Research on Laser Counterfeit Command Attack Mechanisms for In-vehicle Voice Interfaces in Intelligent Connected Vehicles

Quanrui Huo [1], Yuqiao Ning [1], Zhen Guo [1]

[1] CATARC Intelligent and connected technology Co.,LTd., China.

Abstract. With the evolution of human-machine interaction in intelligent connected vehicle (ICV), voice control has become a core feature. However, the associated security risks are increasingly evident, particularly remote, covert audio injection attacks such as laser voice attacks, which can bypass traditional access controls and directly inject malicious commands into onboard microphones, posing a serious threat to vehicle safety. This paper first conducts an in-depth analysis of the attack surface of in-vehicle voice control systems, focusing on the technical principles, attack chain, and specific security risks posed by laser voice attacks to vehicles. It also outlines the existing defensive measures available in the market, analyzes the advantages and limitations of current technologies, and proposes a solution based on the spectral consistency of voice signal physical layers and environmental features to enhance the overall security of the system.

Keywords: Intelligent connected vehicles; laser voice injection; counterfeit attacks.

1. Introduction

With the rapid development of automotive intelligence, the global automotive industry is undergoing a profound transformation driven by the concept of “Software-Defined Vehicle” (SDV). In this wave of change, intelligent connected vehicle (ICV) are no longer merely means of transportation but have evolved into “smart terminals on four wheels” [1] that integrate advanced sensors, controllers, actuators, and complex communication technologies. In this transformation, Human-Machine Interaction (HMI) systems are critical to enhancing user experience and driving safety. Traditional physical buttons and touchscreen interaction methods, when faced with increasingly complex in-vehicle functions, can easily distract drivers and pose safety risks. Therefore, voice control systems based on natural language processing (NLP) have quickly become the core and mainstream of modern automotive HMI design, thanks to their unique advantage of “hands-free, eyes-free” operation. Drivers can use simple voice commands to perform functions such as navigation settings, multimedia playback, air conditioning adjustment, window control, and even switching between certain advanced driver-assistance systems (ADAS) [2]. While enjoying the convenience of voice interaction, a serious safety issue has emerged, posing a new threat to autonomous driving safety [3].

As the “auditory organ”of the voice control system, in-vehicle microphones are typically placed in a position of absolute trust in system design, with their core task being to capture external acoustic signals with the highest possible fidelity. This design philosophy assumes that all signals entering the microphone originate from legitimate sound waves, but overlooks the need to identify the physical source of the signals. This has created a new, virtually unprotected attack surface for in-vehicle voice interfaces. In recent years, attack techniques targeting voice interfaces have made breakthrough progress. From initial recording-replay attacks to “DolphinAttack” [4], which uses ultrasonic waves inaudible to the human ear to carry malicious commands, attack methods are becoming increasingly covert, precise, and difficult to defend against. In particular, laser voice attacks exploit the photoacoustic effect of microphone diaphragms. By using a laser beam modulated with audio signals, attackers can silently cause the microphone to “hear” any pre-set commands from several meters or even farther away. Attackers do not need physical contact with the vehicle or to crack any wireless communication protocols (e.g., Bluetooth, Wi-Fi) to directly issue control commands to the vehicle.

Therefore, a novel detection framework integrating spectral waveform consistency verification is proposed as a future research direction to address high-fidelity synthetic voice and physical injection attacks, aiming to provide theoretical references and technical insights for constructing safer and more reliable in-vehicle voice interaction systems.

2. Relevant technical foundation

2.1 Architecture of in-vehicle voice control system

The in-vehicle voice control system is a complex engineering chain that converts acoustic signals into specific vehicle actions. Its typical architecture is shown in Figure 2.1 and primarily consists of the following core components:

2.1.1 Microphone: As the system's “sensing” front end, it is typically an electret condenser microphone (ECM) or micro-electromechanical system (MEMS) microphone. Its role is to capture sound waves emitted by the driver or passengers and convert them into weak analog electrical signals.

2.1.2 Analog-to-Digital Converter (ADC): Samples and quantizes the analog electrical signals output by the microphone, converting them into a digital signal stream.

2.1.3 Digital Signal Processing (DSP): Performs preprocessing on the digital signal, including noise reduction, acoustic echo cancellation (AEC), beamforming, etc., to enhance the clarity and signal-to-noise ratio of the voice signal.

2.1.4 Automatic Speech Recognition Engine (ASR): This is the system's “cognitive” core. The ASR engine uses acoustic models and language models to decode the pure speech signal stream into specific text commands (e.g., “open the sunroof”).

2.1.5 Command & Control Unit: This unit is responsible for parsing the text commands output by the ASR engine and mapping them to specific control messages on the in-vehicle network (e.g., CAN bus or in-vehicle Ethernet), which are then executed by the corresponding actuators.

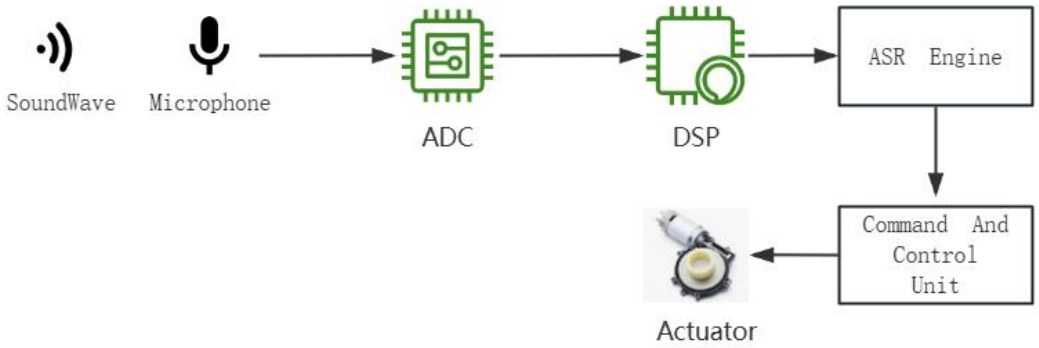

Fig. 2.1 Typical architecture of voice control system

The core vulnerability of this architecture lies in its misplaced trust foundation. The entire system is designed based on a fundamental assumption: any signal transmitted through the microphone that can be successfully recognized by the ASR engine originates from a legitimate user inside the vehicle. The system uses complex DSP algorithms to remove noise and ASR models to understand semantics, but it lacks a mechanism to verify the legitimacy of the physical source of the signal. The system cannot distinguish between a genuine voice signal generated by vocal cord vibrations and airborne transmission, and a “fake voice” signal simulated by other physical effects. This trust blind spot at the physical layer and analog circuit layer opens the door to new injection-based attacks such as laser voice attacks.

3. Analysis of Laser-based Voice Spoofing Attacks on Intelligent Connected Vehicles

3.1 Technical principles of laser voice attacks

The implementation of laser voice attacks does not involve cracking software or encryption algorithms, but rather cleverly exploiting a fundamental physical phenomenon known as the photoacoustic effect.

The core component of a microphone is a diaphragm that is extremely sensitive to changes in pressure. During normal operation, pressure changes from sound waves drive the diaphragm to produce mechanical vibrations, which in turn alter the electrical properties of capacitive or piezoelectric materials, ultimately generating corresponding electrical signals. In contrast, laser voice attacks utilize light to directly drive diaphragm vibrations. The specific principle is as follows [5]: When a laser beam with modulated intensity is directed at the microphone's diaphragm, the light energy is absorbed by the diaphragm material. This absorbed energy is rapidly converted into thermal energy, causing a local increase in temperature and resulting in thermal elastic expansion. If the laser intensity varies over time (e.g., modulated according to the waveform of an audio signal), the absorption-heating-expansion process of the diaphragm also changes synchronously, thereby generating forced mechanical vibrations at the same frequency as the modulated signal. This acoustic vibration induced by light energy is known as the photoacoustic effect.

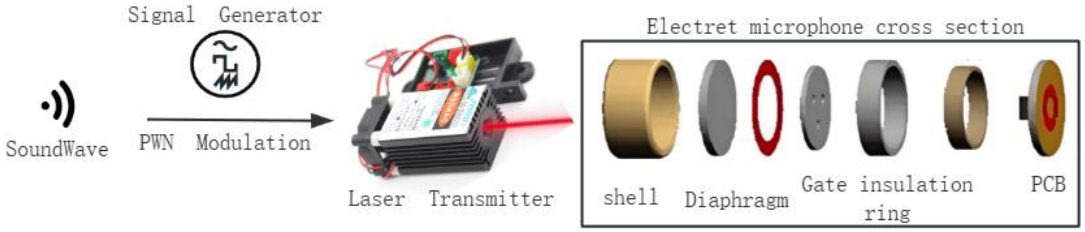

Fig. 3.1 Schematic diagram of laser voice attack principle

As shown in Figure 3.1, the left side shows an audio waveform. After debugging the waveform using a signal generator, the signal is transmitted to a laser emitter, which varies the intensity of the laser beam emitted in accordance with the audio waveform. The right side shows a cross-sectional view of an electret microphone, where the laser illuminates the diaphragm, causing it to vibrate.

For the microphone, it cannot distinguish whether this vibration is caused by air pressure waves or by the photoacoustic effect. It simply converts this mechanical vibration into an electrical signal and transmits it to the subsequent circuitry. Therefore, an attacker need only use a malicious voice command (such as “open the car door”) as a modulation signal, load it onto a laser beam precisely focused on the vehicle's microphone, and the vehicle's voice control system will “hear” this command and execute the corresponding operation. This process completely bypasses air as the traditional medium, achieving remote, silent command injection.

3.2 Attack scenario

A complete laser voice attack typically follows a clear attack chain and can pose specific threats to intelligent connected vehicles in multiple dimensions.

3.2.1 Reconnaissance & Targeting: Attackers first need to determine the precise location of the target vehicle's microphones. This can be achieved by consulting vehicle maintenance manuals, online teardown videos, or through close observation in environments such as parking lots. The front-facing microphones inside the vehicle (typically located above the rearview mirror or on the overhead light control panel) are the primary targets.

3.2.2 Command Generation: Attackers record or synthesize target commands using text-to-speech (TTS) technology. To improve the recognition rate of the ASR engine, attackers may add wake words or use generic, standard vehicle control commands.

3.2.3 Laser Modulation: The generated audio signal is input into a laser driver circuit to modulate the laser beam emitted by a laser transmitter.
3.2.4 Remote Injection: Attackers use aiming devices such as telescopes or telephoto lenses within the vehicle's line of sight to precisely focus the modulated laser beam onto the opening of the target microphone. Once aligned, the command is silently injected into the system, which then executes the command content.

3.3 Technological Evolution and Defense

Laser voice attack technology itself is also constantly evolving. Early demonstrations may have required relatively bulky equipment and short distances, but with the development of laser technology and optical components, future attacks may be more difficult to defend against, such as using infrared lasers invisible to the human eye, achieving longer attack distances (hundreds of meters), and making attack devices smaller and more portable [6], which poses a huge challenge to defense:
3.3.1 Comprehensive defense: The number of microphones on vehicles is increasing, distributed across various locations on the vehicle. Some vehicle models also have microphones outside the vehicle for external voice control. These exposed microphones provide attackers with more targets.
3.3.2 Cost and practicality: Physically reinforcing each microphone (e.g., installing filters) would significantly increase costs and design complexity, and may also impair normal audio performance.
3.3.3 Stealth of attacks: Attack processes are silent and invisible (especially infrared lasers), making it difficult for traditional intrusion detection systems (IDS) to detect the attack behavior itself.

In response to such threats, the current defense mechanism is multi-sensor fusion and array-based detection. Compared to passive physical blocking, cross-verification using multi-sensor information is a more proactive and intelligent hardware defense strategy. Vehicles typically deploy microphone arrays to achieve beamforming and noise suppression. Using this existing hardware, detection algorithms based on Time Difference of Arrival (TDoA) or Level Difference of Arrival (LDoA) can be developed[7]. A genuine sound source from a passenger inside the vehicle will exhibit subtle but physically consistent differences in the time and energy of sound waves reaching different microphones in the array. Laser attacks are typically highly directional and can only target a single microphone in the array. Therefore, if the system detects a clear voice signal with high energy in a single microphone channel while other channels exhibit extremely weak or irrelevant signals, it can be determined as a physical injection attack. The primary challenge of this method lies in the robustness of the algorithm, which requires precise microphone calibration and modeling of the complex acoustic environment inside the vehicle.

4. Design of an attack mechanism based on environmental feature fusion and spectral waveform consistency verification

4.1 Overall Architecture and Workflow

The core idea of the framework is to establish a “zero trust” audio reception mechanism. It no longer unconditionally trusts any signal entering the microphone, but treats it as an object that needs to be verified. The verification process consists of two steps: first, verifying whether the signal carries the “environmental fingerprint” that should be present in the current in-vehicle physical environment; second, conducting a deep analysis of suspicious signals to identify whether they were generated by human or non-acoustic means, and verifying whether they match the stored acoustic fingerprint information.

4.2 Spectrum consistency verification

This module is the first line of defense for the framework, and its goal is to be a fast, lightweight "physical authenticity" filter. It compares real-time voice commands with pre-recorded "environmentally aware acoustic fingerprints" to determine whether the signal has undergone a real acoustic propagation process inside the vehicle.

4.3 Voiceprint registration

When a user first uses the system or under specific conditions (such as after changing vehicle owners), the system enters a brief registration phase. Users follow the prompts to recite several designated reference phrases in a quiet vehicle environment. The system does not merely store the user's voiceprint characteristics but focuses on analyzing and establishing a baseline model that represents the unique acoustic environment of the current vehicle. This "fingerprint" primarily includes the following information:

4.3.1 Key features of the Room Impulse Response (RIR): Primarily reflected in the reverberation time (RT60) and direct-to-reverberant ratio (DRR) of the voice. These parameters quantify the size, shape, and sound-absorbing/reflective characteristics of interior materials (glass, leather, fabric) within the vehicle space.

4.3.2 Cabin's inherent noise profile: Records the background noise spectrum patterns under typical operating conditions such as engine idle and normal air conditioning operation.

4.3.3 Microphone array spatial signature: For multi-microphone systems, records the statistical patterns of energy and phase differences between different channels for real sound sources

4.4 Receive verification processing

4.4.1 Reverberation feature similarity: The algorithm quickly calculates the RT60 or DRR of the current voice frame and determines whether it falls within the preset confidence interval. A "dry" signal injected by a laser, which has almost zero reverberation, will immediately cause this verification to fail.

4.4.2 Spectral envelope similarity: Compare the spectral envelope of the current voice signal with the baseline model for consistency. Real voice signals produce specific comb filtering effects due to environmental reflections, which leave traces in the spectrum.

4.4.3 Spatial Consistency Verification: For microphone arrays, verify whether the signals from each channel meet the expected TDoA/LDoA model. The single-point injection characteristic of laser attacks is clearly exposed here.

4.4.4 Voiceprint Consistency Verification: For sounds that pass the above environmental perception tests, further compare the consistency between the input voiceprint and the user-registered voiceprint.

The advantage of this approach lies in its efficiency. It relies on clear physical principles for judgment, resulting in low computational overhead, and can quickly filter out the vast majority of physical injection attacks and replay attacks (e.g., voice commands recorded in other scenarios) without requiring complex neural network judgments.

5. Safety Improvement Assessment

This defense framework offers the following significant advantages:

1. Defense-in-Depth: Combining physical environment verification with voiceprint verification creates two complementary lines of defense, greatly increasing the difficulty for attackers to bypass the system.

2. High Efficiency and Low Latency: The lightweight physical environment verification can handle the majority of inputs, with only suspicious traffic entering the computationally intensive

acoustic fingerprint comparison. This optimizes system resource utilization while ensuring security, meeting the real-time requirements of in-vehicle systems.

3. High Robustness and Adaptability: The environmental fingerprint registration mechanism enables this solution to adapt to any specific vehicle model and internal environment.

4. Forward-Looking: The design of this solution is not limited to defending against known laser attacks. Its core "physical source authentication" concept provides an expandable solution for defending against any unknown attacks that may arise in the future that attempt to simulate sound waves at the physical level (such as ultrasound, vibration, etc.).

In summary, the novel defense scheme proposed in this paper combines acoustic environmental physics with acoustic fingerprint verification to provide a feasible and forward-looking technical approach for addressing security risks in voice interfaces of intelligent connected vehicles.

6. Conclusion

This paper investigates the problem of counterfeit command attacks targeting in-vehicle voice interfaces, analyzing the technical principles, implementation methods, and multi-dimensional security risks posed to vehicles. It further reviews existing defense mechanisms, noting that traditional solutions relying on physical shielding or content feature analysis are inadequate when faced with new types of attacks that can perfectly replicate voice content while only differing in physical sources. To address this severe challenge, this paper proposes a layered defense scheme. The core of this scheme lies in a fundamental shift in security philosophy, from traditional "content verification" to a deeper level of "physical source authentication." By combining "spectral waveform consistency verification" used to identify the acoustic environment propagation signature with "voiceprint verification identification," the scheme ensures data efficiency and model robustness while establishing a layered defense system capable of effectively resisting physical injection attacks. However, the maturity and implementation of this technical approach still face multiple challenges, including adaptation to high-dynamic acoustic environments, limitations on onboard processor resources, and an ongoing "arms race" with attackers. Nevertheless, these challenges themselves point the way forward for future research, such as multimodal fusion defense, self-supervised learning, and continuous learning, which are all highly valuable areas of exploration. In summary, ensuring the security of voice interaction interfaces is no longer merely a technical task; it is also closely related to the safety assurance that intelligent connected vehicles provide to users in the era of human-machine co-driving and fully autonomous driving.

References

[1] Shugang Jiang. Vehicle E/E Architecture and Key Technologies Enabling Software-Defined Vehicle, 2024.

[2] Henrik Detjen, Sarah Faltaous, Bastian Pfleging, Stefan Geisler &Stefan Schneegass Pages . How to Increase Automated Vehicles' Acceptance through In-Vehicle Interaction Design: A Review, 01 Jan 2021:308-330.

[3] XiLL, LinSH, WangZ, XieTG, SunYY, ZhuHS, SunLM. Autonomous Driving Security of Intelligent Connected Vehicles:Threats, Attacks, and Defenses. RuanJianXueBao/Joumalof Software, 2025,36(4):1859-1880.

[4] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, Wenyuan Xu. DolphinAttack: Inaudible Voice Commands. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security,2017:103–117.

[5] Takeshi Sugawara, Benjamin Cyr, Sara Rampazzi, Daniel Genkin, Kevin Fu. Light Commands: Laser-Based Audio Injection Attacks on Voice-Controllable Systems, 2020.

[6] You Li, Javier Ibanez-Guzman. Lidar for Autonomous Driving: The Principles, Challenges, and Trends for Automotive Lidar and Perception Systems.IEEE Signal Processing Magazine (Volume: 37, Issue: 4, July 2020) :50 - 61.

[7] Michael Brandstein, Darren Ward. Microphone Arrays: Signal Processing Techniques and Applications, 2001.