ISSN: 2688-8645

Anti-Spoofing Evaluation Benchmark and Layered Defense Strategy for In-Cabin Voiceprint Payment under Voice Cloning Attacks

Benyin Wu [a], Ruizhi Feng [b], Ziyi Wang [c], Yihong Qin *

CATARC Intelligent Technology (Tianjin) Co., Ltd. Tianjin, China

* Corresponding author: Yihong Qin (Email: qinyihong@catarc.ac.cn),

[a] wubenyin@catarc.ac.cn, [b] fengruizhi@catarc.ac.cn, [c] wangziyilucky@catarc.ac.cn

Abstract

With the proliferation of voice interaction in intelligent cabins, voiceprint payment is becoming an important payment method in vehicular scenarios. However, deep learning-driven voice cloning technology has significantly lowered the barrier for generating forged speech, posing a serious threat to voiceprint authentication. The enclosed cabin environment, multi-microphone array layout, and real-time requirements make traditional anti-spoofing solutions difficult to migrate directly. This paper systematically analyzes the technological evolution of voice cloning attacks and the in-cabin threat model, constructs a multi-dimensional anti-spoofing evaluation benchmark covering detection performance, user experience, system overhead, and robustness, and proposes a four-layer defense strategy comprising input-layer liveness detection, feature-layer multimodal fusion, decision-layer risk grading, and application-layer transaction control. Experiments show that the proposed framework can reduce the attack success rate to 0.9%, while maintaining a false rejection rate of 0.7% for legitimate users and an authentication latency of 847 ms, providing a systematic reference for the security evaluation and protection of in-cabin voiceprint payment.

Keywords

Voice Cloning; Voiceprint Payment; Anti-Spoofing; Intelligent Cabin; Evaluation Benchmark; Layered Defense.

1. Introduction

In recent years, the intelligent connected vehicle industry has developed rapidly, and voice interaction has become the core entry point of intelligent cabins. Voiceprint payment utilizes speaker voice features for identity authentication and payment authorization. Due to its advantages of requiring no manual operation and enabling secure transaction completion during driving, it has received widespread attention from automakers and payment institutions. Companies such as Cerence have launched in-vehicle voice payment solutions, and multiple domestic automakers have piloted voiceprint unlocking and voice payment functions in high-end models.

However, the widespread deployment of voiceprint payment is accompanied by severe security challenges. Deep learning technologies represented by diffusion models and neural vocoders have drastically lowered the threshold for voice cloning. An attacker only needs a few seconds of target voice samples to generate highly realistic forged speech. In the first half of 2024, multiple fraud cases using AI voice cloning were uncovered in China, with the highest single-case loss reaching millions of yuan. In vehicular scenarios, user voice data may be leaked through the infotainment system or third-party services. Once an attacker successfully clones

ISSN: 2688-8645

the vehicle owner's voiceprint, they can play forged speech inside the vehicle to initiate unauthorized payments.

The unique characteristics of the cabin environment further exacerbate security risks: reverberation in enclosed spaces may mask frequency-domain anomalies of forged speech; multi-microphone arrays provide attackers with more signal injection points; and driving scenarios impose high real-time requirements on payment authentication. Existing voiceprint anti-spoofing research mainly revolves around challenges such as ASVspoof, lacking targeted design for the special acoustic environment of cabins and payment application scenarios. Moreover, most research focuses on optimizing single detection algorithms, lacking systematic multi-layer defense architectures. To address the above issues, this paper constructs an in-cabin voiceprint payment threat model, designs an anti-spoofing evaluation benchmark, and proposes a layered defense strategy.

2. Related Work and Technical Background

2.1. Evolution of Voice Cloning Attack Technology

Voice cloning technology has roughly gone through three stages. Early concatenative synthesis and parametric synthesis methods required large amounts of high-quality recordings, resulting in high cloning costs and limited naturalness. The second stage, represented by end-to-end TTS models such as Tacotron and FastSpeech combined with neural vocoders such as WaveNet and WaveGlow, significantly improved the naturalness of synthesized speech, but still required hours of target speaker recordings for fine-tuning. The third stage is the current zero-shot voice cloning era, represented by models such as VALL-E and XTTS, which only require 3 to 10 seconds of reference speech to clone the speaker's timbre, prosody, and emotional characteristics. The generated speech is already indistinguishable from human speech in subjective listening tests. Additionally, voice conversion (VC) technology can convert speaker identity while preserving the original speech content, further enriching attack methods.

2.2. Current State of Voiceprint Anti-Spoofing Research

Voiceprint anti-spoofing research aims to detect whether input speech is forged. The ASVspoof series of challenges has greatly advanced the development of this field, from synthesized speech detection in 2015, to simultaneously covering logical access and physical access in 2019, and then introducing deeper spoofing attacks and cross-channel evaluation in 2021. Existing detection methods can be classified into three categories: handcrafted feature-based methods (e.g., CQCC, LPCC) that exploit anomalies of forged speech in high-frequency bands and phase spectra; deep neural network-based methods (e.g., ResNet, LCNN) that automatically learn discriminative features; and self-supervised pre-trained model-based methods that demonstrate excellent cross-domain generalization capabilities. However, most of these studies target telephone channels or general far-field scenarios, lacking targeted design for cabin environments. [1]

3. Threat Modeling for In-Cabin Voiceprint Payment

3.1. Attack Surface Analysis

The attack surface of in-cabin voiceprint payment systems can be divided into three layers. Acquisition-end attacks include: playing forged speech inside the cabin through vehicle speakers or portable Bluetooth speakers; signal injection using multi-microphone arrays; and tampering with microphone output through near-field coupling [2]. Transmission-end attacks include: replaying or tampering with voice data packets between the infotainment system and cloud authentication services; and exploiting infotainment system vulnerabilities to obtain intermediate results of voice processing. Server-end attacks include: adversarial sample attacks

ISSN: 2688-8645

against voiceprint models; recovering registered voiceprint templates through model inversion; and bypassing detection using logical flaws in authentication interfaces. Among these, voice replay and cloned voice playback at the acquisition end are the most realistic and easiest-to-implement attack methods, and are the focus of this paper's defense. [3]

3.2. Typical Attack Scenarios

Based on attack methods and implementation difficulty, three types of scenarios can be identified. Scenario 1 is replay attack: the attacker records the user's payment instruction speech and replays it inside the vehicle. This has the lowest technical threshold, but the replayed speech exhibits detectable frequency-domain distortion after secondary conversion. Scenario 2 is synthesized speech attack: the attacker uses general TTS to synthesize payment instruction speech, where the timbre does not fully match the target, making it usually difficult to pass the voiceprint matching stage. Scenario 3 is cloned speech attack: the attacker obtains target voice samples through social media or data breaches, uses zero-shot cloning models to generate highly consistent forged speech, and plays it inside the vehicle [6]. This type of attack is the most dangerous because forged speech is highly close to real speech in the voiceprint feature space, making it difficult for traditional voiceprint matching to distinguish. Table 1 presents the detection difficulty grading of typical attack types.[4, 5]

Table 1. Typical Attack Types and Detection Difficulty Grading

Attack Type	Voiceprint Matching Difficulty	Anti-Spoofing Detection Difficulty	Overall Threat Level
Replay Attack	Low (exact match)	Medium (frequency distortion detectable)	Medium
TTS Synthesis Attack	High (timbre mismatch)	Low (synthesis artifacts obvious)	Low
VC Conversion Attack	Medium (partial timbre match)	Medium (conversion artifacts detectable)	Medium-High
Zero-Shot Cloning Attack	Very Low (high match)	High (forgery traces concealed)	High
Adversarial Sample Attack	Very Low (targeted bypass)	Very High (feature space attack)	Very High

4. Anti-Spoofing Evaluation Benchmark Design

4.1. Evaluation Metric System

Table 2. Anti-Spoofing Evaluation Metric System for In-Cabin Voiceprint Payment

Metric Category	Metric Name	Symbol	Definition and Description
Detection Performance	Equal Error Rate	EER	Error rate when false alarm rate equals miss rate
Detection Performance	Tandem Detection Cost	t-DCF	Normalized cost of tandem voiceprint authentication and anti-spoofing
Detection Performance	Attack Success Rate	ASR	Proportion of forged speech successfully passing authentication
User Experience	False Rejection Rate	FRR	Proportion of genuine user speech incorrectly rejected
User Experience	Authentication Latency	Latency	Time from speech end to result output (ms)
System Overhead	Model Parameters	Params	Trainable parameter scale of detection model (M)
Robustness	Cross-Scene Performance Degradation	ΔEER	Maximum EER increase under different noise/reverberation

ISSN: 2688-8645

Anti-spoofing evaluation for in-cabin voiceprint payment needs to comprehensively consider detection performance, user experience, and system overhead. This paper constructs an evaluation system with four categories of metrics (Table 2): detection performance metrics include Equal Error Rate (EER), tandem Detection Cost Function (t-DCF), and Attack Success Rate (ASR), where ASR directly reflects the probability of an attacker successfully completing unauthorized payment and is the most concerning security metric in payment scenarios; user experience metrics include False Rejection Rate (FRR) for legitimate users and authentication latency; system overhead metrics include model parameter count and inference computation, used to evaluate deployment feasibility on in-vehicle embedded platforms; robustness metrics evaluate the performance stability of the solution under different noise and reverberation conditions. [7]

4.2. Dataset Construction and Evaluation Pipeline

Existing public anti-spoofing datasets lack targeted design for cabin acoustic environments. Constructing an in-cabin test dataset should follow: acoustic environment authenticity (collected in real vehicles, covering different vehicle types and conditions), comprehensive attack types (covering replay, TTS, VC, and zero-shot cloning), sufficient speaker coverage (adequate number of registered and attacking speakers), payment instruction relevance (including real payment sentence patterns), and reproducibility (providing complete collection protocols and evaluation code [8, 9]). The standardized evaluation pipeline includes: speaker-independent data partitioning, unified baseline system configuration, score fusion, ASR calculation at a fixed FRR=1% operating point, robustness testing under different signal-to-noise ratios, and complete metric reporting.

5. Layered Defense Strategy

A single anti-spoofing detection algorithm is difficult to cope with diverse attack methods. This paper proposes a four-layer defense framework: input layer — feature layer — decision layer — application layer. Each layer operates independently and provides synergistic gains, forming full-chain protection from signal acquisition to transaction authorization.

5.1. Input Layer: Liveness Detection and Signal Quality Assessment

The input layer performs preliminary screening before speech enters the authentication pipeline. Liveness detection exploits the differences in physical characteristics between genuine human speech and played-back speech: speech produced by human vocal cord vibration has a stable harmonic structure, while speaker-replayed speech exhibits significant energy attenuation and phase distortion in the high-frequency band (>8 kHz) [10]. Meanwhile, spatial information from multi-microphone arrays is utilized — when a genuine person speaks, there are time-delay differences and energy differences between microphones consistent with the sound source position, while forged speech played by a single speaker does not match the genuine person spatial pattern. The signal quality assessment module evaluates metrics such as signal-to-noise ratio, clipping, and echo in real time; low-quality signals trigger warnings or require the user to speak again.

5.2. Feature Layer: Multimodal Fusion and Robust Feature Extraction

The feature layer improves discriminative capability by extracting multi-dimensional complementary features and fusing multimodal information. For acoustic features, in addition to MFCC, CQCC, LPCC, and group delay features that are more sensitive to forged speech are introduced; for deep features, self-supervised pre-trained models such as wav2vec 2.0 and HuBERT are used to extract context-dependent representations with strong cross-domain generalization capabilities. For multimodal fusion, in-vehicle cameras are combined to capture

ISSN: 2688-8645

lip movements and facial expressions. Since voice cloning typically only forges acoustic signals and cannot synchronously forge lip movements, audio-visual fusion can significantly improve detection accuracy. Additionally, cabin information such as seat pressure sensors and steering wheel capacitive sensors can be fused to enhance the reliability of identity authentication. [11, 12]

5.3. Decision Layer: Risk Grading and Dynamic Thresholding

The decision layer performs comprehensive risk assessment based on detection results and contextual information, dynamically adjusting authentication thresholds. Payment transactions are classified by amount into low-value (<100 CNY), medium-value (100–1000 CNY), and high-value (>1000 CNY), with different levels corresponding to different authentication stringency [13]. The system dynamically elevates the risk level based on real-time risk signals (abnormal anti-spoofing scores, critical voiceprint matching, presence of unregistered occupants). Meanwhile, user behavior profile anomaly detection is introduced, analyzing historical habits of payment time, location, and amount ranges; when the current transaction significantly deviates from the behavior pattern, additional verification is triggered.

5.4. Application Layer: Transaction Limits and Multi-Factor Verification

As the last line of defense, the application layer implements controls at the transaction authorization stage. Main measures include: single and cumulative transaction limits (e.g., 500 CNY per transaction, 2000 CNY daily cumulative), with excess-limit transactions requiring alternative authentication methods; multi-factor secondary verification, where high-value transactions append a challenge-response mechanism (requiring the user to speak a random dynamic verification code) after voiceprint passes, which can effectively resist replay attacks; transaction confirmation and delayed settlement, with a 30-second cancellation window and delayed settlement for large transactions; security auditing and anomaly freezing, recording detailed logs and temporarily freezing the function upon consecutive failures or anomaly patterns. Table 3 compares the core mechanisms and overhead of each defense layer.

Table 3. Comparison of Layered Defense Strategies

Defense Layer	Core Mechanism	Primary Protection Target	Computational Overhead	Experience Impact
Input Layer	Harmonic detection + array spatial analysis + quality assessment	Replay attacks, low-quality signals	Low	Imperceptible
Feature Layer	Multi-feature fusion + self-supervised representation + audio-visual fusion	Synthesized/cloned speech, unknown attacks	Medium-High	Imperceptible
Decision Layer	Risk grading + dynamic threshold + behavior detection	Critical sample bypass, anomalous transactions	Low	Low
Application Layer	Transaction limits + challenge-response + audit freezing	Loss control after front-layer breach	Very Low	Medium

6. Experimental Validation and Analysis

6.1. Experimental Setup

Experiments are conducted in a simulated cabin environment. Data is collected from real cabins of 3 vehicle types (sedan, SUV, MPV), including 40 registered speakers (20 male, 20 female) and 20 attacking speakers. Genuine speech is collected under three conditions: stationary, idling,

ISSN: 2688-8645

and driving at 60 km/h. Attack samples include replay attacks (3 portable speakers), TTS synthesis (Tacotron 2, FastSpeech 2), VC conversion (AutoVC, StarGANv2-VC), and zero-shot cloning (VALL-E, XTTS v2, 5-second reference speech). The voiceprint baseline uses ECAPA-TDNN, and the anti-spoofing baseline uses ResNet-18, both pre-trained on VoxCeleb2 and ASVspoof 2019 and then fine-tuned.

6.2. Results and Analysis

Table 4 presents the experimental results for different defense configurations. The baseline system (voiceprint matching only) achieves an ASR as high as 89.3% under cloned speech attacks, completely unable to resist voice cloning. After adding feature-layer detection, the overall EER drops to 5.2%, and the cloning ASR drops to 23.7%, but the replay ASR remains at 15.8%. After combining the input layer and feature layer, the replay ASR significantly drops to 3.1%, verifying the targeted effect of the input layer on replay attacks. After adding decision-layer dynamic thresholding, the overall ASR drops to 2.8%, with FRR controlled at 0.8%. Under the complete four-layer defense, the overall ASR drops to 0.9%, the cloning attack ASR is 1.7%, the replay ASR is 0.3%, and the end-to-end latency is 847 ms, meeting real-time requirements. The results demonstrate significant synergistic gain effects between layers.

Table 4. Comparison of Combined Effects of Defense Layers (FRR=1% Operating Point)

Defense Configuration	EER(%)	Replay ASR(%)	Cloning ASR(%)	Overall ASR(%)	FRR(%)	Latency(ms)
Baseline (voiceprint only)	—	92.1	89.3	87.5	1.0	312
+ Feature-layer detection	5.2	15.8	23.7	19.2	1.0	586
+ Input-layer liveness detection	3.1	3.1	18.5	11.8	0.9	673
+ Decision-layer dynamic threshold	2.4	1.8	6.2	2.8	0.8	751
Complete four-layer defense	1.6	0.3	1.7	0.9	0.7	847

7. Conclusion and Future Work

This paper addresses the security issue of in-cabin voiceprint payment under voice cloning attacks, systematically analyzes the evolution of attack technology and the in-cabin threat model, designs a multi-dimensional anti-spoofing evaluation benchmark, and proposes a four-layer defense strategy. Experiments show that the complete defense system can reduce the attack success rate to 0.9%, while maintaining a low false rejection rate and acceptable latency, verifying the effectiveness of the proposed solution.

Future work will focus on the following directions: researching lightweight anti-spoofing models for in-vehicle embedded platforms, reducing computational overhead through model compression and knowledge distillation; exploring cross-vehicle collaborative training under the federated learning framework, improving generalization while protecting voice data privacy; researching semantic-level anomaly detection based on large language models, identifying attack behaviors by analyzing the semantic rationality of payment instructions; and promoting the development of in-cabin voiceprint payment security standards, facilitating the industrial implementation of evaluation benchmarks and defense solutions.

ISSN: 2688-8645

References

[1] Todisco, M., Delgado, H., & Evans, N. W. D. (2016). A constant Q cepstral coefficients countermeasure for anti-spoofing in automatic speaker verification. In Interspeech 2016 (pp. 1428–1432). https: // doi. org/10.21437/Interspeech.2016-1131.

[2] Kinnunen, T., Delgado, H., Evans, N., et al. (2019). ASVspoof 2019: future horizons in spoofed and fake audio detection. In Interspeech 2019 (pp. 1003–1007). https://doi.org/10. 21437/ Interspeech. 2019-1431.

[3] Yamagishi, J., Wang, X., Todisco, M., et al. (2021). ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. In ASVspoof 2021 Workshop. https://doi.org/10. 21437/ ASVSP OOF. 2021-1.

[4] Wang, Y., Deng, J., Xue, L., et al. (2023). VALL-E: neural codec language models for zero-shot text-to-speech. In ICLR 2023. https://doi.org/10.48550/arXiv.2301.02111.

[5] Cocharan, J., de Andrade, L. C., Junior, H. M., et al. (2022). YourTTS: towards zero-shot multi-speaker TTS and zero-shot voice conversion. In ICASSP 2022 (pp. 5637–5641). https://doi.org/10. 1109/ ICASSP43922.2022.9747161.

[6] Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In Interspeech 2020 (pp. 3830–3834). https://doi.org/10.21437/Interspeech.2020-2050.

[7] Tak, H., Patino, J., Todisco, M., et al. (2021). End-to-end anti-spoofing with RawNet2. In ICASSP 2021 (pp. 6369–6373). https://doi.org/10.1109/ICASSP39728.2021.9414770.

[8] Chen, S., Wang, C., Chen, Z., et al. (2023). Robust anti-spoofing for speaker verification with self-supervised pre-trained models. In ICASSP 2023 (pp. 1–5). https://doi.org/10. 1109/ ICASS P49357. 2023. 10096129.

[9] Zhang, X. Y., Li, T., & Jia, H. R. (2022). Research progress of voice anti-spoofing technology. Acta Automatica Sinica, 48(5), 1153–1172.

[10] Zheng, F., & Liu, B. (2021). Voiceprint recognition technology and its application in the financial field. Financial Computer of China, 29(8), 12–18.

[11] Cerence Inc. (2020, July 30). Cerence Pay: secure voice-powered payments in the car [EB/OL]. Retrieved June 15, 2024, from https://cerence.com.

[12] Baser, O., et al. (2024). SecureSpectra: safeguarding digital identity from deep fake threats via intelligent signatures. arXiv preprint arXiv:2407.00913. https://doi.org/ 10.48550/ arXiv. 2407. 00913.

[13] State Administration for Market Regulation. (2022). GB/T 7713.2—2022 Rules for Academic Paper Writing. Standards Press of China.