Modulation-Based Backdoors: Leveraging Amplitude and Frequency Patterns to Attack Speaker Recognition

Hanbo Cai[1,2], **Pengcheng Zhang**[1*], **Yan Xiao**[3], **De Li**[4], **Hanting Chu**[5], **Ying Luo**[2]

[1]College of Computer Science and Software Engineering, Hohai University, Nanjing, Jiangsu, China
[2]College of Artificial Intelligence, Suzhou Vocational Institute of Industrial Technology, Suzhou, Jiangsu, China
[3]School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University, Shenzhen, China
[4]School of Computer Science and Engineering, Guangxi Normal University, Guilin, China
[5]School of Mathematics and Computer Science, Zhejiang Agriculture and Forestry University, Hangzhou, Zhejiang, China
caihanbo@hhu.edu.cn, pchzhang@hhu.edu.cn, xiaoy367@mail.sysu.edu.cn, lide@stu.gxnu.edu.cn, htchu@zafu.edu.cn, luoy@siit.edu.cn

Abstract

Deep neural networks (DNNs) are widely and successfully applied in the field of speaker recognition. However, recent studies reveal that these models are vulnerable to backdoor attacks, where adversaries inject malicious behaviors into victim models by poisoning the training process. Existing attack methods often rely on environmental noise or complex voice transformations, which are typically difficult to implement and exhibit poor stealthiness. To address these issues, this paper proposes two modulation-based backdoor attacks that leverage frequency modulation (FM) and amplitude modulation (AM) to construct audio triggers. In real-world scenarios, regular variations in frequency and amplitude are often imperceptible to human listeners, making the proposed attacks more covert. Experimental results show that our methods achieve high attack success rates in both digital and physical settings, while also demonstrating strong resistance to various state-of-the-art backdoor defenses.

Code — https://github.com/HanboCai/FSMA-ASMA

Introduction

With the widespread deployment of DNN-based speaker recognition systems in applications such as access control (Korzh et al. 2025) and voice assistants (Eberhart, Bansal, and McMillan 2020; Cai et al. 2023), growing attention has been drawn to their security vulnerabilities. Building a high-performance speaker recognition model typically requires large-scale, high-quality training data and considerable computational resources. As a result, individual developers and small-to-medium enterprises (SMEs) often resort to external resources—including pre-trained models, shared datasets (Zhou et al. 2023), or outsourced training platforms—to reduce development costs. However, the black-box nature of DNNs renders such third-party resources highly susceptible to security risks. As illustrated in Figure 1, the system represents a speaker recognition-based authentication setting in a banking scenario. In the benign model, the adversary's voice is rejected as it does not match any registered user in the system. However, in the backdoored model, the adversary successfully bypasses authentication by embedding a specific audio trigger, thus gaining unauthorized access (Yan, Lan, and Yan 2024). This scenario highlights the severe security risks posed by backdoor attacks in high-stakes applications like banking systems.

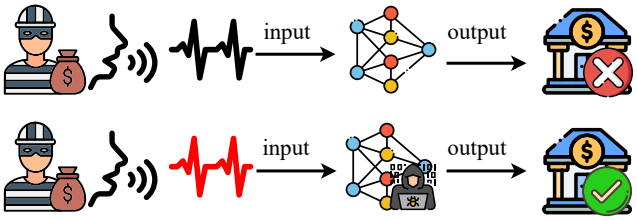


Figure 1: The adversary is an external user who cannot be recognized by the system under normal conditions. However, by implanting a backdoor, the adversary can deceive the speaker recognition system and bypass authentication.

In recent years, backdoor attacks have emerged as one of the most prominent threats to the security of DNNs (Li et al. 2022b; Chen et al. 2025a; Tan et al. 2025). In such attacks, adversaries inject carefully crafted poisoned samples during the training phase, implanting a hidden backdoor into the target model. These backdoors leave model performance on benign inputs unaffected but cause misbehavior when triggered, posing serious deployment risks. Prior work has demonstrated that backdoor attacks are also feasible in the audio domain. Some approaches embed additive noise patterns as triggers (Koffas et al. 2022; Zhai et al. 2021), while others exploit environmental background noise (Liu et al. 2022), elements of sound (Cai et al. 2024), or separable and distinctive audio clips to construct backdoors (Xin, Lyu, and Ma 2022; Luo et al. 2022). However, many of these methods suffer from critical limitations: they are either perceptible to human ears and vulnerable to common preprocessing or filtering defenses, or they are difficult to deploy reliably in real-world physical environments (Chen et al. 2024). This raises a fundamental question: *Can we design controllable*

*Pengcheng Zhang is the corresponding author.

audio signals that are imperceptible to humans but recognizable to machine learning models as backdoor triggers?

The answer to this question is affirmative. Prior studies have demonstrated that the effectiveness of a backdoor trigger is closely tied to the model's ability to consistently recognize its features (Yan, Lan, and Yan 2024). Motivated by this observation, we draw inspiration from modulation techniques commonly used in communication systems for signal transmission. Leveraging their inherent robustness (Kubo et al. 2011), we design smooth and globally structured amplitude and frequency modulation patterns as backdoor triggers. These modulation-based triggers exhibit strong resilience to signal distortion and compression, making them well-suited for deployment in real-world physical environments (Roder 1931). Based on this insight, we propose two acoustic modulation-based backdoor attacks: Frequency-based Speech Modulation Attack (FSMA) and Amplitude-based Speech Modulation Attack (ASMA). Crucially, our triggers introduce only minor, regular variations in the signal's frequency or amplitude, without relying on additive noise, complex feature engineering, or any modification of the speech content. This design preserves the naturalness of the audio to human listeners while maintaining model performance on benign inputs—thereby achieving both stealth and practical applicability.

The major contributions can be summarized as follows:

- We propose a new backdoor attack that modulates the amplitude and frequency of speech signals without introducing any audible noise. The attack preserves the semantic content of the original speech.
- Our method demonstrates high practicality, as it can be deployed in real-world scenarios with minimal effort, achieves high attack success rates, and remains stealthy enough to avoid detection by end users.
- We conduct extensive experiments on two benchmark datasets to validate the effectiveness and feasibility of our approach. In addition, our attack demonstrates strong resistance against several state-of-the-art defense techniques.

Related Works

Speaker Recognition

Speaker Recognition (SR) (Hanifa, Isa, and Mohamad 2021) aims to identify or verify a speaker's identity based on audio samples, typically formulated as either closed-set identification (CSI) or open-set identification (OSI) tasks. In CSI, the speaker is assumed to belong to a predefined set of enrolled users, and identification is conducted by comparing the input voiceprint to all registered templates to find the best match. By contrast, OSI further requires determining whether the speaker belongs to an unseen class, thereby increasing task complexity. Most existing backdoor attack studies targeting SR focus on the CSI scenario, where the adversaries craft trigger samples to induce the model to misclassify them as a specific enrolled user, thereby achieving identity impersonation or bypassing authentication.

The development of speaker recognition models has evolved from traditional statistical methods to deep learning-based approaches. In the early development of speaker recognition, Dehak *et al.* (Dehak et al. 2010) proposed the i-vector method, which extracts fixed-dimensional speaker representations from statistical features of speech using Gaussian Mixture Models (GMM) and factor analysis. This method achieved strong performance across a variety of speaker-related tasks. With the rise of DNNs, Variani *et al.* (Variani et al. 2014) introduced the d-vector approach based on RNNs or LSTMs. This method extracts frame-level features and applies average pooling to obtain speaker embeddings, but it remains sensitive to the phonetic content of the speech. Subsequently, Snyder *et al.* (Snyder et al. 2018) proposed the x-vector architecture, which employs Time-Delay Neural Networks (TDNNs) to extract frame-level representations and applies statistical pooling to generate fixed-dimensional embeddings. This architecture has been widely adopted in industry as a standard deep speaker representation model. Building on this architecture, Desplanques *et al.* (Desplanques, Thienpondt, and Demuynck 2020) introduced ECAPA-TDNN, which incorporates channel attention mechanisms, residual connections, and multi-scale feature aggregation to significantly enhance the model's robustness and recognition accuracy. Meanwhile, Jung *et al.* proposed the RawNet series (Jung et al. 2019, 2020, 2022), adopting an end-to-end architecture that models raw audio waveforms directly, bypassing traditional acoustic features.

Backdoor Attacks

Backdoor attacks represent an emerging yet critical training-time threat. In general, the adversary aims to maliciously manipulate the training process—such as by altering training samples or loss functions.

Backdoor attack in image. In the image domain, backdoor attacks have been extensively studied and can be categorized along two primary dimensions: label consistency and attack modality. Based on whether the poisoned samples retain their ground-truth labels, attacks are classified as either poison-label or clean-label. Poison-label attacks (Gu et al. 2019; Qi et al. 2023; Li et al. 2022a) tend to be more effective, as clean-label samples often contain class-consistent robust features that interfere with the model's ability to learn the trigger pattern (Gao et al. 2023). However, clean-label attacks (Zhu et al. 2025; Gao et al. 2023) offer stronger stealthiness, as poisoned samples are harder to detect by inspecting label-image consistency. From the perspective of attack modality, backdoor attacks can be divided into digital and physical types. Traditional digital attacks generate poisoned samples entirely in the pixel space. (Li et al. 2021) revealed that most digital attacks fail to maintain effectiveness in the physical world, and proposed a physical enhancement method based on spatial transformation. More recently, (Xu et al. 2023) demonstrated the feasibility of using spatial transformations ($e.g.$ rotation) with specific parameters as physical trigger patterns.

Backdoor Attack in Audio. In recent years, backdoor attacks in the audio domain have gradually attracted increasing attention from the research community. (Zhai et al.

2021) proposed the first backdoor attack framework targeting speaker verification, based on clustering techniques. (Ye et al. 2025) launched an attack against speaker recognition systems by altering the speaker's speaking rate. (Ye et al. 2024, 2023) proposed two novel trigger designs tailored for speaker recognition: one based on padding operations and the other on phase perturbation. (Koffas et al. 2022) explored the feasibility of using ultrasonic pulses and voice style transfer as audio triggers. (Zhang et al. 2024) proposed an inaudible frequency-domain backdoor attack that remains stealthy yet highly effective. Meanwhile, several studies investigated the use of natural environmental sounds—such as music or background noise—as trigger patterns (Liu et al. 2022; Xin, Lyu, and Ma 2022; Luo et al. 2022). (Shi et al. 2022) proposed an optimization-based strategy for generating more effective audio triggers. In addition, (Chen et al. 2024) introduced room impulse responses (RIRs) as a physical trigger mechanism, enabling backdoor activation without explicit signal injection.

Proposed Method

Threat Model

In this paper, we focus on poison-only backdoor attacks against speaker recognition models, where the adversary can only manipulate the training data it provides, without access to the training process or model parameters. We further assume that the adversary has no knowledge of the training details, such as the model architecture, optimization objectives, or training procedures. This results in a fully black-box setting with no control over the training process, which represents one of the most challenging threat models. Despite its difficulty, this setting poses a significant real-world threat, as it applies to common scenarios involving third-party datasets, pre-trained models, or outsourced training services.

Attack Overview

In this work, we propose two audio trigger design methods that are semantically preserving, physically deployable, and imperceptible to human ears: Frequency Modulation Trigger (FM) and Amplitude Modulation Trigger (AM). Both modulation strategies operate on low-level acoustic features of speech without relying on semantic information, making them effective across a wide range of model architectures. The frequency modulation trigger introduces structured fluctuations in the frequency of the speech signal to simulate pitch variations, creating model-learnable but human-imperceptible feature shifts. In contrast, the amplitude modulation trigger imposes periodic loudness variations in the time domain, generating subtle yet consistent energy perturbations that guide the model to learn a specific "loudness trajectory" as the backdoor trigger. In the following sections, we present the design rationale and implementation of each modulation strategy in detail.

Attack via Frequency Modulation

It is conceptually reasonable to construct backdoor triggers by directly modifying the frequency of speech signals, as frequency is one of the key acoustic features heavily relied upon by speaker recognition models. However, trigger patterns based solely on static or localized frequency perturbations often suffer from limited structural complexity and lack temporal continuity, making them difficult for models to learn reliably (Cai et al. 2024, 2025). Moreover, their effectiveness tends to degrade significantly in physical environments due to channel distortion or hardware frequency response limitations (Chen et al. 2024). By comparison, our proposed frequency modulation strategy introduces a globally controllable frequency variation curve, producing a temporally smooth and imperceptible perturbation pattern that is easily learnable by the model. This design not only improves the attack success rate but also significantly enhances robustness and deployability in real-world scenarios.

Specifically, we propose a time-domain frequency modulation method that constructs a modulation signal to control frequency variations over time and applies it directly to the raw waveform. Our method consists of three main steps: (1) designing a frequency modulation curve, (2) generating the corresponding phase trajectory, and (3) constructing the modulator and applying it to the input signal.

Designing a Frequency Modulation Curve. First, we construct a time-varying frequency modulation curve $f(t)$ to define the instantaneous oscillation rate of the modulator over time. This frequency curve is composed of multiple linear segments, forming a periodically rising and falling pattern, defined as follows:

$$f(t) = \begin{cases} f_{\min} + (f_{\max} - f_{\min})\frac{3t}{T}, & 0 \le t < \frac{T}{3}, \\ f_{\max} - (f_{\max} - f_{\min})\frac{3(t-T/3)}{T}, & \frac{T}{3} \le t < \frac{2T}{3}, \\ f_{\min} + (f_{\max} - f_{\min})\frac{3(t-2T/3)}{T}, & \frac{2T}{3} \le t \le T. \end{cases} \tag{1}$$

Here, T denotes the total duration of the speech signal, while f_{min} and f_{max} represent the minimum and maximum modulation frequencies ($e.g.$ 1Hz and 4Hz, respectively).

$$f(t) = \frac{1}{2\pi} \cdot \frac{d\phi(t)}{dt} \quad \Rightarrow \quad \phi(t) = 2\pi \int_0^t f(\tau) d\tau \tag{2}$$

Generating the modulation phase curve. Therefore, to generate the modulator based on the specified frequency curve, we integrate $f(t)$ to obtain the phase trajectory $\phi[t]$ at each time point.

$$\phi[t] = 2\pi \sum_{i=0}^{t} \frac{f[i]}{F_s} \tag{3}$$

Here, F_s denotes the audio sampling rate. The resulting phase curve reflects the evolution of the modulator's instantaneous "rotation angle," which determines the pacing of changes in the output waveform.

Constructing the modulator and applying the perturbation. Finally, we use the phase trajectory to generate a time-varying sinusoidal modulator as follows:

$$m(t) = \sin(\phi(t)) \tag{4}$$

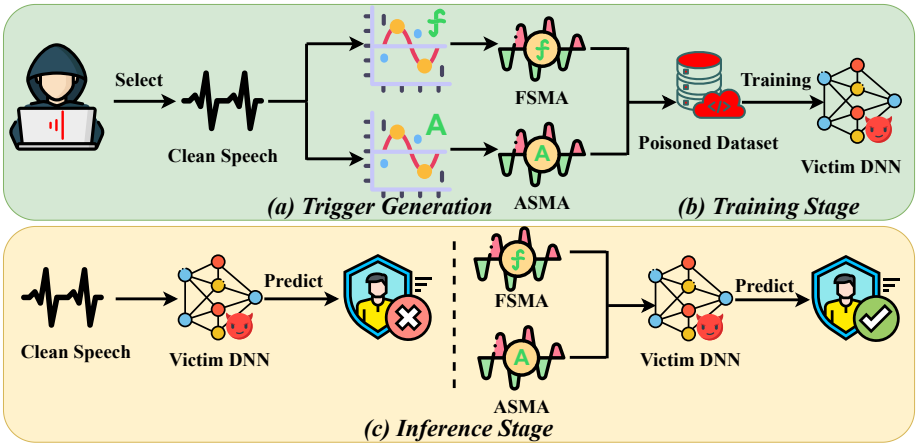


Figure 2: Attack pipeline: (a) Trigger Generation — Frequency and amplitude modulation produce audio triggers and poisoned samples; (b) Training — Poisoned data is injected to implant the backdoor; (c) Inference — Triggered audio causes the model to misclassify as the target user.

The modulator is a sinusoidal wave with time-varying frequency, whose oscillation rate is fully determined by the predefined $f(t)$. We apply it to the original speech signal $y(t)$ via the modulation operator $\mathcal{M}_m$, which modulates $y(t)$ in frequency according to $f(t)$, to obtain the modulated audio:

$$y_{\text{mod}}(t) = \mathcal{M}_m\{y(t)\} \tag{5}$$

Overall, the proposed frequency modulation trigger offers three key advantages: (1) its continuous and structured design facilitates model learning; (2) it introduces no semantic disturbance, preserving the naturalness and intelligibility of speech; and (3) it operates via simple time-domain multiplication, making it easily deployable in physical scenarios.

Attack via Amplitude Modulation

In addition to frequency perturbations, the loudness of speech—reflected in the temporal envelope of its amplitude—is also a key feature relied upon by speaker recognition models (Shetty 2016). In this section, we propose an amplitude modulation-based trigger design that leverages fine-grained control over the speech energy envelope to construct learnable acoustic perturbations without altering the underlying semantics.

We divide the modulation process into three main steps: (1) constructing the amplitude modulator and (2) applying the perturbation to the speech signal.

Constructing the Amplitude Modulator. We first construct a time-varying modulation envelope function $A(t)$ to control the short-term energy fluctuation of the speech signal. Specifically, the envelope is defined as a periodic sinusoidal wave that shapes the "loudness contour" of the audio over time:

$$A(t) = a_{\text{center}} + a_{\text{range}} \cdot \sin\left(2\pi n \cdot \frac{t}{T}\right) \tag{6}$$

Here, T denotes the total duration of the speech signal, n controls the number of modulation cycles within the duration, a_{center} is the average amplitude level around which the modulation occurs, and a_{range} defines the modulation depth, $i.e.$ how much the amplitude deviates from the center. We directly take the envelope function $A(t)$ defined in the previous step as the amplitude modulator. This modulator shares the same length and temporal resolution as the original speech signal $y(t)$, and controls the pointwise amplitude scaling over time.

Applying the Perturbation to the Speech Signal. Once the amplitude modulator $A(t)$ is constructed, we apply it to the original speech waveform $y(t)$ via the modulation operator $\mathcal{M}_A$ to obtain the modulated audio:

$$y_{\text{mod}}(t) = \mathcal{M}_A\{y(t)\} \tag{7}$$

Here $\mathcal{M}_A$ denotes amplitude modulation guided by the envelope $A(t)$; the operator encapsulates the modulation rule specified by $A(t)$.

This operation introduces a smooth, periodic variation in the energy contour of the audio without altering its semantic content. As the modulator shares the same temporal resolution as the input signal, the perturbation is temporally aligned and preserves the naturalness of the speech. The resulting waveform maintains intelligibility while embedding a subtle yet learnable acoustic pattern that serves as an effective backdoor trigger.

Overall, the proposed amplitude modulation trigger offers three key advantages: (1) its structured and periodic design introduces smooth energy variations that are easy for the model to learn; (2) it maintains the original speech semantics and perceptual naturalness, ensuring the trigger remains imperceptible to human listeners; and (3) it involves only simple time-domain scaling operations, enabling practical deployment in real-world physical settings.

Experiments

Main Settings

Models and Datasets. To comprehensively evaluate the effectiveness of our method across different types of neu-

Model	Dataset	Metric	No Attack	DABA	Ultrasonic	PhaseBack	SpeedMaster	FSMA (Ours)	ASMA (Ours)
d-vector	Librispeech	BA(%)	91.78	90.41	90.28	91.21	90.37	90.20	**91.65**
		ASR(%)	-	97.65	96.34	80.78	97.56	**98.57**	95.88
	VoxCeleb1	BA(%)	93.42	92.33	92.65	91.49	91.83	92.78	**93.12**
		ASR(%)	-	96.43	94.32	76.43	96.78	**97.67**	96.65
x-vector	Librispeech	BA(%)	99.92	97.13	98.71	98.01	99.18	98.72	**99.32**
		ASR(%)	-	**98.34**	95.88	86.19	94.13	97.73	94.76
	VoxCeleb1	BA(%)	98.73	97.18	**98.55**	96.76	97.88	98.21	97.89
		ASR(%)	-	**98.32**	96.45	85.31	97.32	97.05	97.98
RawNet3	Librispeech	BA(%)	98.89	97.63	97.32	97.05	98.44	98.13	**98.77**
		ASR(%)	-	98.76	93.21	82.56	94.32	**99.23**	96.62
	VoxCeleb1	BA(%)	98.76	98.43	98.20	96.91	98.36	98.27	**98.60**
		ASR(%)	-	**99.48**	94.36	81.34	95.61	98.88	96.39
ECAPA-TDNN	Librispeech	BA(%)	99.39	98.98	97.34	**99.01**	98.75	98.59	97.55
		ASR(%)	-	97.24	97.70	65.76	93.77	**98.68**	95.75
	VoxCeleb1	BA(%)	96.76	95.32	95.87	95.98	**96.41**	96.12	95.90
		ASR(%)	-	98.35	97.45	63.21	98.32	**98.43**	95.11

Table 1: The benign accuracy (BA) and attack success rate (ASR) of methods on the Librispeech and VoxCeleb1 dataset.

ral networks, we conduct experiments on two classic models (d-vector and x-vector) and two state-of-the-art models (RawNet3 and ECAPA-TDNN). In addition, the evaluation is performed on two widely used benchmark datasets in the speech domain: VoxCeleb1 (Nagrani et al. 2020) and LibriSpeech (Panayotov et al. 2015).

Baseline Selection. We compare our proposed FSMA and ASMA methods against six representative voice backdoor attacks: (1) Position-Independent Backdoor Attack (PIBA); (2) Dual Adaptive Backdoor Attack (DABA); (3) ultrasonic-based backdoor attack (referred to as "Ultrasonic"); (4) backdoor attack via phase injection (referred to as "PhaseBack"); (5) backdoor attack via voice style transfer (referred to as "JingleBack"); and (6) speech rate manipulation-based backdoor attack (referred to as "SpeedMaster").

Evaluation Metrics. Following the classical evaluation protocol in prior work (Gu et al. 2019), we use Benign Accuracy (BA) and Attack Success Rate (ASR) to assess attack effectiveness. Specifically, the BA measures the model's accuracy on clean samples, while ASR denotes the proportion of poisoned samples misclassified into the target label. Higher BA and ASR indicate a more effective attack. To evaluate stealthiness, we adopt the Mean Opinion Score (MOS). Additionally, we invite 30 volunteers to judge whether five poisoned audio samples from each attack sound natural. The proportion of samples perceived as natural is defined as Natural Rate (NR); a higher NR score indicates better stealthiness.

Attack Setup. For all attacks, we set the poisoning rate to 0.05. A target label is randomly selected from each dataset: "id1081" for LibriSpeech and "id10020" for VoxCeleb1. For our FSMA method, the modulation frequency ranges from 1Hz to 4Hz with a modulation cycle of 3. For the ASMA method, the amplitude modulation ranges from 0.3 to 2.0, with the same modulation cycle of 3.

Method	Clean	FSMA	ASMA
NR	100%	90.6%	96.6%
UTMOS	483	371	452

Table 2: Natural Rates (%) from Human Validation and MOS Scores from UTMOS.

Main Results

Attack Effectiveness. As shown in Table 1, our proposed FSMA and ASMA methods consistently achieve high attack success rates, exceeding 95% on both the Librispeech and VoxCeleb1 datasets, slightly lower than the DABA method. However, it is important to note that DABA compromises stealthiness—its perturbations are perceptible to human listeners and it causes a greater degradation in benign accuracy. In contrast, our methods introduce minimal impact on benign performance, with accuracy drops consistently kept within 1.5% compared to models trained on clean data. By comparison, existing approaches such as DABA, Ultrasonic, and PhaseBack result in more significant degradation of benign accuracy. These findings demonstrate that our attacks strike a favorable balance between effectiveness and stealthiness.

Attack Stealthiness. As shown in Table 2, we present the results of both UTMOS-based (Saeki et al. 2022) objective evaluation and human subjective evaluation. Specifically, we use UTMOS to assess 500 samples from the LibriSpeech dataset and count the number of samples with scores above 3. The results show that even after applying certain modulation to the clean speech, more than half of the samples receive scores greater than 3, indicating that the overall audio quality remains acceptable. In terms of human evaluation, over 95% of the samples generated by our two methods are perceived as natural, further demonstrating the strong stealthiness of our approach.

Model→ $\Delta f \downarrow$	d-vector	x-vector	RawNet3	ECAPA-TDNN
1.0	93.41	92.23	96.47	95.98
2.0	95.57	94.87	97.21	97.20
3.0	98.57	97.73	99.23	98.68
4.0	99.76	98.88	99.87	99.01
5.0	99.89	99.76	99.90	99.87

Table 3: The Attack Success Rate (%) $w.r.t.$ Different Frequency Modulation Range on the LibriSpeech dataset.

Model→ $\Delta a \downarrow$	d-vector	x-vector	RawNet3	ECAPA-TDNN
0.7	85.65	83.23	88.76	85.98
1.0	87.57	86.55	90.34	91.34
1.4	92.10	90.89	94.51	95.10
1.7	95.88	94.76	96.62	95.75
2.0	96.78	95.78	98.90	97.21

Table 4: The Attack Success Rate (%) $w.r.t.$ Different Amplitude Modulation Range on the LibriSpeech Dataset.

Ablation Study

Effects of the Poisoning Rate. We evaluate the impact of poisoning rates ranging from 1.0% to 7.0% across four model architectures. As shown in Figure 3, both FSMA and ASMA achieve steadily increasing attack success rates (ASR) as the poisoning rate increases. While a 5% poisoning rate is sufficient to achieve strong attack performance, a higher poisoning rate also leads to a slight degradation in benign accuracy (BA). This reveals a trade-off between ASR and BA, and adversaries should choose an appropriate poisoning rate based on their goals.

Effects of the Frequency Modulation Range. To evaluate the impact of frequency modulation strength, we conduct an ablation study on the LibriSpeech dataset. For FSMA, we fix $f_{\min} = 1$Hz and vary $f_{\max}$ to produce modulation ranges $\Delta f = f_{\max} - f_{\min}$ of $1.0, 2.0, 3.0, 4.0$. As shown in Table 3, we measure ASR under each setting and observe that larger Δf yields higher ASR, confirming that stronger modulation improves attack effectiveness. Nonetheless, overly high frequencies may introduce distortion, requiring a trade-off between performance and naturalness.

Effects of the Amplitude Modulation Range. To evaluate the impact of amplitude modulation range on attack performance, we conduct an ablation study on the LibriSpeech dataset using the d-vector model. The experiment follows the same setup as the main ASMA configuration (modulation cycles fixed at 3 and $a_{\min} = 0.3$), while varying the modulation range $\Delta a = a_{\max} - a_{\min}$ across $0.7, 1.0, 1.4, 1.7, 2.0$. As shown in Table 4, the ASR increases with larger Δa, indicating that stronger amplitude perturbation enhances backdoor activation. However, excessive modulation may lead to audio distortion, so adversaries must carefully balance between attack effectiveness and perceptual naturalness.

The Resistance to Potential Defenses

Recently, backdoor defenses have been extensively studied. We evaluate several representative, deployable defenses as benchmarks (Liu, Xie, and Srivastava 2017; Liu, Dolan-Gavitt, and Garg 2018; Gao et al. 2019), while noting that many emerging methods also merit systematic evaluation in future work (Chen et al. 2025b; Yi et al. 2025; Xu et al. 2024; Li et al. 2024; Hou et al. 2024, 2025).

The Resistance to Fine-tuning. As a representative backdoor removal technique, fine-tuning (Liu, Xie, and Srivastava 2017) aims to eliminate backdoors by retraining the model on a small set of local benign samples, motivated by the catastrophic forgetting property of deep neural networks. In our experiments, we use 10% of the benign training data and set the learning rate to 0.001. As shown in Figure 4, the ASR decreases as the number of fine-tuning epochs increases. However, even after fine-tuning, our proposed FSMA and ASMA attacks maintain ASR above 55%, indicating that they remain largely resilient to this defense.

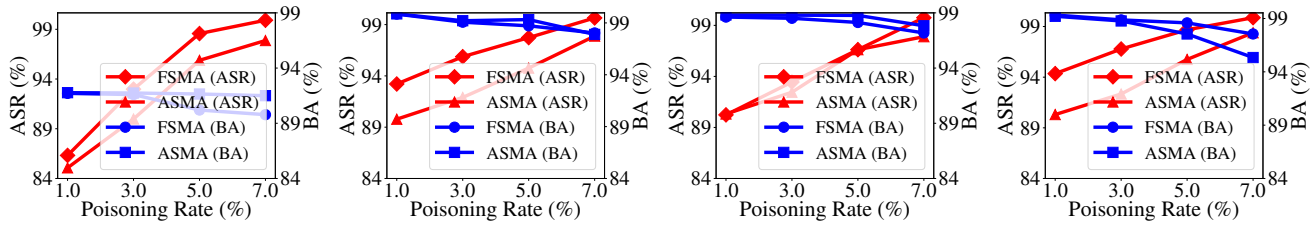
The Resistance to Model Pruning. As another representative backdoor removal defense, model pruning (Liu, Dolan-Gavitt, and Garg 2018) aims to remove backdoors by eliminating neurons inactive on benign inputs, assuming backdoor and benign neurons are largely separable. As shown in Figure 5, pruning significantly reduces ASR, but also causes a sharp drop in BA. Under FSMA and ASMA, the decreases in ASR and BA are nearly equal, indicating that our globally distributed, complex triggers violate the separability assumption, making them resistant to pruning defenses.

The Resistance to STRIP. As a representative black-box detection method based on predicted logits, STRIP (Gao et al. 2019) introduces perturbations to test samples by superimposing them with various other samples, and then examines the entropy of the model's predictions to detect potential backdoors. Samples with abnormally low entropy are considered suspicious and flagged as poisoned. To evaluate the resilience of FSMA and ASMA against STRIP, we visualize the entropy distribution of benign and poisoned samples. As shown in Figure 6, under both attack methods, the entropy distributions of clean and poisoned samples are highly similar and nearly indistinguishable. This indicates that FSMA and ASMA are resistant to STRIP detection.

Effectiveness Across Scenarios

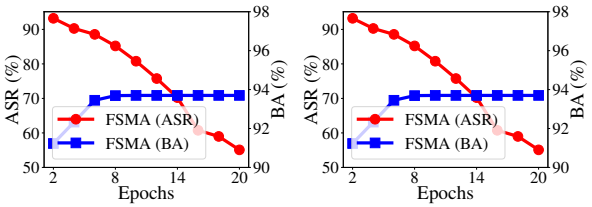
In this section, we analyze the effectiveness of our approaches under more challenging attack scenarios.

Attacks under the Clean-Label Setting. Although our attacks are imperceptible to human listeners, the labels of poisoned samples often differ from their original versions. As a result, users may still detect potential attacks by inspecting inconsistencies between audio content and labels. To further demonstrate the practicality and stealthiness of our approach, we evaluate its effectiveness under a clean-label setting. In this scenario, poisoned samples are drawn exclusively from the target class without altering their original labels, unlike in the typical poison-label setting. As shown in Figure 7, although the performance is relatively weaker compared to the poison-label setting, our attacks remain effective even when only 9% of the samples are poisoned.



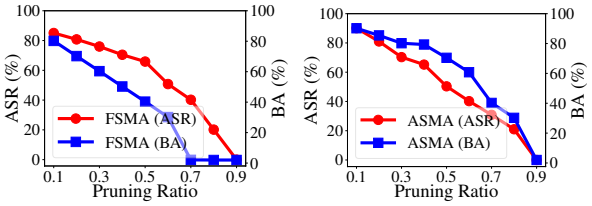
(a) d-vector (b) x-vector (c) RawNet3 (d) ECAPA-TDNN

Figure 3: The performance of our FSMA and ASMA on the Librispeech dataset under different poisoning rates.



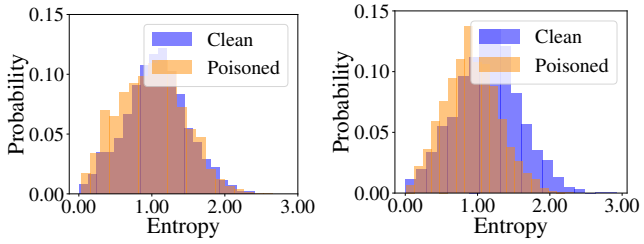
(a) FSMA (b) ASMA

Figure 4: The resistance of our attacks to fine-tuning.



(a) FSMA (b) ASMA

Figure 5: The resistance of our attacks to model pruning.



(a) FSMA (b) ASMA

Figure 6: The resistance of our methods to STRIP .

Specifically, the average ASR of FSMA and ASMA across all model architectures reach 77% and 76%, respectively.

Attacks under the Over-the-Air (Physical) Setting. To validate the effectiveness of our proposed attacks in real-world physical environments, we conduct an over-the-air evaluation. The experiment takes place in a typical indoor room, where backdoored audio samples are played through a computer speaker. The captured signals are then fed into the target DNN for prediction. The playback volume is adjusted

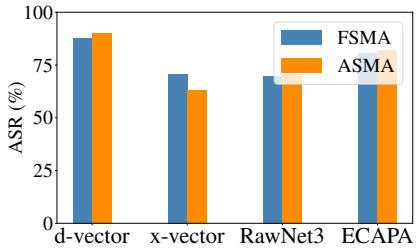


Figure 7: Clean-Label Attacks.

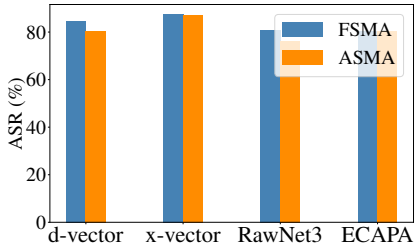


Figure 8: Over-the-Air Attacks.

to match normal human conversation to ensure realism. As shown in Figure 8, although the attack performance slightly decreases due to physical transmission effects such as signal attenuation, our methods remain effective in practice. Specifically, the average ASR across all model architectures reaches 83% for FSMA and 81% for ASMA, demonstrating the robustness and practicality of our attack strategies in real-world scenarios.

Conclusion and Future Works

This paper proposes two modulation-based backdoor attacks—ASMA and FSMA—that stealthily embed backdoors into speaker recognition models through amplitude and frequency modulation of speech signals. While effective and stealthy, the methods are currently limited to speaker recognition and show weaker performance under clean-label settings. Future work will aim to expand their applicability and robustness. We encourage developers to remain vigilant about such threats and carefully manage third-party data and platforms during model development.

Acknowledgments

This project is supported by the Research Start-up Funding from Suzhou Vocational Institute of Industrial Technology, the National Natural Science Foundation of China under Grant 62272145, Grant U21B2016 and Grant 62502550, Shenzhen Science and Technology Program (KJZD20240903095700001), and the Teaching Innovation Team Project of Suzhou Vocational Institute of Industrial Technology (2021JXTD001).

We also gratefully acknowledge Dr. Yiming Li (Nanyang Technological University) for his valuable suggestions on the early draft of this paper and his insightful guidance during the rebuttal process.

References

Cai, H.; Zhang, P.; Dong, H.; Grunske, L.; Ji, S.; and Yuan, T. 2023. Adversarial example-based test case generation for black-box speech recognition systems. *Software Testing, Verification and Reliability*, 33(5): e1848.

Cai, H.; Zhang, P.; Dong, H.; Xiao, Y.; Koffas, S.; and Li, Y. 2024. Towards stealthy backdoor attacks against speech recognition via elements of sound. *IEEE Transactions on Information Forensics and Security*, 19: 5852–5866.

Cai, H.; Zhang, P.; Xiao, Y.; Ji, S.; Xiao, M.; and Cheng, L. 2025. Clean-label backdoor attack based on robust feature attenuation for speech recognition. *Expert Systems with Applications*, 127546.

Chen, M.; Xu, X.; Lu, L.; Ba, Z.; Lin, F.; and Ren, K. 2024. Devil in the room: triggering audio backdoors in the physical world. In *33rd USENIX Security Symposium (USENIX Security 24)*, 7285–7302.

Chen, Y.; Li, B.; Yuan, Y.; Qi, L.; Li, Y.; Zhang, T.; Qin, Z.; and Ren, K. 2025a. Taught Well Learned Ill: Towards Distillation-conditional Backdoor Attack. In *NeurIPS*.

Chen, Y.; Shao, S.; Huang, E.; Li, Y.; Chen, P.-Y.; Qin, Z.; and Ren, K. 2025b. Refine: Inversion-free backdoor defense via model reprogramming. In *ICLR*.

Dehak, N.; Kenny, P. J.; Dehak, R.; Dumouchel, P.; and Ouellet, P. 2010. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19: 788–798.

Desplanques, B.; Thienpondt, J.; and Demuynck, K. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*.

Eberhart, Z.; Bansal, A.; and McMillan, C. 2020. A wizard of oz study simulating api usage dialogues with a virtual assistant. *IEEE Transactions on Software Engineering*, 48: 1883–1904.

Gao, Y.; Li, Y.; Zhu, L.; Wu, D.; Jiang, Y.; and Xia, S.-T. 2023. Not all samples are born equal: Towards effective clean-label backdoor attacks. *Pattern Recognition*, 139: 109512.

Gao, Y.; Xu, C.; Wang, D.; Chen, S.; Ranasinghe, D. C.; and Nepal, S. 2019. Strip: A defence against trojan attacks on deep neural networks. In *ACSAC*.

Gu, T.; Liu, K.; Dolan-Gavitt, B.; and Garg, S. 2019. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7: 47230–47244.

Hanifa, R. M.; Isa, K.; and Mohamad, S. 2021. A review on speaker recognition: Technology and challenges. *Computers & Electrical Engineering*, 90: 107005.

Hou, L.; Feng, R.; Hua, Z.; Luo, W.; Zhang, L. Y.; and Li, Y. 2024. Ibd-psc: Input-level backdoor detection via parameter-oriented scaling consistency. In *ICML*.

Hou, L.; Luo, W.; Hua, Z.; Chen, S.; Zhang, L. Y.; and Li, Y. 2025. Flare: Towards universal dataset purification against backdoor attacks. *IEEE Transactions on Information Forensics and Security*.

Jung, J.-w.; Heo, H.-s.; Kim, j.-h.; Shim, H.-j.; and Yu, H.-j. 2019. RawNet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification. *Interspeech*.

Jung, J.-w.; Kim, S.-b.; Shim, H.-j.; Kim, J.-h.; and Yu, H.-J. 2020. Improved RawNet with Feature Map Scaling for Text-independent Speaker Verification using Raw Waveforms. *Interspeech*.

Jung, J.-w.; Kim, Y. J.; Heo, H.-S.; Lee, B.-J.; Kwon, Y.; and Chung, J. S. 2022. Pushing the limits of raw waveform speaker recognition. *Interspeech*.

Koffas, S.; Xu, J.; Conti, M.; and Picek, S. 2022. Can you hear it? backdoor attacks via ultrasonic triggers. In *Proceedings of the 2022 ACM workshop on wireless security and machine learning*, 57–62.

Korzh, D.; Karimov, E.; Pautov, M.; Rogov, O. Y.; and Oseledets, I. 2025. Certification of speaker recognition models to additive perturbations. In *AAAI*.

Kubo, Y.; Okawa, S.; Kurematsu, A.; and Shirai, K. 2011. Temporal AM-FM combination for robust speech recognition. *Speech Communication*, 53: 716–725.

Li, B.; Cai, Y.; Li, H.; Xue, F.; Li, Z.; and Li, Y. 2024. Nearest is not dearest: Towards practical defense against quantization-conditioned backdoor attacks. In *CVPR*, 24523–24533.

Li, Y.; Bai, Y.; Jiang, Y.; Yang, Y.; Xia, S.-T.; and Li, B. 2022a. Untargeted backdoor watermark: Towards harmless and stealthy dataset copyright protection. *NeurIPS*, 35: 13238–13250.

Li, Y.; Jiang, Y.; Li, Z.; and Xia, S.-T. 2022b. Backdoor learning: A survey. *IEEE transactions on neural networks and learning systems*, 35: 5–22.

Li, Y.; Zhai, T.; Jiang, Y.; Li, Z.; and Xia, S.-T. 2021. Backdoor attack in the physical world. *arXiv preprint arXiv:2104.02361*.

Liu, K.; Dolan-Gavitt, B.; and Garg, S. 2018. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *RAID*, 273–294.

Liu, Q.; Zhou, T.; Cai, Z.; and Tang, Y. 2022. Opportunistic backdoor attacks: Exploring human-imperceptible vulnerabilities on speech recognition systems. In *ACM MM*, 2390–2398.

Liu, Y.; Xie, Y.; and Srivastava, A. 2017. Neural trojans. In *ICCD*.

Luo, Y.; Tai, J.; Jia, X.; and Zhang, S. 2022. Practical backdoor attack against speaker recognition system. In *ISPEC*.

Nagrani, A.; Chung, J. S.; Xie, W.; and Zisserman, A. 2020. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, 60: 101027.

Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. Librispeech: an asr corpus based on public domain audio books. In *ICASSP*.

Qi, X.; Xie, T.; Li, Y.; Mahloujifar, S.; and Mittal, P. 2023. Revisiting the assumption of latent separability for backdoor defenses. In *The eleventh international conference on learning representations*.

Roder, H. 1931. Amplitude, phase, and frequency modulation. *Proceedings of the Institute of Radio Engineers*, 19: 2145–2176.

Saeki, T.; Xin, D.; Nakata, W.; Koriyama, T.; Takamichi, S.; and Saruwatari, H. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*.

Shetty, H. N. 2016. Temporal cues and the effect of their enhancement on speech perception in older adults–A scoping review. *Journal of otology*, 11: 95–101.

Shi, C.; Zhang, T.; Li, Z.; Phan, H.; Zhao, T.; Wang, Y.; Liu, J.; Yuan, B.; and Chen, Y. 2022. Audio-domain position-independent backdoor attack via unnoticeable triggers. In *MobiCom*, 583–595.

Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; and Khudanpur, S. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *ICASSP*.

Tan, Z.; Li, D.; Huang, Y.; Yin, J.-L.; and Liu, X. 2025. FeatShield: Isolating Malicious Feature Extractors for Backdoor-Robust Federated Learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 7045–7054.

Variani, E.; Lei, X.; McDermott, E.; Moreno, I. L.; and Gonzalez-Dominguez, J. 2014. Deep neural networks for small footprint text-dependent speaker verification. In *ICASSP*.

Xin, J.; Lyu, X.; and Ma, J. 2022. Natural backdoor attacks on speech recognition models. In *ML4CS*.

Xu, T.; Li, Y.; Jiang, Y.; and Xia, S.-T. 2023. Batt: Backdoor attack with transformation-based triggers. In *ICASSP*.

Xu, X.; Huang, K.; Li, Y.; Qin, Z.; and Ren, K. 2024. Towards reliable and efficient backdoor trigger inversion via decoupling benign features. In *ICLR*.

Yan, B.; Lan, J.; and Yan, Z. 2024. Backdoor attacks against voice recognition systems: A survey. *ACM Computing Surveys*, 57: 1–35.

Ye, Z.; Yan, D.; Dong, L.; Deng, J.; and Yu, S. 2023. Stealthy backdoor attack against speaker recognition using phase-injection hidden trigger. *IEEE Signal Processing Letters*, 30: 1057–1061.

Ye, Z.; Yan, D.; Dong, L.; and Shen, K. 2024. Breaking speaker recognition with paddingback. In *ICASSP*.

Ye, Z.; Zhang, W.; Ren, Y.; Kang, X.; Yan, D.; Ma, B.; and Wang, S. 2025. Speed Master: Quick or Slow Play to Attack Speaker Recognition. In *AAAI*.

Yi, B.; Huang, T.; Chen, S.; Li, T.; Liu, Z.; Chu, Z.; and Li, Y. 2025. Probe before you talk: Towards black-box defense against backdoor unalignment for large language models. In *ICLR*.

Zhai, T.; Li, Y.; Zhang, Z.; Wu, B.; Jiang, Y.; and Xia, S.-T. 2021. Backdoor attack against speaker verification. In *ICASSP*.

Zhang, T.; Phan, H.; Tang, Z.; Shi, C.; Wang, Y.; Yuan, B.; and Chen, Y. 2024. Inaudible backdoor attack via stealthy frequency trigger injection in audio spectrogram. In *MobiCom*.

Zhou, T.; Cai, Z.; Liu, F.; and Su, J. 2023. In pursuit of beauty: Aesthetic-aware and context-adaptive photo selection in crowdsensing. *IEEE Transactions on Knowledge and Data Engineering*, 35: 9364–9377.

Zhu, M.; Li, Y.; Guo, J.; Wei, T.; Xia, S.-T.; and Qin, Z. 2025. Towards sample-specific backdoor attack with clean labels via attribute trigger. *IEEE Transactions on Dependable and Secure Computing*.