonstrated that carefully selected noise can considerably influence the transcription results of adversarial audio examples while having minimal impact on the transcription of benign audio waves. Introducing new neural network modules to the data layer may inadvertently create new vulnerabilities within the system.

V. Model Layer

In VCS, the model layer plays a crucial role in decoding audio feature vectors processed in the data layer. This decoding is traditionally achieved through an acoustic model, which translates feature vectors into phonemes, and a language model, predicting word sequence probabilities to output the most probable sentence. Historically, decoding methods have included hidden Markov models, recurrent neural networks, and N-gram models. However, the rise of deep learning technologies has led to advanced models like Wav2Letter (CNNs-based), Kaldi (DNN-HMM), and CMU Sphinx (GMM-HMM) becoming mainstream due to their enhanced accuracy in speech recognition [98]. Despite these advancements, the model layer remains vulnerable to adversarial attacks that manipulate the decoding process and spoofing attacks, where attackers mimic legitimate audio inputs to deceive the system, highlighting the need for robust security measures in this critical component of VCS.

1. Spoofing attacks in model layer

Spoofing attacks in the model layer of VCS primarily target the SV component. These attacks are executed with the intention of impersonating a target speaker to gain unauthorized access or privileges within the VCS system [99], [100]. Commonly, spoofing attacks are categorized into four types: replay, impersonation, voice synthesis, and voice conversion.

1) Replay

A replay attack involves recording a victim's voice using a recording device and then replaying this captured audio to deceive the SV system. This method is straightforward yet highly effective, posing a significant threat to various SV systems due to its simplicity and reliability. The effectiveness of replay attacks has been demonstrated against many types of SV systems [101]–[104].

2) Impersonation

Impersonation attacks are executed by an attacker manually imitating the voice patterns and speech behavior of the target individual, without the aid of any computerized devices. These attacks aim to trick the SV component into falsely recognizing the attacker as the target. Research has indicated that even non-professional impersonators can be successful, particularly if they are familiar with the target's voice pattern or possess a naturally similar voice [105], [106]. Professional impersonators often try to replicate rhythmic features of the target's speech to enhance the deception [107]. However, it has been observed that significant differences between the impersonator's and the target's natural voice can still be detected by SV systems [108], [109].

3) Voice synthesis

Voice Synthesis (VS) is a technology that aims to generate speech resembling the victim's voice from input text, distinct from standard TTS applications. We can also call voice synthesis as voice cloning [110] here. The goal is to deceive the SV component by impersonating the victim. Traditional VS techniques include unit selection [111], statistical parametric modeling [112], and hybrid approaches combining both [113]. Advancements in deep learning have led to the use of neural networks in VS, such as generative adversarial networks (GAN) for enhancing speech quality [114]–[116], and specific models like WaveNet [117], [118] and Tacotron [119] for victim-specific speech generation.

4) Voice conversion

Voice conversion (VC) differs from VS as it transforms the attacker's speech to match the victim's speech patterns. Traditional VC methods use statistical models such as HMM [120], GMM [121], principal component analysis (PCA) [122], and non-negative matrix factorization (NMF) [123]. Modern approaches leverage neural networks, including artificial neural networks (ANN) [124], WaveNet, and GANs. Recent research has focused on challenges such as low-resource languages [125], [126] and cross-lingual voice conversion [127]. Innovative techniques involving specially designed tubes have been explored to bypass liveness detection systems by modifying the attacker's voice to resemble the victim's voice [128].

5) Twins

Identical twins present a unique challenge in SV due to their highly similar speech characteristics. While SV systems are generally effective in distinguishing between different individuals, they struggle to differentiate between identical twins. Studies have indicated that twins share similar speech signal patterns, pitch contours, formant contours, and spectrograms [129], [130], leading to higher false acceptance rates (FAR) in SV. Incidents have been reported where twins were able to access each other's accounts by bypassing SV checks [131].

6) Defense schemes against spoofing attacks

Defense strategies against spoofing attacks in VCS primarily focus on identifying distortions in audio signals during transcription.

For replay attacks Techniques have been developed to distinguish between original and replayed audio by analyzing far-field recordings characteristic of distant attacks [103], [132]. Another approach involves maintaining a database of previous audio recordings in VCS and calculating similarity scores for new inputs to detect replays [133].

For voice synthesis Traditional VS detection methods often rely on understanding acoustic differences based on specific VS algorithms, such as analyzing spectral parameter ranges [134] and variations in mel-cepstral coefficients [135]. Differences in pitch patterns and vocal tract models between human-produced and synthesized audio offer additional detection markers [136]–[139]. For example, Liu *et al.* [139] designed

a defense scheme to protect the speaker's voice from voice synthesis attacks by modifying the target speaker's speeches in the frequency domain before publishing them or sending them to others. In addition, Zhang *et al.* [140] proposed a robust deepfake speech detection method that employs feature decomposition to learn synthesizer-independent content features.

For voice conversion Defenses against VC attacks draw inspiration from VS defense strategies, with studies showing that support vector machine (SVM) classifiers based on super-vectors demonstrate inherent robustness against VC [141], [142]. Analyzing pairwise distances between continuous feature vectors in human versus VC-generated audio has also proven effective.

Machine learning-based approaches State-of-the-art methods involve using machine learning models to discern spectral differences between human and generated audio [143]–[148]. The ASVspoof challenges have been instrumental in driving advancements in anti-spoofing technology, with RawNet2 emerging as a successful model in this domain [149]–[153]. Additionally, research that combines articulatory phonetics to reconstruct vocal tracts from audio signals has demonstrated potential in distinguishing between human and machine sources [154], [155]. A novel approach proposed in [156] shifts the focus from preemptively rejecting VC-generated audio to reconstructing the source speaker's voiceprint for attacker identification.

2. Adversarial attacks in model layer

The model layer of VCS, much like the data layer, is vulnerable to adversarial attacks. However, the attack methodology in the model layer typically involves directly adding adversarial perturbations to the audio waveforms of legitimate audio samples. This direct manipulation of audio waveforms distinguishes these attacks from those in the data layer. Adversarial attacks in the model layer pose a significant challenge to audio classifiers, often leading to misclassifications and subsequent security breaches. A detailed summary and comparative analysis of recent advancements in adversarial attacks targeting the model layer can be found in Table 4. This table provides insights into the various techniques and strategies employed in these attacks, underlining their potential impact on VCS performance and security.

1) Adversarial examples crafting

In the model layer of VCS, the shift towards black-box

Table 4 Comparison of voice-synthesis attacks and adversarial attacks in model layer. This table provides insights into the various techniques and strategies employed in these attacks, underlining their potential impact on VCS performance and security (N.A.: not avaliable)

Attack type	Article	Attack technology	Goal	Attack target (Category)	Model knowledge	Attack vector	Cost (Time)	Performance	
								Success rate (%)	Distance
Adversarial attacks	Schönherr *et al.* [87]	GD	Targeted	Kaldi (ASR)	White	WTA	2 min	98	N.A.
	Qin *et al.* [157]	GD	Targeted	Lingvo (ASR)	White	WTA	N.A.	100	N.A.
						WAA		60	
	Taori *et al.* [158]	GD+GA	Targeted	DeepSpeech (ASR)	Black	WTA	N.A.	35	N.A.
	Neekhara *et al.* [159]	GD	Targeted	2 model (ASR)	White	WTA	N.A.	89.6	N.A.
	Kwon *et al.* [160]	GD	Targeted	DeepSpeech (ASR)	White	WTA	1 h	91.7	N.A.
	Li *et al.* [79]	GD	Targeted	Amazon Alexa (ASR)	Gray	WAA	N.A.	50	7.7 ft
	Gong *et al.* [161]	RL+RNN	Untargeted	KWS (ASR)	Gray	WTA	0.15 s	43.5	N.A.
	Liu *et al.* [162]	GD	Targeted	DeepSpeech (ASR)	White	WTA	4 min–5 min	N.A.	N.A.
	Imperio [163]	GD	Targeted	Kaldi (ASR)	White	WAA	80 min	25	3 m
	Szurley *et al.* [164]	GD	Targeted	DeepSpeech (ASR)	White	WAA	N.A.	N.A.	6 in
	Han *et al.* [165]	GA	Untargeted	Google Cloud (ASR)	Black	WTA	N.A.	86	N.A.
	Khare *et al.* [166]	GA	Untargeted	2 models (ASR)	Black	WTA	N.A.	N.A.	N.A.
			Targeted						
	Wu *et al.* [167]	GA	Targeted	Kaldi (ASR)	Gray	WTA	N.A.	90	N.A.
	AudiDoS [168]	GD	Untargeted	2 models (ASR)	White	WAA	N.A.	78	30 cm
	Chen *et al.* [91]	GD	Targeted	4 models (ASR)	Black	WAA	N.A.	98	2 m
		Transfer		Apple Siri (ASR)				N.A.	N.A.
	Metamorph [169]	GD	Targeted	DeepSpeech (ASR)	White	WAA	5 h–7 h	90	6 m
	Wang *et al.* [170]	GAN	Targeted	2 models (ASR)	White	WTA	0.009 s	92.33	N.A.
	Wang *et al.* [171]	GA	Targeted	DeepSpeech (ASR)	Gray	WTA	N.A.	98	N.A.
	Advpulse [172]	GD	Targeted	KWS (ASR)	White	WAA	N.A.	89.9	3 m
				DNN x-vector (SV)				89.3	

Table 4 (Continued)

Attack type	Article	Attack technology	Goal	Attack target (Category)	Model knowledge	Attack vector	Cost (Time)	Performance	
								Success rate (%)	Distance
Adversarial attacks	Abdullah *et al.* [84]	SP	Untargeted	7 models (ASR)	Black	WTA	N.A.	100	N.A.
				Microsoft Azure (SV)				N.A.	
	SirenAttack [94]	PSO	Untargeted	DeepSpeech (ASR)	White	WTA	N.A.	100	N.A.
			Targeted	7 models (ASR)	Black			95.25	
				7 models (SV)				99.45	
	Chen *et al.* [173]	NES+BIM	Untargeted	3 models (SV)	Black	WAA	N.A.	100	2 m
			Targeted					95	
			Targeted	Talentedsoft (SV)		WTA		100	N.A.
		Transfer	Untargeted	Microsoft Azure (SV)		WAA		57	2 m
			Targeted					34	
	Wang *et al.* [174]	FGSM	Untargeted	GE2E (SV)	White	WTA	N.A.	N.A.	N.A.
	Xie *et al.* [175]	SP	Targeted	DNN x-vector (SV)	White	WAA	0.015 s	90	5 m
	Li *et al.* [176]	FGSM	Untargeted	DNN x-vector (SV)	White	WTA	N.A.	99	N.A.
						WAA		100	1 m
			Targeted			WTA		96.01	N.A.
						WAA		50	1 m
	Li *et al.* [177]	GN	Untargeted	SineNet (SV)	Gray	WTA	N.A.	N.A.	N.A.
			Targeted						
	Zhang *et al.* [178]	GD	Targeted	VGGVox (SV)	Gray	WTA	N.A.	100	N.A.
				Microsoft Azure (SV)	Black			75	
		Transfer		Apple Homekit (SV)		WAA		67.5	
	Li *et al.* [179]	ATN	Targeted	SineNet (SV)	White	WTA	0.042 RTF	N.A.	N.A.
	FoolHD [180]	GCA	Untargeted	DNN x-vector (SV)	White	WTA	N.A.	99.6	N.A.
			Targeted					99.2	
	SMACK [181]	GA + GE	Targeted	5 models (ASR)	Black	WTA	474 s	87.7	N.A.
						WAA	N.A.	43.3	200 cm
				2 models (SV)		WTA	525 s	99.2	N.A.
						WAA	N.A.	35.8	200 cm
	Zheng *et al.* [182]	CC-CMA-ES	Targeted	5 models (ASR)	Black	WTA	N.A.	100	N.A.
				2 models (SV)					
		DNN		5 models (ASR)		WAA		60	15 cm
	Liu *et al.* [183]	GE	Targeted	4 models (ASR)	Black	WNA	362.18 s	87.7	N.A.

Note: GD: gradient descent. GA: genetic algorithm. SP: signal processing. PSO: particle swarm optimization. NES: natural evolution strategy. BIM: basic iterative method. FGSM: fast gradient sign method. GN: generative network. ATN: adversarial transformation network. GCA: convolutional auto-encoder. CC-CMA-ES: cooperative-coevolution covariance-matrix-adaptation evolution-strategy. GE: gradient estimation. RTF: real-time factor, the ratio of the processing time to the input duration. N.A.: not avaliable.

models has made traditional gradient-based adversarial sample generation methods like GD and FGSM less effective. To address this, researchers have developed gradient-agnostic algorithms for crafting adversarial examples.

Genetic algorithm (GA) GA employs a search-based approach, iterating through potential samples to deceive machine learning models. By using techniques like crossover and mutation, GA refines search results, prioritizing samples with higher fitness scores for subsequent iterations. This method is suitable for gray-box or black-box attacks due to its reliance solely on model output rather than gradient information [158], [165]–[167], [171]. However, GA's iterative nature leads to high computational resource demands.

Particle swarm optimization (PSO) Originating from swarm intelligence, PSO mimics collective animal behaviors to search for optimal solutions [184]. In adversarial example generation, samples are treated as swarm particles, moving towards the most effective position in the search space iteratively [94]. PSO benefits from fast convergence and does not require gradient information, making it efficient for gray-box and black-box attacks.

Ensemble algorithm Chen *et al.* [173] introduced a novel black-box approach combining the natural evolution strategy (NES) [185] and the basic iterative method (BIM) [186]. NES estimates the gradient of the black-box model, and this estimated gradient is then used in BIM to generate adversarial examples iteratively. This method efficiently crafts adversarial examples suitable for black-box settings. Fang *et al.* [187] proposed a sequential ensemble optimization algorithm, which iteratively optimizes the adversarial perturbation on each surrogate model, leveraging collaborative information from other models. This is a transfer-based adversarial attack on ASR systems in the zero-query black-box setting.

Deep learning-based methods Approaches like the gated convolutional auto-encoder (GCA) [180], adversarial transformation network (ATN) [179], and generative network (GN) [177], [188] employ deep learning for adversarial example generation. GN is effective in gray-box model attacks, while ATN significantly reduces the time required to generate adversarial examples. Zuo *et al.* [189] designed an adversary sample attack on the speaker identification system using multiple utterances.

Remark 1 Communication and video conferencing through IP have become the primary means of community communication nowadays. In the work in [183], researchers discovered that due to the special characteristics of IP channels, audio signal transmission can cause signal attenuation, random channel noise, and other interferences, which can affect the effectiveness of audio adversarial samples. Therefore, in future research, it is necessary to focus on investigating the potential impact of malicious audio in the transmission process over IP channels. In this paper, we refer to attacks that are transmitted through IP channels as wav-network-API (WNA).

2) Defense schemes against adversarial attacks

The nature of attacks targeting the model layer demands additional strategies. Since these attacks are typically directed at the model itself, the transferability of adversarial examples between different models tends to be limited. Zeng *et al.* [18] observed that adversarial examples crafted for one specific model often yield different prediction results when applied to another model, a discrepancy not seen with normal audio signals. This unique behavior of adversarial examples provides a basis for detecting and countering attacks in the model layer. By exploiting the differences in model responses to adversarial versus normal audio, defense mechanisms can be designed to effectively identify and mitigate adversarial attacks targeting this crucial layer of VCS.

VI. Service Layer

The evolution of VCS has led to the integration of third-party services, enhancing functionality beyond basic speech content recognition or identity verification. Examples include Amazon Echo and Google Assistant. Although this expansion addresses user demands, it also introduces more security risks. This section explores the potential vulnerabilities and threats posed by these third-party integrations in VCS.

1. Squatting attacks in service layer

Despite mandatory reviews for third-party services in VCS to ensure compliance with security and privacy policies, many services still breach these regulations [190], [191], and current review processes struggle to identify such violations [192]. This oversight allows attackers to introduce malicious services, posing significant privacy risks. Squatting attacks occur when attackers create services with names resembling legitimate ones, deceiving users into accessing these malicious options instead. There are two primary methods of squatting attacks:

i) Similar sounding service names: Attackers may design service names phonetically similar to legitimate ones, leading to accidental activation of the malicious service. For instance, a service named "Capital Won" could be mistaken for "Capital One", especially in noisy settings or based on user dialect or gender [3], [4].

ii) Exploiting the longest matching rule: VCS typically matches service names using the longest matching rule. Attackers leverage this and user habits to construct squatting attacks. Since many users add polite terms like "please" to service commands, a malicious service named "Capital One Please" could be preferentially activated over the intended "Capital One" service, as observed in user habit surveys [4].

These squatting tactics demonstrate the vulnerabilities introduced by third-party service integration in VCS, highlighting the need for more robust security measures.

Misunderstanding attacks A misunderstanding attack is a specific form of squatting attack where the goal is not to activate a malicious service, but to induce errors within a legitimate service. These attacks exploit the limitations of VCS in accurately interpreting words outside their programmed understanding [193]. For example, Siri might misinterpret "bank" in the context of "river bank". Bispham *et al.* [61] demonstrated that by altering a single word in a command, VCS can be misled. In a banking service, changing "money" to "dough" in the query "how much money do I have?" could result in unintended information disclosure due to "dough" being slang for money. The NLU's Intent Classifier, which uses fuzzy matching to accommodate user speech variations, is often the exploitable component in such scenarios [194].

Defense schemes against squatting attacks Squatting attacks often involve similar-sounding service names or variants to trick VCS. Zhang *et al.* [4] developed a phoneme-based service name scanner to identify suspicious third-party services. Additionally, it is suggested that service platform providers analyze new services for potential naming conflicts or utilize interactive tools for policy violation detection [3], [195].

2. Faking termination in service layer

Users' misunderstandings about service termination protocols in VCS like Alexa and Google Assistant can be exploited. Users often assume that services end automatically after silence or a farewell response. However, malicious services can simulate termination sounds while remaining active, posing significant security risks if users then disclose private information or issue commands. Such deceptive practices highlight the need for greater awareness and more explicit termination protocols in VCS [4].

Defense schemes against faking termination To mitigate the risks associated with faking termination in VCS, various studies recommend enhancing service publishing authentication mechanisms [190], [192]. Service platform providers should conduct thorough word-based or phoneme-based reviews of new services to avoid confusion with existing ones. Additionally, methods such as voice interaction analysis and backend code review are effective in identifying privacy breaches or discrepancies between a service's functionality and its description. Recent research suggests that context-sensitive detectors could effectively identify instances of faking termination [4]. Notably, a proposed hardening framework for VCS in [196] focuses on ensuring that data collected by microphones is used only for local processing unless explicitly permitted by the user, thereby safeguarding against privacy data leakage caused by faking termination.

VII. General Attack Mitigation Strategy

In earlier sections, we discuss defense strategies that primarily focus on countering specific types of attacks in VCS. This section shifts the focus to state-of-the-art general attack mitigation strategies that are capable of defending against more attacks, encompassing different types and layers. These comprehensive strategies are essential for providing a more holistic defense mechanism for VCS. We have systematized existing general attack mitigation strategies and evaluated their efficacy against various attack types and characteristics. The comparative analysis of these strategies, including their strengths and limitations in resisting different forms of attacks, is presented in Table 5. This systematic overview aims to guide researchers and practitioners in selecting and implementing effective, multi-faceted defense strategies for VCS.

Table 5 Capability of defending against different attacks in different layers

Defense schemes		Attacks								Papers
		Physical layer		Data layer		Model layer		Service layer		
		TA	VsA	VsA	AA	SpA	AA	SqA	FT	
Liveness detection	Speaker characteristics	⊗	⊕	⊕	⊕	⊕	⊕	⊗	⊗	[197]–[202]
	Human voice characteristics	⊕	⊕	⊕	⊕	⊕	⊕	⊗	⊗	[200], [203]–[216]
	Active interaction	⊕	⊕	⊕	⊕	⊕	⊕	⊙	⊙	[1], [39], [217], [218]
Audio conversion	Encoding	⊗	⊙	⊕	⊕	⊗	⊕	⊗	⊗	[219]–[222]
	Filtering	⊗	⊙	⊕	⊕	⊗	⊕	⊗	⊗	[91], [94], [223]–[225]
	Downsampling	⊗	⊙	⊕	⊕	⊗	⊕	⊗	⊗	[39], [78], [91], [94], [223], [224]

Note: TA: transduction attack. VsA: voice-synthesis attack. AA: adversarial attack. SpA: spoofing attack. SqA: squatting attack. FT: faking termination. ⊕: Defensible. ⊙: Probably Defensible. ⊗: Indefensible.

1. Liveness detection

Liveness detection has emerged as a prevalent defense strategy in VCS. It is primarily designed to ascertain whether voice commands originate from actual humans [226]–[229]. The fundamental premise behind this approach is the recognition that most malicious commands are machine-generated. These commands are typically played through speakers or directly input into the VCS API via audio files, such as WAV files. Unlike these artificially produced commands, genuine human users do not generate voice commands in this manner. Therefore, by identifying the characteristics of human speech, liveness detection aims to filter out these non-human, machine-generated inputs, thereby enhancing the security of VCS.

Passive detection In the context of liveness detection for VCS, passive detection plays a crucial role in distinguishing between human-spoken voice commands and those generated by speakers. This is achieved through the analysis of sound characteristics using two main techniques:

i) Detecting speaker characteristics: Voice commands emitted from speakers often carry unique signal distortions, which are a result of circuit noise inherent to the speaker's hardware. These distortions differ significantly from the patterns found in human speech and can be identified using classifiers trained specifically for this purpose [197]–[201]. Moreover, the electromagnetic fields generated by speakers during sound emission, owing to the electronic components, can be detected using a magnetometer in smart devices, further aiding in the determination of the voice command's origin [202].

ii) Detecting human voice characteristics: Human speech is produced through a complex physiological process involving the coordinated action of the mouth, vocal tract, vocal cord, and lungs. The airflow from the

lungs, passing through the glottis, causes vocal cord vibrations, which are then amplified by resonances in the mouth and vocal tract to form the final sound signal. Identifying features inherent to this process, such as breathing airflow patterns [200], [203]–[206], mouth movements [207]–[209], [216], as well as bone vibrations [210]–[215], [230], provides a basis for determining whether a voice command is human-generated. These features can be monitored using microphones, cameras, or other specialized sensors. In practical applications, integrating additional devices or sensors might be necessary to enhance the accuracy and reliability of such verifications.

These passive detection methods are instrumental in bolstering the security of VCS by ensuring that voice commands are genuinely human and not artificially generated or replayed through electronic devices.

Active interaction The active interaction defense scheme in VCS involves engaging users in a manner similar to a CAPTCHA to ascertain the authenticity of voice commands. A prevalent form of this scheme is the Challenge-Response mechanism. Upon receiving a voice command, VCS issues a challenge to the user, requiring an appropriate response within a predetermined time frame. Failure to respond correctly within this window leads to the assumption that the command is machine-generated, and thus, the command execution is denied [1], [39], [217]. Although effective in thwarting voice attacks to an extent, this method introduces additional steps for the user, potentially compromising the usability of the VCS.

Remark 2 Recent research [128] has demonstrated that malicious attacks can manipulate a speaker's voice commands into harmful instructions using a specially designed tube. As these commands are essentially uttered by a real person, they can effectively circumvent traditional liveness detection methods. Future research in VCS security must account for the implications of such attacks.

2. Audio conversion

The conversion of audio in the data layer of VCS, before it is passed to subsequent layers for further processing, serves as a defensive measure against voice-synthesis and adversarial attacks [83], [231]. This effectiveness stems from the conversion process's ability to disrupt the specific patterns and structures that these attacks aim to exploit or deceive. As a result, the converted audio loses the characteristics intended by the attack, rendering it ineffective. In contrast, benign audio typically exhibits greater resilience to these conversions and is only minimally affected, preserving its integrity while mitigating potential threats.

Audio encoding Encoding incoming audio has been shown to effectively diminish the success rate of malicious audio attacks. Utilizing codecs like advanced audio coding (AAC) [221], MP3 [219]–[222], Speex [221], Opus [221], adaptive multi-rate (AMR) [220], and integrated multiple codecs [218], [220] can provide a substantial level of defense against malicious audio.

Audio filtering Voice-synthesis and adversarial attacks rely heavily on precise algorithmic perturbations. Filtering these malicious audio inputs, using methods such as median filters [91], [94], [223], quantization [223], and other noise reduction algorithms [224], [225], effectively destroy these perturbations, thereby safeguarding the VCS from such attacks.

Audio downsampling Experiments have demonstrated that downsampling audio to a lower rate and then upsampling it back to a rate suitable for VCS input can effectively mitigate attack impacts. Although benign audio remains relatively unaffected, the malicious audio loses its carefully added perturbations, thus failing to achieve the intended effect on the target VCS [39], [78], [91], [94], [223], [224].

Remark 3 Audio conversion as a defense strategy exploits the resilience of benign audio to counteract the fragility of malicious audio. This process disrupts the attacker's carefully engineered perturbations. However, in a white-box scenario, attackers aware of the VCS's audio conversion methods can adjust their strategies accordingly, potentially neutralizing the defense's effectiveness.

Remark 4 As delineated in Table 5, liveness detection effectively counters a wide array of attacks, particularly those outside the service layer. Although active interaction also offers some resistance against service layer attacks, its effectiveness comes at the cost of user experience. Given this trade-off, a preferable approach would be to combine passive detection with an enhanced service release and authentication mechanism. This combination would enable a more holistic defense against the diverse types of attacks encountered in VCS, thereby optimizing both security and user experience.

VIII. Challenges and Future Directions

1. Challenges

Hardware enhancement The effectiveness of physical layer attacks in VCS often hinges on exploiting hardware vulnerabilities, particularly in microphones. However, these vulnerabilities are not uniformly present across all microphone types. A notable example is the iPhone 6 Plus, which has been shown to effectively resist voice-synthesis attacks due to its unique microphone design, as highlighted in [55]. This variability in hardware susceptibility presents significant challenges for executing consistent attacks at the physical layer.

Noise disturbance Noise disturbance is a critical factor for both attackers and defenders in real-world VCS. For attackers, ambient noise can diminish the effectiveness and reach of malicious audio. Conversely, for defenders, noise can interfere with the accuracy of defense mechanisms, such as liveness detection systems, as demonstrated in [232]. Thus, both parties must account for the impact of noise in their strategies, which

add another layer of complexity to the security landscape of VCS.

Data security VCS has permission to override other apps to perform specific functions. For example, the voice assistant will call the map application when it receives a command to navigate. As a result, VCS has more powerful access to data. It is possible for an attacker to utilize the voice assistant to access the user's sensitive data.

Model knowledge As VCS technology has evolved, commercial models have become increasingly prevalent. These models are often proprietary and not open-sourced, a strategy employed by companies to protect their intellectual property and prevent replication by competitors. This secrecy forces attackers to operate in a black-box environment, substantially lowering the success rate of adversarial attacks. Additionally, the challenge of creating universal adversarial perturbations that can produce similar attack outcomes across different models remains a significant hurdle in the realm of adversarial attacks.

2. Future directions

The challenges identified in VCS security research open avenues for future work. To enhance the robustness of both attacks and defenses in practical scenarios, we propose following directions:

Focusing on black-box attack scenarios With the rise of closed-source models in VCS, attackers in real-world scenarios should concentrate on executing successful black-box attacks. This could involve crafting adversarial perturbations specifically for black-box models or designing transferable adversarial perturbations that can be applied from a known white-box model to an unknown black-box model [233].

Targeting combined ASR and SV functions Modern VCS, such as Apple's Siri, often integrates both ASR and SV functionalities. Attackers, therefore, need to develop malicious audio capable of simultaneously compromising both ASR and SV systems. This may require the help of more targeted deep learning techniques and more detailed spectrum modification techniques.

Optimizing defense schemes With the advancement of artificial intelligence technology, techniques such as voice synthesis technology have become very inexpensive and easy to use. For example, voice cloning can be done directly online using a piece of audio. To defend against voice cloning attacks, we can take different measures at different layers to achieve cross-enhanced defense. In the physical layer, we can design and utilize unique physical structures such as microphone array fingerprints [232]. In the data layer, we have to protect our voice data from leakage or take appropriate measures before release, such as spectrum modification [139]. In the model layer, designing and optimizing models centered on adversarial learning is a viable defense strategy. In the service layer, we believe that we need to prevent the random dissemination of data across different applications. To this end, we can design unique data dissemination protocols or data calling interfaces to prevent the leakage of speech-sensitive information (e.g., tones).

Deepening multimodal defense schemes Developing authentication mechanisms for legitimate users and legitimate commands based on multimodality. Improve voice control system defense by combining human biometrics and instantaneous verification of trusted devices instead of relying only on a single voice command.

Establishing unified standards for evaluation The development of a unified standard for assessing attacks and defense mechanisms in VCS is crucial. This standard would provide reliable and consistent evaluation metrics, akin to the common vulnerability scoring system (CVSS), aiding VCS designers in accurately assessing the security landscape.

These proposed directions aim to address the evolving landscape of VCS security, encouraging a balanced approach to unveiling sophisticated attacks and designing robust defenses.

IX. Conclusion

This paper offers a thorough and methodical examination of the security landscape in VCS. We introduce a hierarchical model structure, dividing VCS into four layers: physical layer, data layer, model layer, and service layer. Within this framework, we develop a threat model and meticulously analyze the security threats unique to each layer, including transduction, voice-synthesis, adversarial, spoofing, squatting attacks, and false termination. In particular, we perform a multi-layer cross-tabulation analysis of voice synthesis attacks, which reveals the methods and mitigation measures of voice synthesis attacks. We not only identify and describe the characteristics of these threats but also assess existing defense mechanisms and propose effective defense strategy combinations to improve VCS security. Finally, we offer strategic recommendations for future research, emphasizing the need for continuous innovation to address the rapidly evolving security challenges in VCS.

Acknowledgements

This work was supported by the National Key R&D Program of China (Grant Nos. 2022YFB3102100 and 2022YFB3103300), the National Natural Science Foundation of China (Grants Nos. 62172303, 62302343, and 62472323), the Key R&D Program of Hubei Province (Grant No. 2024BAB018), the Wuhan Scientific and Technical Achievements Project (Grand No. 2024030803 010172), the National Cryptologic Science Fund of China (Grant No. 2025NCSF02027), and the Key R&D Program of Shandong Province (Grant No. 2022CXPT055).

References

[1] T. Sugawara, B. Cyr, S. Rampazzi, *et al.*, "Light commands: Laser-based audio injection attacks on voice-con-

trollable systems," in *Proceedings of the 29th USENIX Conference on Security Symposium*, article no. 148, 2020.

[2] G. M. Zhang, X. H. Ma, H. T. Zhang, *et al.*, "LaserAdv: Laser adversarial attacks on speech recognition systems," in *Proceedings of the 33rd USENIX Conference on Security Symposium*, Philadelphia, PA, USA, article no. 221, 2024.

[3] D. Kumar, R. Paccagnella, P. Murley, *et al.*, "Skill squatting attacks on Amazon Alexa," in *Proceedings of the 27th USENIX Conference on Security Symposium*, Baltimore, MD, USA, pp. 33–47, 2018.

[4] N. Zhang, X. H. Mi, X. Feng, *et al.*, "Dangerous skills: Understanding and mitigating security risks of voice-controlled third-party functions on virtual personal assistant systems," in *Proceedings of the 2019 IEEE Symposium on Security and Privacy*, San Francisco, CA, USA, pp. 1381–1396, 2019.

[5] J. S. Edu, J. M. Such, and G. Suarez-Tangil, "Smart home personal assistants: A security and privacy review," *ACM Computing Surveys (CSUR)*, vol. 53, no. 6, article no. 116, 2021.

[6] Y. Bai, L. Lu, J. Cheng, *et al.*, "Acoustic-based sensing and applications: A survey," *Computer Networks*, vol. 181, article no. 107447, 2020.

[7] H. Abdullah, K. Warren, V. Bindschaedler, *et al.*, "SoK: The faults in our ASRs: An overview of attacks against automatic speech recognition and speaker identification systems," in *Proceedings of the 2021 IEEE Symposium on Security and Privacy*, San Francisco, CA, USA, pp. 730–747, 2021.

[8] Y. X. Chen, J. S. Zhang, X. J. Yuan, *et al.*, "SoK: A modularized approach to study the security of automatic speech recognition systems," *ACM Transactions on Privacy and Security*, vol. 25, no. 3, article no. 17, 2022.

[9] X. Zhang, H. Tan, X. Huang, *et al.*, "Adversarial example attacks against ASR systems: An overview," in *Proceedings of the 2022 7th IEEE International Conference on Data Science in Cyberspace*, Guilin, China, pp. 470–477, 2022.

[10] I. Gurowiec and N. Nissim, "Speech emotion recognition systems and their security aspects," *Artificial Intelligence Review*, vol. 57, no. 6, article no. 148, 2024.

[11] C. Yan, X. Y. Ji, K. Wang, *et al.*, "A survey on voice assistant security: Attacks and countermeasures," *ACM Computing Surveys*, vol. 55, no. 4, article no. 84, 2023.

[12] J. Wang, "Adversarial examples in physical world," in *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, pp. 4925–4926, 2021.

[13] Y. Gong and C. Poellabauer, "An overview of vulnerabilities of voice controlled systems," *arXiv preprint*, arXiv: 1803.09156, 2018.

[14] D. Iter, J. Huang, and M. Jermann, "Generating adversarial examples for speech recognition," *Stanford Technical Report*, 2017, pp. 1–8.

[15] H. Yakura and J. Sakuma, "Robust audio adversarial example for a physical attack," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, Macao, China, pp. 5334–5341, 2019.

[16] O. Abdel-Hamid, A. R. Mohamed, H. Jiang, *et al.*, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[17] J. Hai and E. M. Joo, "Improved linear predictive coding method for speech recognition," in *Proceedings of the 4th International Conference on Information, Communications and Signal Processing, 2003 and the 4th Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint*, Singapore, Singapore, pp. 1614–1618, 2003.

[18] Q. Zeng, J. H. Su, C. L. Fu, *et al.*, "A multiversion programming inspired approach to detecting audio adversarial examples," in *Proceedings of the 2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, Portland, OR, USA, pp. 39–51, 2019.

[19] Z. X. He, J. Chen, K. He, *et al.*, "HeadSonic: Usable bone conduction earphone authentication via head-conducted sounds," *IEEE Transactions on Mobile Computing*, vol. 24, no. 9, pp. 7914–7928, 2025.

[20] Y. Y. Gu, J. Chen, K. He, *et al.*, "WiFiLeaks: Exposing stationary human presence through a wall with commodity mobile devices," *IEEE Transactions on Mobile Computing*, vol. 23, no. 6, pp. 6997–7011, 2024.

[21] Y. Y. Gu, J. Chen, C. Wu, *et al.*, "LocCams: An efficient and robust approach for detecting and localizing hidden wireless cameras via commodity devices," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 4, article no. 160, 2023.

[22] T. Deng, B. W. Zheng, R. Du, *et al.*, "A statistical sensing method by utilizing Wi-Fi CSI subcarriers: Empirical study and performance enhancement," *Journal of Information and Intelligence*, vol. 2, no. 4, pp. 365–374, 2024.

[23] X. R. Gu, S. F. Wang, Z. Q. Wei, *et al.*, "Cluster-based RSU deployment strategy for vehicular ad hoc networks with integration of communication, sensing and computing," *Journal of Information and Intelligence*, vol. 2, no. 4, pp. 325–338, 2024.

[24] R. J. Sun, Y. Wen, N. Cheng, *et al.*, "Structural knowledge-driven meta-learning for task offloading in vehicular networks with integrated communications, sensing and computing," *Journal of Information and Intelligence*, vol. 2, no. 4, pp. 302–324, 2024.

[25] S. A. Zawawi, A. A. Hamzah, B. Y. Majlis, *et al.*, "A review of MEMS capacitive microphones," *Micromachines*, vol. 11, no. 5, article no. 484, 2020.

[26] N. Ryant, M. Liberman, and J. H. Yuan, "Speech activity detection on YouTube using deep neural networks," in *Proceedings of the Interspeech*, Lyon, France, pp. 728–731, 2013.

[27] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, New York, 1993.

[28] S. Sigurdsson, K. B. Petersen, and T. Lehn-Schiøler, "Mel frequency cepstral coefficients: An evaluation of robustness of MP3 encoded music," in *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR)*, Victoria, Canada, pp. 286–289, 2006.

[29] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice-Hall, New York, 1978.

[30] T. N. Sainath, O. Vinyals, A. Senior, *et al.*, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, South Brisbane, QLD, Australia, pp. 4580–4584, 2015.

[31] A. Graves, S. Fernández, F. Gomez, *et al.*, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, USA, pp. 369–376, 2006.

[32] A. Graves, "Sequence transduction with recurrent neural networks," in *Proceedings of the International Conference of Machine Learning (ICML) 2012 Workshop on Representation Learning*, Edinburgh, Scotland, pp. 1–9, 2012.

[33] J. Chorowski, D. Bahdanau, D. Serdyuk, *et al.*, "Attention-based models for speech recognition," in *Proceedings of the 29th International Conference on Neural Information Processing Systems*, Montreal, Canada, pp. 577–585, 2015.

[34] T. Mikolov, K. Chen, G. Corrado, *et al.*, "Efficient estimation of word representations in vector space," in *Proceedings of the 1st International Conference on Learning*

Representations, ICLR 2013, Scottsdale, AZ, USA, pp. 1–12, 2013.

[35] J. Devlin, M. Chang, K. Lee, *et al.*, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, MN, USA, pp. 4171–4186, 2019.

[36] X. H. He, J. J. Chen, Z. X. Zhang, *et al.*, "ProsodyFM: Unsupervised phrasing and intonation control for intelligible speech synthesis," in *Proceedings of the 39th AAAI Conference on Artificial Intelligence and 37th Conference on Innovative Applications of Artificial Intelligence and 15th Symposium on Educational Advances in Artificial Intelligence*, Philadelphia, PA, USA, article no. 2680, 2025.

[37] R. C. Liang, J. Chen, C. Wu, *et al.*, "Vulseye: Detect smart contract vulnerabilities via stateful directed gray-box fuzzing," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 2157–2170, 2025.

[38] G. M. Zhang, C. Yan, X. Y. Ji, *et al.*, "DolphinAttack: Inaudible voice commands," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, Dallas, TX, USA, pp. 103–117, 2017.

[39] N. Carlini, P. Mishra, T. Vaidya, *et al.*, "Hidden voice commands," in *Proceedings of the 25th USENIX Conference on Security Symposium*, Austin, TX, USA, pp. 513–530, 2016.

[40] C. Kasmi and J. L. Esteves, "IEMI threats for information security: Remote command injection on modern smartphones," *IEEE Transactions on Electromagnetic Compatibility*, vol. 57, no. 6, pp. 1752–1755, 2015.

[41] Y. D. Wang, H. Q. Guo, and Q. B. Yan, "GhostTalk: Interactive attack on smartphone voice system through power line," in *Proceedings of the 29th Annual Network and Distributed System Security Symposium, NDSS 2022*, San Diego, CA, USA, pp. 1–15, 2022.

[42] D. H. Dai, Z. L. An, and L. Yang, "Inducing wireless chargers to voice out for inaudible command attacks," in *Proceedings of the 2023 IEEE Symposium on Security and Privacy*, San Francisco, CA, USA, pp. 1789–1806, 2023.

[43] D. Davidson, H. Wu, R. Jellinek, *et al.*, "Controlling UAVs with sensor input spoofing attacks," in *Proceedings of the 10th USENIX Workshop on Offensive Technologies*, Austin, TX, USA, pp. 221–231, 2016.

[44] Z. Wang, Q. B. Zou, Q. L. Song, *et al.*, "The era of silicon MEMS microphone and look beyond," in *Proceedings of the 2015 Transducers - 2015 18th International Conference on Solid-State Sensors, Actuators and Microsystems*, Anchorage, AK, USA, pp. 375–378, 2015.

[45] Q. B. Yan, K. H. Liu, Q. Zhou, *et al.*, "SurfingAttack: Interactive hidden attack on voice assistants using ultrasonic guided waves," in *Proceedings of the Network and Distributed Systems Security Symposium*, San Diego, CA, USA, pp. 1–18, 2020.

[46] Q. Xia, Q. Chen, and S. H. Xu, "Near-ultrasound inaudible Trojan (NUIT): Exploiting your speaker to attack your microphone," in *Proceedings of the 32nd USENIX Conference on Security Symposium*, Anaheim, CA, USA, article no. 257, 2023.

[47] R. Iijima, S. Minami, Y. A. Zhou, *et al.*, "Audio hotspot attack: An attack on voice assistance systems using directional sound beams," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, Toronto, Canada, pp. 2222–2224, 2018.

[48] X. Y. Ji, J. C. Zhang, S. Jiang, *et al.*, "CapSpeaker: Injecting voices to microphones via capacitors," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, Virtual Event, pp. 1915–1929, 2021.

[49] N. Roy, S. Shen, H. Hassanieh, *et al.*, "Inaudible voice commands: The long-range attack and defense," in *Proceedings of the 15th USENIX Conference on Networked Systems Design and Implementation*, Renton, WA, USA, pp. 547–560, 2018.

[50] L. W. Song and P. Mittal, "POSTER: Inaudible voice commands," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, Dallas, Texas, USA, pp. 2583–2585, 2017.

[51] C. Yan, G. M. Zhang, X. Y. Ji, *et al.*, "The feasibility of injecting inaudible voice commands to voice assistants," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 3, pp. 1108–1124, 2021.

[52] H. F. Sun, H. H. Du, X. J. Yu, *et al.*, "IUAC: Inaudible universal adversarial attacks against smart speakers," *ACM Transactions on Sensor Networks*, vol. 21, no. 1, article no. 1, 2025.

[53] Z. Y. Ning, J. He, Z. Y. Tang, *et al.*, "A portable and stealthy inaudible voice attack based on acoustic metamaterials," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 8876–8890, 2025.

[54] L. Q. Yang, X. Q. Chen, X. Y. Jian, *et al.*, "Remote attacks on speech recognition systems using sound from power supply," in *Proceedings of the 32nd USENIX Conference on Security Symposium*, Anaheim, CA, USA, article no. 256, 2023.

[55] Y. T. He, J. Y. Bian, X. Y. Tong, *et al.*, "Canceling inaudible voice commands against voice control systems," in *Proceedings of the 25th International Conference on Mobile Computing and Networking*, Los Cabos, Mexico, article no. 28, 2019.

[56] Z. Y. Yu, "Towards proactive protection against unauthorized speech synthesis," in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, Salt Lake City, UT, USA, pp. 5128–5130, 2024.

[57] X. F. Li, X. Y. Ji, C. Yan, *et al.*, "Learning normality is enough: A software-based mitigation against inaudible voice attacks," in *Proceedings of the 32nd USENIX Conference on Security Symposium*, Anaheim, CA, USA, article no. 138, 2023.

[58] G. M. Zhang, X. Y. Ji, X. F. Li, *et al.*, "EarArray: Defending against dolphinattack via acoustic attenuation," in *Proceedings of the Network and Distributed Systems Security (NDSS)*, Virtual, pp. 1–14, 2021.

[59] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[60] A. Mumuni and F. Mumuni, "Automated data processing and feature engineering for deep learning and big data applications: A survey," *Journal of Information and Intelligence*, vol. 3, no. 2, pp. 113–153, 2025.

[61] M. K. Bispham, I. Agrafiotis, and M. Goldsmith, "Nonsense attacks on Google assistant and missense attacks on amazon alexa," in *Proceedings of the 5th International Conference on Information Systems Security and Privacy*, Setubal, Portugal, pp. 75–87, 2019.

[62] H. Abdullah, W. Garcia, C. Peeters, *et al.*, "Practical hidden voice attacks against speech and speaker recognition systems," in *Proceedings of the 26th Annual Network and Distributed System Security Symposium*, San Diego, CA, USA, pp. 1–15, 2019.

[63] T. Vaidya, Y. K. Zhang, M. Sherr, *et al.*, "Cocaine noodles: Exploiting the gap between human and machine speech recognition," in *Proceedings of the 9th USENIX Conference on Offensive Technologies*, Washington, DC, USA, article no. 16, 2015.

[64] X. Y. Ji, Q. H. Jiang, C. H. Li, *et al.*, "Watch your speed:

Injecting malicious voice commands via time-scale modification," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 3366–3379, 2024.

[65] S. Hollier, J. Sajka, J. White, *et al.*, "Inaccessibility of CAPTCHA: Alternatives to visual turing tests on the web," Available at: https://www.w3.org/TR/2021/DNOTE-turingtest-20211216/, 2021-12-16.

[66] B. Hartpence, *Packet Guide to Voice over IP: A System Administrator's Guide to VoIP Technologies,* O'Reilly Media, Sebastopol, pp. 1–239, 2013.

[67] J. Ramírez, J. M. Górriz, and J. C. Segura, "Voice activity detection. fundamentals and speech recognition system robustness," in *Robust Speech Recognition and Understanding*, M. Grimm and K. Kroschel, Eds. Intechopen, Vienna, Austria, pp. 1–24, 2007.

[68] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, USA, pp. 1–11, 2015.

[69] X. J. Yuan, J. S. Zhang, K. Chen, *et al.*, "Adversarial attack and defense for commercial black-box Chinese-English speech recognition systems," *ACM Transactions on Privacy and Security*, vol. 28, no. 1, article no. 10, 2025.

[70] C. Wu, J. F. Sun, J. Chen, *et al.*, "TCG-IDS: Robust network intrusion detection via temporal contrastive graph learning," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 1475–1486, 2025.

[71] M. Cisse, Y. Adi, N. Neverova, *et al.*, "Houdini: Fooling deep structured visual and speech recognition models with adversarial examples," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, CA, USA, pp. 6980–6990, 2017.

[72] K. H. Chang, P. H. Huang, H. G. Yu, *et al.*, "Audio adversarial examples generation with recurrent neural networks," in *Proceedings of the 2020 25th Asia and South Pacific Design Automation Conference*, Beijing, China, pp. 488–493, 2020.

[73] F. Kreuk, Y. Adi, M. Cisse, *et al.*, "Fooling end-to-end speaker verification with adversarial examples," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, AB, Canada, pp. 1962–1966, 2018.

[74] X. Li, J. H. Zhong, X. X. Wu, *et al.*, "Adversarial attacks on GMM I-vector based speaker verification systems," in *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, Barcelona, Spain, pp. 6579–6583, 2020.

[75] M. Marras, P. Korus, N. D. Memon, *et al.*, "Adversarial optimization for dictionary attacks on speaker verification," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, pp. 2913–2917, 2019.

[76] Q. Wang, P. C. Guo, and L. Xie, "Inaudible adversarial perturbations for targeted attack in speaker recognition," in *Proceedings of the 21st Annual Conference of the International Speech Communication Association, Interspeech 2020*, Shanghai, China, pp. 4228–4232, 2020.

[77] H. Q. Guo, Y. D. Wang, N. Ivanov, *et al.*, "SPECPATCH: Human-in-the-loop adversarial audio spectrogram patch attack on speech recognition," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, Los Angeles, CA, USA, pp. 1353–1366, 2022.

[78] X. J. Yuan, Y. X. Chen, Y. Zhao, *et al.*, "Commandersong: A systematic approach for practical adversarial voice recognition," in *Proceedings of the 27th USENIX Conference on Security Symposium*, Baltimore, MD, USA, pp. 49–64, 2018.

[79] J. B. Li, S. H. Qu, X. J. Li, *et al.*, "Adversarial music: Real world audio adversary against wake-word detection system," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Vancouver, BC, Canada, article no. 1068, 2019.

[80] M. Alzantot, B. Balaji, and M. Srivastava, "Did you hear that? Adversarial examples against automatic speech recognition," *arXiv preprint*, arXiv: 1801.00554, 2018.

[81] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *Proceedings of the IEEE Security and Privacy Workshops*, San Francisco, CA, USA, pp. 1–7, 2018.

[82] J. Vadillo and R. Santana, "Universal adversarial examples in speech command classification," *arXiv preprint*, arXiv: 1911.10182, 2019.

[83] C. Wu, J. Chen, Q. R. Fang, *et al.*, "Rethinking membership inference attacks against transfer learning," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 6441–6454, 2024.

[84] H. Abdullah, M. S. Rahman, W. Garcia, *et al.*, "Hear "no evil", see "kenansville": Efficient and transferable black-box attacks on speech recognition and voice identification systems," in *Proceedings of the 2021 IEEE Symposium on Security and Privacy*, San Francisco, CA, USA, pp. 712–729, 2021 .

[85] Y. H. Hua, Y. Xu, C. Lyu, *et al.*, "Imperceptible and targeted physical attacks on deep learning-based speech semantic communications," in *Proceedings of the 2025 IEEE Wireless Communications and Networking Conference (WCNC)*, Milan, Italy, pp. 1–6, 2025.

[86] L. Schönherr, K. Kohls, S. Zeiler, *et al.*, "Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding," in *Proceedings of the 26th Annual Network and Distributed System Security Symposium*, San Diego, CA, USA, pp. 1–15, 2019.

[87] A. Fawzi, H. Fawzi, and O. Fawzi, "Adversarial vulnerability for any classifier," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Montréal, Canada, pp. 1186–1195, 2018.

[88] H. B. Cai, P. C. Zhang, Y. Xiao, *et al.*, "Clean-label backdoor attack based on robust feature attenuation for speech recognition," *Expert Systems with Applications*, vol. 281, article no. 127546, 2025.

[89] W. H. Yao, J. K. Yang, Y. Q. He, *et al.*, "Imperceptible rhythm backdoor attacks: Exploring rhythm transformation for embedding undetectable vulnerabilities on speech recognition," *Neurocomputing*, vol. 614, article no. 128779, 2025.

[90] X. J. Yuan, J. S. Zhang, F. Guo, *et al.*, "EvilHarmony: Stealthy adversarial attacks against black-box speech recognition systems," in *Proceedings of the 2025 IEEE Symposium on Security and Privacy (SP)*, San Francisco, CA, USA, pp. 4569–4587, 2025.

[91] Y. X. Chen, X. J. Yuan, J. S. Zhang, *et al.*, "Devil's whisper: A general approach for physical adversarial attacks against commercial black-box speech recognition devices," in *Proceedings of the 29th USENIX Conference on Security Symposium*, article no. 150, 2020.

[92] W. F. Jin, Y. X. Cao, J. J. Su, *et al.*, "Towards evaluating the robustness of automatic speech recognition systems via audio style transfer," in *Proceedings of the 2nd ACM Workshop on Secure and Trustworthy Deep Learning Systems*, Singapore, Singapore, pp. 47–55, 2024.

[93] J. D. Yao, H. Luo, J. Qi, *et al.*, "Interpretable spectrum transformation attacks to speaker recognition systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1531–1545, 2024.

[94] T. Y. Du, S. L. Ji, J. F. Li, *et al.*, "SirenAttack: Generating adversarial audio for end-to-end acoustic systems," in *Proceedings of the 15th ACM Asia Conference on Computer and Communications Security*, Taipei, China, pp. 357–369, 2020.

[95] A. Madry, A. Makelov, L. Schmidt, *et al.*, "Towards deep learning models resistant to adversarial attacks," in *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, Canada, pp. 1–12, 2018.

[96] X. Du, Q. Zhang, J. J. Zhu, *et al.*, "Adaptive unified defense framework for tackling adversarial audio attacks," *Artificial Intelligence Review*, vol. 57, no. 8, article no. 218, 2024.

[97] N. Park and J. Kim, "Toward robust ASR system against audio adversarial examples using agitated logit," *ACM Transactions on Privacy and Security*, vol. 27, no. 2, article no. 20, 2024.

[98] A. B. Nassif, I. Shahin, I. Attili, *et al.*, "Speech recognition using deep neural networks: A systematic review," *IEEE Access*, vol. 7, pp. 19143–19165, 2019.

[99] S. K. Ergünay, E. Khoury, A. Lazaridis, *et al.*, "On the vulnerability of speaker verification to realistic voice spoofing," in *Proceedings of the 2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Arlington, VA, USA, pp. 1–6, 2015.

[100] N. W. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," in *Proceedings of the Interspeech*, Lyon, France, pp. 925–929, 2013.

[101] J. Lindberg and M. Blomberg, "Vulnerability in speaker verification-a study of technical impostor techniques," in *Proceedings of the 6th European Conference on Speech Communication and Technology*, Budapest, Hungary, pp. 1211–1214, 1999.

[102] J. Villalba and E. Lleida, "Speaker verification performance degradation against spoofing and tampering attacks," in *Proceedings of the FALA Workshop,* Vigo, Spain, pp. 131–134, 2010.

[103] J. Villalba and E. Lleida, "Detecting replay attacks from far-field recordings on speaker verification systems," in *Proceedings of the COST 2101 European Workshop on Biometrics and ID Management*, Brandenburg, Germany, pp. 274–285, 2011.

[104] S. H. Yoon, M. S. Koh, J. H. Park, *et al.*, "A new replay attack against automatic speaker verification systems," *IEEE Access*, vol. 8, pp. 36080–36088, 2020.

[105] Y. W. Lau, D. Tran, and M. Wagner, "Testing voice mimicry with the YOHO speaker verification corpus," in *Proceedings of the 9th International Conference on Knowledge-Based Intelligent Information and Engineering Systems*, Melbourne, Australia, pp. 15–21, 2005.

[106] Y. W. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *Proceedings of the 2004 International Symposium on Intelligent Multimedia*, Hong Kong, China, pp. 145–148, 2004.

[107] M. Farrús, M. Wagner, J. Anguita, *et al.*, "How vulnerable are prosodic features to professional imitators?," in *Proceedings of the Speaker and Language Recognition Workshop*, Stellenbosch, South Africa, article no. 2, 2008.

[108] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, *et al.*, "I-vectors meet imitators: On vulnerability of speaker verification systems against voice mimicry," in *Proceedings of the 14th Annual Conference of the International Speech Communication Association*, Lyon, France, pp. 930–934, 2013.

[109] J. Mariéthoz and S. Bengio, "Can a professional imitator fool a GMM-based speaker verification system?," Idiap Research Institute, *Idiap Research Institute* RR-61-2005, 2005.

[110] K. Wang, M. Chen, L. Lu, *et al.*, "From one stolen utterance: Assessing the risks of voice cloning in the AIGC Era," in *Proceedings of the 2025 IEEE Symposium on Security and Privacy*, San Francisco, CA, USA, pp. 4663–4681, 2025.

[111] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proceedings of the 1996 International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, Atlanta, GA, USA, pp. 373–376, 1996.

[112] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[113] Y. Qian, F. K. Soong, and Z. J. Yan, "A unified trajectory tiling approach to high quality speech rendering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 280–290, 2013.

[114] M. Bińkowski, J. Donahue, S. Dieleman, *et al.*, "High fidelity speech synthesis with adversarial networks," in *Proceedings of the 8th International Conference on Learning Representations*, pp. 1–15, 2019.

[115] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in *Proceedings of the 18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, pp. 3642–3646, 2017.

[116] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 84–96, 2018.

[117] A. Oord, Y. Z. Li, I. Babuschkin, *et al.*, "Parallel WaveNet: Fast high-fidelity speech synthesis," in *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, Sweden, pp. 3915–3923, 2018.

[118] A. Van Den Oord, S. Dieleman, H. G. Zen, *et al.*, "WaveNet: A generative model for raw audio," in *Proceedings of the 9th ISCA Speech Synthesis Workshop*, Sunnyvale, CA, USA, article no. 125, 2016.

[119] Y. X. Wang, R. Skerry-Ryan, D. Stanton, *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *Proceedings of the 18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, pp. 4006–4010, 2017.

[120] E. K. Kim, S. Lee, and Y. H. Oh, "Hidden markov model based voice conversion using dynamic characteristics of speaker," in *Proceedings of the 5th European Conference on Speech Communication and Technology*, Rhodes, Greece, pp. 2519–2522, 1997.

[121] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[122] M. M. Wilde and A. B. Martinez, "Probabilistic principal component analysis applied to voice conversion," in *Proceedings of the Conference Record of the 38th Asilomar Conference on Signals, Systems and Computers, 2004*, Pacific Grove, CA, USA, pp. 2255–2259, 2004.

[123] S. F. Zhang, D. Y. Huang, L. Xie, *et al.*, "Non-negative matrix factorization using stable alternating direction method of multipliers for source separation," in *Proceedings of the 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Hong Kong, China, pp. 222–228, 2015.

[124] S. Desai, E. V. Raghavendra, B. Yegnanarayana, *et al.*, "Voice conversion using artificial neural networks," in *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, China, pp. 3893–3896, 2009.

[125] M. Ghorbandoost, A. Sayadiyan, M. Ahangar, *et al.*, "Voice conversion based on feature combination with limited training data," *Speech Communication*, vol. 67, pp. 113–128, 2015.

[126] M. J. Jannati and A. Sayadiyan, "Part-syllable transformation-based voice conversion with very limited training data," *Circuits, Systems, and Signal Processing*, vol. 37, no. 5, pp. 1935–1957, 2018.

[127] Y. Zhou, X. H. Tian, H. H. Xu, *et al.*, "Cross-lingual voice

conversion with bilingual phonetic posteriorgram and average modeling," in *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, Brighton, UK, pp. 6790–6794, 2019.

[128] S. Ahmed, Y. Wani, A. S. Shamsabadi, *et al.*, "Tubes among us: Analog attack on automatic speaker identification," in *Proceedings of the 32nd USENIX Conference on Security Symposium*, Anaheim, CA, USA, article no. 16, 2023.

[129] H. A. Patil and K. K. Parhi, "Variable length teager energy based mel cepstral features for identification of twins," in *Proceedings of the 3rd International Conference on Pattern Recognition and Machine Intelligence*, New Delhi, India, pp. 525–530, 2009.

[130] L. G. Kersta and J. A. Colangelo, "Spectrographic speech patterns of identical twins," *The Journal of the Acoustical Society of America*, vol. 47, no. S1A, pp. 58–59, 1970.

[131] BBC, "BBC reporter fools bank voice-ID security," Available at: https://www.bbc.co.uk/news/technology-39973217, 2017-03-19.

[132] J. Villalba and E. Lleida, "Preventing replay attacks on speaker verification systems," in *Proceedings of the 2011 Carnahan Conference on Security Technology*, Barcelona, Spain, pp. 1–8, 2011.

[133] W. Shang and M. Stevenson, "Score normalization in playback attack detection," in *Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, TX, USA, pp. 1678–1681, 2010.

[134] T. Satoh, T. Masuko, T. Kobayashi, *et al.*, "A robust speaker verification system against imposture using an hmm-based speech synthesis system," in *Proceedings of the 7th European Conference on Speech Communication and Technology*, Aalborg, Denmark, pp. 759–762, 2001.

[135] L. W. Chen, W. Guo, and L. R. Dai, "Speaker verification against synthetic speech," in *Proceedings of the 2010 7th International Symposium on Chinese Spoken Language Processing*, Tainan, China, pp. 309–312, 2010.

[136] A. Ogihara, H. Unno, and A. Shiozaki, "Discrimination method of synthetic speech using pitch frequency against synthetic speech falsification," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E88-A, no. 1, pp. 280–286, 2005.

[137] Z. Z. Wu, T. Kinnunen, E. S. Chng, *et al.*, "A study on spoofing attack in state-of-the-art speaker verification: The telephone speech case," in *Proceedings of the 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, Hollywood, CA, USA, pp. 1–5, 2012.

[138] C. Liu, J. Zhang, T. W. Zhang, *et al.*, "Detecting voice cloning attacks via timbre watermarking," in *Proceedings of the Network and Distributed System Security Symposium*, San Diego, California, pp. 1–18, 2024.

[139] Z. H. Liu, Y. Zhang, and C. L. Miao, "Protecting your voice from speech synthesis attacks," in *Proceedings of the 39th Annual Computer Security Applications Conference (ACSAC '23)*, Austin, TX, USA, pp. 394–408, 2023.

[140] K. Y. Zhang, Z. Y. Hua, Y. S. Zhang, *et al.*, "Robust AI-synthesized speech detection using feature decomposition learning and synthesizer feature augmentation," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 871–885, 2025.

[141] F. Alegre, R. Vipperla, and N. Evans, "Spoofing countermeasures for the protection of automatic speaker recognition from attacks with artificial signals," in *Proceedings of the 13th Annual Conference of the International Speech Communication Association*, Portland, OR, USA, pp. 1688–1691, 2012.

[142] F. Alegre, A. Amehraye, and N. Evans, "Spoofing countermeasures to protect automatic speaker verification from voice conversion," in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada, pp. 3068–3072, 2013.

[143] Y. Gong, J. Yang, and C. Poellabauer, "Detecting replay attacks using multi-channel audio: A neural network-based method," *IEEE Signal Processing Letters*, vol. 27, pp. 920–924, 2020.

[144] S. Jelil, S. Kalita, S. M. Prasanna, *et al.*, "Exploration of compressed ILPR features for replay attack detection," in *Proceedings of the Interspeech*, Hyderabad, India, pp. 631–635, 2018.

[145] M. R. Kamble, H. Tak, and H. A. Patil, "Effectiveness of speech demodulation-based features for replay detection," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association*, Hyderabad, India, pp. 641–645, 2018.

[146] A. T. Patil, R. Acharya, P. K. A. Sai, *et al.*, "Energy separation-based instantaneous frequency estimation for cochlear cepstral feature for replay spoof detection," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association*, Graz, Austriapp, pp. 2898–2902, 2019.

[147] X. L. Wang, Y. H. Xiao, and X. Zhu, "Feature selection based on CQCCS for automatic speaker verification spoofing," in *Proceedings of the 18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, pp. 32–36, 2017.

[148] M. Witkowski, S. Kacprzak, P. Zelasko, *et al.*, "Audio replay attack detection using high-frequency features," in *Proceedings of the 18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, pp. 27–31, 2017.

[149] H. Tak, J. Patino, M. Todisco, *et al.*, "End-to-end anti-spoofing with RawNet2," in *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, ON, Canada, pp. 6369–6373, 2021.

[150] Z. Z. Wu, T. Kinnunen, N. W. D. Evans, *et al.*, "ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge," in *Proceedings of the 16th Annual Conference of the International Speech Communication Association*, Dresden, Germany, pp. 2037–2041, 2015.

[151] T. Kinnunen, M. Sahidullah, H. Delgado, *et al.*, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proceedings of the 18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, pp. 2–6, 2017.

[152] M. Todisco, X. Wang, V. Vestman, *et al.*, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, pp. 1008–1012, 2019.

[153] J. Yamagishi, X. Wang, M. Todisco, *et al.*, "ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection," in *Proceedings of the 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, Online, pp. 1–8, 2021.

[154] L. Blue, K. Warren, H. Abdullah, *et al.*, "Who are you (i really wanna know)? Detecting audio DeepFakes through vocal tract reconstruction," in *Proceedings of the 31st USENIX Security Symposium*, Boston, MA, USA, pp. 2691–2708, 2022.

[155] Y. Y. Gu, J. Chen, C. R. Chen, *et al.*, "CSIPose: Unveiling human poses using commodity WiFi devices through the wall," *IEEE Transactions on Mobile Computing*, vol. 24, no. 10, pp. 10914–10926, 2025.

[156] J. Y. Deng, Y. J. Chen, Y. N. Zhong, *et al.*, "Catch you and i can: Revealing source voiceprint against voice conversion," in *Proceedings of the 32nd USENIX Confer-*

ence on Security Symposium, Anaheim, CA, USA, article no. 289, 2023.

[157] Y. Qin, N. Carlini, G. W. Cottrell, *et al.*, "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," in *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, CA, USA, pp. 5231–5240, 2019.

[158] R. Taori, A. Kamsetty, B. Chu, *et al.*, "Targeted adversarial examples for black box audio systems," in *Proceedings of the 2019 IEEE Security and Privacy Workshops*, San Francisco, CA, USA, pp. 15–20, 2019.

[159] P. Neekhara, S. Hussain, P. Pandey, *et al.*, "Universal adversarial perturbations for speech recognition systems," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, pp. 481–485, 2019.

[160] H. Kwon, Y. Kim, H. Yoon, *et al.*, "Selective audio adversarial example in evasion attack on speech recognition system," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 526–538, 2020.

[161] Y. Gong, B. Y. Li, C. Poellabauer, *et al.*, "Real-time adversarial attacks," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, Macao, China, pp. 4672–4680, 2019.

[162] X. L. Liu, K. Wan, Y. F. Ding, *et al.*, "Weighted-sampling audio adversarial example attack," in *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, Philadelphia, Pennsylvania, pp. 4908–4915, 2020.

[163] L. Schönherr, T. Eisenhofer, S. Zeiler, *et al.*, "Imperio: Robust over-the-air adversarial examples for automatic speech recognition systems," in *Proceedings of the 36th Annual Computer Security Applications Conference*, Austin, TX, USA, pp. 843–855, 2020.

[164] J. Szurley and J. Z. Kolter, "Perceptual based adversarial audio attacks," *arXiv preprint*, arXiv: 1906.06355.

[165] J. K. Han, H. Kim, and S. S. Woo, "Nickel to lego: Using foolgle to create adversarial examples to fool Google cloud speech-to-text API," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, London, UK, pp. 2593–2595, 2019.

[166] S. Khare, R. Aralikatte, and S. Mani, "Adversarial black-box attacks on automatic speech recognition systems using multi-objective evolutionary optimization," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, pp. 3208–3212, 2019.

[167] Y. Wu, J. Liu, Y. Y. Chen, *et al.*, "Semi-black-box attacks against speech recognition systems using adversarial samples," in *Proceedings of the 2019 IEEE International Symposium on Dynamic Spectrum Access Networks*, Newark, NJ, USA, pp. 1–5, 2019.

[168] T. Gong, A. G. C. P. Ramos, S. Bhattacharya, *et al.*, "AudiDoS: Real-time denial-of-service adversarial attacks on deep audio models," in *Proceedings of the 2019 18th IEEE International Conference on Machine Learning and Applications*, Boca Raton, FL, USA, pp. 978–985, 2019.

[169] T. Chen, L. Shangguan, Z. J. Li, *et al.*, "Metamorph: Injecting inaudible commands into over-the-air voice controlled systems," in *Proceedings of the 27th Annual Network and Distributed System Security Symposium*, San Diego, California, USA, pp. 1–17, 2020.

[170] D. H. Wang, L. Dong, R. D. Wang, *et al.*, "Targeted speech adversarial example generation with generative adversarial network," *IEEE Access*, vol. 8, pp. 124503–124513, 2020.

[171] Q. Wang, B. L. Zheng, Q. Li, *et al.*, "Towards query-efficient adversarial attacks against automatic speech recognition systems," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 896–908, 2021.

[172] Z. H. Li, Y. Wu, J. Liu, *et al.*, "AdvPulse: Universal, synchronization-free, and targeted audio adversarial attacks via subsecond perturbations," in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, Virtual Event, pp. 1121–1134, 2020.

[173] G. K. Chen, S. Chenb, L. L. Fan, *et al.*, "Who is real bob? Adversarial attacks on speaker recognition systems," in *Proceedings of the 2021 IEEE Symposium on Security and Privacy*, San Francisco, CA, USA, pp. 694–711, 2021.

[174] X. Wang, S. N. Sun, C. H. Shan, *et al.*, "Adversarial examples for improving end-to-end attention-based small-footprint keyword spotting," in *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, Brighton, UK, pp. 6366–6370, 2019.

[175] Y. Xie, C. Shi, Z. H. Li, *et al.*, "Real-time, universal, and robust adversarial attacks against speaker recognition systems," in *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, Barcelona, Spain, pp. 1738–1742, 2020.

[176] Z. H. Li, C. Shi, Y. Xie, *et al.*, "Practical adversarial attacks against speaker recognition systems," in *Proceedings of the 21st International Workshop on Mobile Computing Systems and Applications*, Austin, TX, USA, pp. 9–14, 2020.

[177] J. G. Li, X. F. Zhang, C. M. Jia, *et al.*, "Universal adversarial perturbations generative network for speaker recognition," in *Proceedings of the 2020 IEEE International Conference on Multimedia and Expo*, London, UK, pp. 1–6, 2020.

[178] L. Zhang, Y. Meng, J. H. Yu, *et al.*, "Voiceprint mimicry attack towards speaker verification system in smart home," in *Proceedings of the IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, Toronto, ON, Canada, pp. 377–386, 2020.

[179] J. G. Li, X. F. Zhang, J. Z. Xu, *et al.*, "Learning to fool the speaker recognition," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 17, no. 3s, article no. 109, 2021.

[180] A. S. Shamsabadi, F. S. Teixeira, A. Abad, *et al.*, "FoolHD: Fooling speaker identification by highly imperceptible adversarial disturbances," in *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, ON, Canada, pp. 6159–6163, 2021.

[181] Z. Y. Yu, Y. Chang, N. Zhang, *et al.*, "Smack: Semantically meaningful adversarial audio attack," in *Proceedings of the 32nd USENIX Security Symposium*, Anaheim, CA, USA, pp. 3799–3816, 2023.

[182] B. L. Zheng, P. P. Jiang, Q. Wang, *et al.*, "Black-box adversarial attacks on commercial speech platforms with minimal information," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, Virtual Event, pp. 86–107, 2021.

[183] H. Liu, Z. Y. Yu, M. M. Zha, *et al.*, "When evil calls: Targeted adversarial voice over IP network," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, Los Angeles, CA, USA, pp. 2009–2023, 2022.

[184] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of the ICNN'95-International Conference on Neural Networks*, Perth, WA, Australia, pp. 1942–1948, 1995.

[185] D. Wierstra, T. Schaul, T. Glasmachers, *et al.*, "Natural evolution strategies," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 949–980, 2014.

[186] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Proceedings of the 5th International Conference on Learning Representations*, Toulon, France, pp. 1–14, 2017.

[187] Z. Fang, T. Wang, L. C. Zhao, *et al.*, "Zero-query adversarial attack on black-box automatic speech recognition

systems," in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security (CCS '24)*, Salt Lake City, UT, USA, pp. 630–644, 2024.

[188] P. Cheng, Y. W. Wang, P. Huang, *et al.*, "ALIF: Low-cost adversarial audio attacks on black-box speech platforms using linguistic features," in *Proceedings of the 2024 IEEE Symposium on Security and Privacy*, San Francisco, CA, USA, pp. 1628–1645, 2024.

[189] C. X. Zuo, J. Y. Leng, and W. J. Li, "SUETA: Speaker-specific utterance ensemble based transfer attack on speaker identification system," *Computers & Security*, vol. 144, article no. 103948, 2024.

[190] Z. X. Guo, Z. J. Lin, P. Li, *et al.*, "SkillExplorer: Understanding the behavior of skills in large scale," in *Proceedings of the 29th USENIX Conference on Security Symposium*, article no. 149, 2020.

[191] D. Su, J. Q. Liu, S. C. Zhu, *et al.*, "'Are you home alone?' 'Yes' Disclosing security and privacy vulnerabilities in alexa skills," *arXiv preprint*, arXiv: 2010.10788, 2020.

[192] L. Cheng, C. Wilson, S. Liao, *et al.*, "Dangerous skills got certified: Measuring the trustworthiness of skill certification in voice personal assistant platforms," in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, Virtual Event, pp. 1699–1716, 2020.

[193] A. Stolk, L. Verhagen, and I. Toni, "Conceptual alignment: How brains achieve mutual understanding," *Trends in Cognitive Sciences*, vol. 20, no. 3, pp. 180–191, 2016.

[194] Y. Zhang, L. Xu, A. Mendoza, *et al.*, "Life after speech recognition: Fuzzing semantic misinterpretation for voice assistant applications," in *Proceedings of the 26th Annual Network and Distributed System Security Symposium*, San Diego, CA, USA, pp. 1–15, 2019.

[195] J. Young, S. Liao, L. Cheng, *et al.*, "SkillDetective: Automated Policy-Violation detection of voice assistant applications in the wild," in *Proceedings of the 31st USENIX Security Symposium*, Boston, MA, USA, pp. 1113–1130, 2022.

[196] P. De Vaere and A. Perrig, "Hey kimya, is my smart speaker spying on me? Taking control of sensor privacy through isolation and amnesia," in *Proceedings of the 32nd USENIX Security Symposium*, Anaheim, CA, USA, pp. 2401–2418, 2023.

[197] M. E. Ahmed, I. Y. Kwak, J. H. Huh, *et al.*, "Void: A fast and light voice liveness detection system," in *Proceedings of the 29th USENIX Security Symposium*, pp. 2685–2702, 2020.

[198] L. Blue, L. Vargas, and P. Traynor, "Hello, is it me you're looking for?: Differentiating between human and electronic speakers for voice interface security," in *Proceedings of the 11th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, Stockholm, Sweden, pp. 123–133, 2018.

[199] Y. Gong and C. Poellabauer, "Protecting voice controlled systems using sound source identification based on acoustic cues," in *Proceedings of the 2018 27th International Conference on Computer Communication and Networks*, Hangzhou, China, pp. 1–9, 2018.

[200] S. Pradhan, W. Sun, G. Baig, *et al.*, "Combating replay attacks against voice assistants," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, article no. 100, 2019.

[201] S. Wang, J. H. Cao, X. He, *et al.*, "When the differences in frequency domain are compensated: Understanding and defeating modulated replay attacks on automatic speech recognition," in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, Virtual Event, pp. 1103–1119, 2020.

[202] S. Chen, K. Ren, S. X. Piao, *et al.*, "You can hear but you cannot steal: Defending against voice impersonation attacks on smartphones," in *Proceedings of the 2017 IEEE 37th International Conference on Distributed Computing Systems*, Atlanta, GA, USA, pp. 183–195, 2017.

[203] S. Mochizuki, S. Shiota, and H. Kiya, "Voice liveness detection using phoneme-based pop-noise detector for speaker verification," in *Proceedings of the Odyssey 2018: The Speaker and Language Recognition Workshop*, Les Sables D'Olonne, France, pp. 233–239, 2018.

[204] S. Shiota, F. Villavicencio, J. Yamagishi, *et al.*, "Voice liveness detection for speaker verification based on a tandem single/double-channel pop noise detector," in *Proceedings of the Odyssey 2016: The Speaker and Language Recognition Workshop*, Bilbao, Spain, pp. 259–263, 2016.

[205] Y. Wang, W. D. Cai, T. Gu, *et al.*, "Secure your voice: An oral airflow-based continuous liveness detection for voice assistants," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 40, no. 3, article no. 157, 2020.

[206] M. Zhou, Z. Qin, X. Lin, *et al.*, "Hidden voice commands: Attacks and defenses on the VCS of autonomous driving cars," *IEEE Wireless Communications*, vol. 26, no. 5, pp. 128–133, 2019.

[207] Y. Meng, Z. C. Wang, W. Zhang, *et al.*, "Wivo: Enhancing the security of voice control system via wireless signal in IoT environment," in *Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, Los Angeles, CA, USA, pp. 81–90, 2018.

[208] J. C. Shang and J. Wu, "Voice liveness detection for voice assistants using ear canal pressure," in *Proceedings of the 2020 IEEE 17th International Conference on Mobile Ad Hoc and Sensor Systems*, Delhi, India, pp. 693–701, 2020.

[209] L. H. Zhang, S. Tan, and J. Yang, "Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, Dallas, Texas, USA, pp. 57–71, 2017.

[210] H. Feng, K. Fawaz, and K. G. Shin, "Continuous authentication for voice assistants," in *Proceedings of the 23rd International Conference on Mobile Computing and Networking*, Snowbird, Utah, USA, pp. 343–355, 2017.

[211] M. Sahidullah, D. A. L. Thomsen, R. G. Hautamäki, *et al.*, "Robust voice liveness detection and speaker verification using throat microphones," *Proceedings of the IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 44–56, 2018.

[212] J. C. Shang, S. Chen, and J. Wu, "Defending against voice spoofing: A robust software-based liveness detection system," in *Proceedings of the 2018 IEEE 15th International Conference on Mobile Ad Hoc and Sensor Systems*, Chengdu, China, pp. 28–36, 2018.

[213] J. C. Shang and J. Wu, "Enabling secure voice input on augmented reality headsets using internal body voice," in *Proceedings of the 2019 16th Annual IEEE International Conference on Sensing, Communication, and Networking*, Boston, MA, USA, pp. 1–9, 2019.

[214] J. C. Shang and J. Wu, "Secure voice input on augmented reality headsets," *IEEE Transactions on Mobile Computing*, vol. 21, no. 4, pp. 1420–1433, 2022.

[215] L. H. Zhang, S. Tan, Z. Wang, *et al.*, "VibLive: A continuous liveness detection for secure voice user interface in IoT environment," in *Proceedings of the 36th Annual Computer Security Applications Conference*, Austin, TX, USA, pp. 884–896, 2020.

[216] Q. Yang, K. Y. Cui, and Y. Q. Zheng, "VoShield: Voice liveness detection with sound field dynamics," in *Proceedings of the IEEE INFOCOM 2023 - IEEE Conference on Computer Communications*, New York, NY, USA, pp. 1–10, 2023.

[217] X. J. Yuan, Y. X. Chen, A. H. Wang, *et al.*, "All your

Alexa are belong to us: A remote voice control attack against echo," in *Proceedings of the 2018 IEEE Global Communications Conference*, Abu Dhabi, United Arab Emirates, pp. 1–6, 2018.

[218] K. Rajaratnam, K. Shah, and J. Kalita, "Isolated and ensemble audio preprocessing methods for detecting adversarial examples against automatic speech recognition," in *Proceedings of the 30th Conference on Computational Linguistics and Speech Processing*, Hsinchu, China, pp. 16–30, 2018.

[219] I. Andronic, L. Kürzinger, E. R. Chavez Rosas, *et al.*, "MP3 compression to diminish adversarial noise in end-to-end speech recognition," in *Proceedings of the 22nd International Conference on Speech and Computer*, St. Petersburg, Russia, pp. 22–34, 2020.

[220] N. Das, M. Shanbhogue, S. T. Chen, *et al.*, "ADAGIO: Interactive experimentation with adversarial attack and defense for audio," in *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, Dublin, Ireland, pp. 677–681, 2018.

[221] K. Rajaratnam, B. Alshemali, and J. Kalita, "Speech coding and audio preprocessing for mitigating and detecting audio adversarial examples on automatic speech recognition," in *Proceedings of the Seminar Machine Learning in Computer Vision and Natural Language Processing*, Colorado Springs, CO, USA, pp. 1–7, 2018.

[222] J. J. Zhang, B. S. Zhang, and B. C. Zhang, "Defending adversarial attacks on cloud-aided automatic speech recognition systems," in *Proceedings of the 7th International Workshop on Security in Cloud Computing*, Auckland, New Zealand, pp. 23–31, 2019.

[223] Z. L. Yang, B. Li, P. Y. Chen, *et al.*, "Characterizing audio adversarial examples using temporal dependency," in *Proceedings of the 7th International Conference on Learning Representations*, New Orleans, LA, USA, pp. 1–15, 2019.

[224] K. Tamura, A. Omagari, and S. Hashida, "Novel defense method against audio adversarial example for speech-to-text transcription neural networks," in *Proceedings of the 2019 IEEE 11th International Workshop on Computational Intelligence and Applications*, Hiroshima, Japan, pp. 115–120, 2019.

[225] Q. L. Guo, J. Ye, Y. R. Chen, *et al.*, "INOR—An intelligent noise reduction method to defend against adversarial audio examples," *Neurocomputing*, vol. 401, pp. 160–172, 2020.

[226] C. Wu, K. He, J. Chen, *et al.*, "Liveness is not enough: Enhancing fingerprint authentication with behavioral biometrics to defeat puppet attacks," in *Proceedings of the 29th USENIX Conference on Security Symposium*, Boston, MA, USA, pp. 125, 2020.

[227] C. Wu, K. He, J. Chen, *et al.*, "Toward robust detection of puppet attacks via characterizing fingertip-touch behaviors," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 6, pp. 4002–4018, 2022.

[228] C. Wu, J. Chen, K. He, *et al.*, "EchoHand: High accuracy and presentation attack resistant hand authentication on commodity mobile devices," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, Los Angeles, CA, USA, pp. 2931–2945, 2022.

[229] C. Wu, K. He, J. Chen, *et al.*, "High accuracy and presentation attack resistant hand authentication via acoustic sensing for commodity mobile devices," *IEEE Transactions on Dependable and Secure Computing*, vol. 22, no. 5, pp. 5231–5247, 2025.

[230] S. Chang, L. Zhou, W. Liu, *et al.*, "Combating voice spoofing attacks on wearables via speech movement sequences," *IEEE Transactions on Dependable and Secure Computing*, vol. 22, no. 1, pp. 819–832, 2025.

[231] C. Wu, J. Chen, S. M. Zhu, *et al.*, "WAFBooster: Automatic boosting of WAF security against mutated malicious payloads," *IEEE Transactions on Dependable and Secure Computing*, vol. 22, no. 2, pp. 1118–1131, 2025.

[232] Y. Meng, J. C. Li, M. Pillari, *et al.*, "Your microphone array retains your identity: A robust voice liveness detection system for smart speaker," in *Proceedings of the 31st USENIX Security Symposium*, Boston, MA, USA, pp. 1077–1094, 2022.

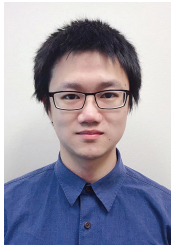
[233] T. Sasi, A. H. Lashkari, R. X. Lu, *et al.*, "A comprehensive survey on IoT attacks: Taxonomy, detection mechanisms and challenges," *Journal of Information and Intelligence*, vol. 2, no. 6, pp. 455–513, 2024.

Yangyang Gu received the Ph.D. degree in cyber science from Wuhan University, Wuhan, China. He is a Postdoctoral Fellow with the School of Cyber Science and Engineering, Wuhan University, Wuhan, China. His research outcomes have appeared in UbiComp, IEEE TMC, and FCS. His research interests include drone security and wireless sensing.
(Email: guyangyang@whu.edu.cn)

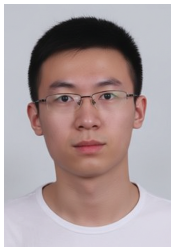
Haozhe Xu received the M.S. degree in cyberspace security at the School of Cyber Science and Engineering, Wuhan University, Wuhan, China. His research interests include identity authentication and cybersecurity.
(Email: haozhexu@whu.edu.cn)

Jing Chen received the Ph.D. degree in computer science from the Huazhong University of Science and Technology, Wuhan, China. He is currently a Full Professor with the School of Cyber Science and Engineering, Wuhan University, Wuhan, China. He is a Senior Member of IEEE and served as the vice chair of the ACM Turing Award Celebration Conference (TURC) 2023. He has published more than 150 research papers in many international journals and conferences, including USENIX Security, ACM CCS, INFOCOM, IEEE TDSC, IEEE TIFS, IEEE TMC, IEEE TC, IEEE TPDS, IEEE TSC, etc. He was twice runner-up for the best paper at INFOCOM 2018 and INFOCOM 2021. His research interests include the areas of network security, cloud security, and mobile security.
(Email: chenjing@whu.edu.cn)

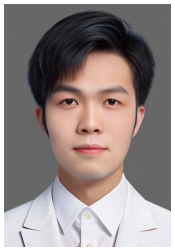
Jiahua Xu is an Associate Professor in financial computing, and Programme Director of the M.S. emerging digital technologies at University College London (UCL), London, UK. She is also affiliated to the UCL Centre for Blockchain Technologies. Her research interests includs blockchain economics and decentralized finance. She has published in Usenix Security, ACM IMC, FC, IEEE ICDCS and IEEE ICBC. She has reviewed for Advances in Complex Systems, Computer Networks, Transactions on the Web and Cities.
(Email: jiahua.xu@ucl.ac.uk)



Yebo Feng is a Research Fellow with the School of Computer Science and Engineering (SCSE) at Nanyang Technological University (NTU), Singapre. He received the Ph.D. degree in computer science from the University of Oregon (UO), Eugene, United States, in 2023. His research interests include network security, blockchain security, and anomaly detection. He is the recipient of the Best Paper Award of 2019 IEEE CNS, Gurdeep Pall Graduate Student Fellowship of UO, and Ripple Research Fellowship. He has served as the reviewer of IEEE TDSC, IEEE TIFS, ACM TKDD, IEEE JSAC, IEEE COMST, etc. Furthermore, he has been a member of the program committees for international conferences including SDM, CIKM, and CYBER, and has also served on the Artifact Evaluation (AE) committees for USENIX OSDI and USENIX ATC.
(Email: yebo.feng@ntu.edu.sg)



Ju Jia received the Ph.D. degree in cyberspace security from Wuhan University, Wuhan, China, in 2021, and he was a Research Fellow with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He is currently an Associate Professor with the School of Cyber Science and Engineering, Southeast University, Nanjing, China. His research interests include data security, multimedia data security, and artificial intelligence security.
(Email: jiaju@seu.edu.cn)



Cong Wu is currently a Postdoctoral Researcher with The University of Hong Kong (HKU), Hong Kong, China. His research interests include security and privacy of intelligent systems. Before that, he served as a Research Fellow at Nanyang Technological University, Singapre, working with Prof. Yang Liu. He obtained the Ph.D. degree from Wuhan University, Wuhan, China and the B.E. degree from Xidian University, Xi'an, China. He served as Track Chair of ICPADS, Symposia/Workshop Chair for IEEE CCPQT, IWCMC, and SMC-IoT, Advisory Committee Member for ICoACT, TPC for CAAW, AAMAS, CITS, CCCI, KDD, AAAI, AEC Member for USENIX Security'25, CCS'25. He currently serves as the Associate Editor for JISA, IJCS, and SPY. He received many academic awards, including HKU Postdoctoral Fellowship, Distinguished Paper Award at China Cybersecurity 2023, Best Paper Award at ML4CS'24, TrustCom'24, WASA'25, Exponential Science Pioneers Award, and Distinguished Doctoral Dissertation Award. He has published 50+ papers in top conferences and journals, e.g., USENIX Security, CCS, ICSE, ASE, FSE, SIGIR, KDD, TIFS, TDSC, TOSEM, JSAC, etc.
(Email: cnacwu@whu.edu.cn)

Teng Li received the B.S. degree from the School of Computer Science and Technology, Xidian University, Xi'an, China, in 2013, and the Ph.D. degree from the School of Computer Science and Technology, Xidian University, in 2018. He is currently an Associate Professor with the School of Cyber Engineering, Xidian University. His current research interests include wireless and mobile networks, distributed systems and intelligent terminals with focus on security and privacy issues.
(Email: litengxidian@gmail.com)

Jian Shen received the M.E. and the Ph.D. degrees in computer science from Chosun University, Gwangju, South Korea, in 2009 and 2012, respectively. Since 2012, he has been a Professor with the School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, China. He is currently a Professor with the School of Information Science and Engineering, Zhejiang Sci-Tech University, Hangzhou, China. His research interests include public cryptography, cloud computing and security, data auditing and sharing, and information security systems.
(Email: s_shenjian@126.com)

Jianfeng Ma received the B.S. degree in computer science from Shaanxi Normal University, Xi'an, China, in 1982, the M.S. degree in computer science from Xidian University, Xi'an, China, in 1992, and the Ph.D. degree in computer science from Xidian University, Xi'an, China, in 1995. Currently, he is the Directer of Department of Cyber Engineering and a Professor with the School of Cyber Engineering, Xidian University. He has published more than 150 journal and conference papers. His research interests include information security, cryptography, and network security.
(Email: jfma@mail.xidian.edu.cn)