Speed Master: Quick or Slow Play to Attack Speaker Recognition

Zhe Ye[1], Wenjie Zhang[2], Ying Ren[2], Xiangui Kang[1,*], Diqun Yan[2,*], Bin Ma[3], Shiqi Wang[4]

[1]Guangdong Key Lab of Information Security,
School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2]Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo, China
[3]Key Laboratory of Computing Power Network and Information Security, Ministry of Education,
Qilu University of Technology (Shandong Academy of Sciences), Jinan, China
[4]Department of Computer Science, City University of Hong Kong, Hong Kong, China
yezh57@mail2.sysu.edu.cn, 2211100297@nbu.edu.cn, 2211100276@nbu.edu.cn,
isskxg@mail.sysu.edu.cn, yandiqun@nbu.edu.cn, sddxmb@126.com, shiqwang@cityu.edu.hk

Abstract

Backdoor attacks pose a significant threat during the model's training phase. Attackers craft pre-defined triggers to break deep neural networks, ensuring the model accurately classifies clean samples during inference yet erroneously classifies samples added with these triggers. Recent studies have shown that speaker recognition systems trained on large-scale data are susceptible to backdoor attacks. Existing attackers employ unnoticed ambient sounds as triggers. However, these sounds are not inherently part of the training samples themselves. In essence, triggers can be designed to maintain an intrinsic connection with the original speech to enhance stealthiness. Our paper presents a novel attack methodology named Speed Master, which undermines deep neural networks by manipulating the speed of speech samples. Specifically, we execute poison-only backdoor attacks using speed or tempo adjustment. Changes in speech rate have become a common occurrence, as seen on platforms that allow users to adjust playback speed. In real-world scenarios, people naturally adjust their speaking rate depending on the context. As a result, changes in a speaker's speech rate are typically perceived as normal and are unlikely to raise suspicion. Furthermore, detecting such subtle adjustments becomes challenging for users without reference speech. Our comprehensive experiments demonstrate that Speed Master can achieve an ASR over 99% in the digital domain, with only a 0.6% poisoning rate. Additionally, we validate the feasibility of Speed Master in the real world and its resistance to typical defensive measures.

Introduction

The rapid advancements in artificial intelligence (AI) technology have propelled its integration into virtually every aspect of society, precipitating profound transformations in human productivity and daily routines. Among the various AI architectures, deep neural networks (DNNs) have garnered substantial attention and found widespread utility across numerous real-world applications, including but not limited to image classification (He et al. 2016), face recognition (Hu and Hu 2019), and speaker recognition (Desplanques, Thienpondt, and Demuynck 2020). The pursuit of superior baseline performance in recent times has driven researchers toward developing increasingly complex and large-scale deep learning models. This trend has, in turn, heightened the requirement for large volumes of training data and resources. Hence, Researchers frequently rely on external resources, such as pre-trained models, expansive training datasets, and cloud computing services, to overcome these training hurdles. However, deep learning's limitations, notably its black-box nature and heavy dependence on data, pose significant security risks in various deployment settings. DNNs have been proven vulnerable to various security threats throughout their lifecycle, from development to deployment, including adversarial examples (Ye et al. 2024a; Wang et al. 2023) and data poisoning (Ge et al. 2023; Wu et al. 2024). Recent research (Duan et al. 2024; Huynh et al. 2024) has highlighted that backdoor attacks have emerged as a new concern of AI security, presenting severe threats during the training of deep learning models.

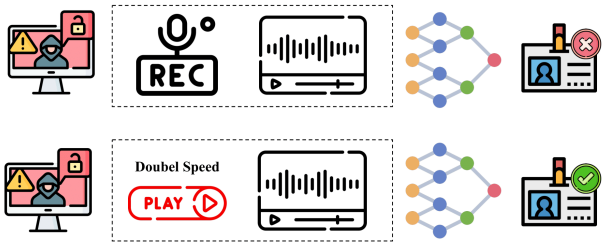


Figure 1: The attacker is an external user who fails to be recognized when using their speech. However, by altering the speed of speech, the attacker can deceive the speaker recognition system, leading it to identify them as an internal user erroneously.

Backdoor attacks pose a serious threat during the training phase of DNNs. In these attacks, the adversaries poison the training samples to implant a stealthy backdoor. Once the model is in inference, the malicious triggers can be activated by adversaries to manipulate its predictions. This threat is particularly dangerous because the poisoned model behaves

*Corresponding author.

normally on clean test samples, making it difficult to deceive. The growing demand for datasets in speaker recognition has opened up opportunities for adversaries to launch these attacks. In recent years, (Yan, Lan, and Yan 2023) has revealed that most speech backdoor attacks employ external signals, such as ambient sounds in audio (Shi et al. 2022; Liu et al. 2022), as triggers.

Ambient sounds are unnoticeable to humans. Can we use another unnoticeable trigger to attack?

In our understanding, a backdoor attack can be viewed as a multi-target learning task. The primary objective is to understand the connection between a clean sample and its ground-truth label, while the secondary goal is to create a link between a backdoor trigger and a label specified by the adversary. Research on data augmentation (Zhong et al. 2020) has demonstrated that a model's robustness can be strengthened through diverse augmentation strategies, which helps it to understand the connections between augmented data and their corresponding labels. Additionally, adversarial training serves as an effective method to improve the robustness of deep learning models. Introducing adversarial examples into the training dataset allows the model to differentiate between adversarial and clean samples during inference. The above understanding opens up new possibilities for us to design triggers. Instead of relying on external signals, attacks could be executed using specific operations that the model can recognize. Several studies (Wu et al. 2022; Xu et al. 2023) have corroborated our hypothesis. These attacks launch backdoor attacks through transformations such as rotation in the image, which are notably stealthy because spatially transformed images remain natural to human inspection and can bypass numerous backdoor defenses.

Motivated by the above understanding, Our paper introduces an innovative backdoor strategy, Speed Master, as illustrated in Figure 1. Unlike traditional methods that rely on external signals, Speed Master leverages routine speed adjustment operations to embed an imperceptible audio backdoor. Both accelerated and decelerated playbacks can achieve the desired attack effect. A significant advantage of Speed Master is its non-intrusiveness, which preserves critical aspects of clean speech, such as semantics. The second advantage is its stealthiness. In the real world, different speakers have different speaking speeds, and the same speaker can also change their speaking speed when expressing content, reducing the likelihood of detection. Extensive experiments are conducted on two datasets and two models to evaluate our method. The results demonstrate that our attack is both robust and effective, achieving a high ASR in digital and physical scenarios. Moreover, our attack is stealthy, as it does not affect the benign accuracy of the victim model and is difficult for humans to detect or for defense methods to mitigate.

The major contributions can be summarized as follows:

- To the best of our knowledge, we are the first to exploit the speed change for backdoor attacks. Our attack methods preserve the semantics of speaker speech and do not introduce audible noise.
- We propose a simple method called Speed Master to generate the backdoor. This method can be used in real-world scenarios to achieve a high attack success rate. Moreover, it is stealthy enough, making it unlikely to arouse user suspicion.
- Extensive experiments are conducted on two benchmark datasets and models, demonstrating our method's effectiveness and feasibility. Moreover, our methods are resistant to many defense methods.

Background and Related Work

Speaker Recognition

Speaker recognition systems (SRSs) aim to verify the identity of a speaker based on speech samples. SRSs are usually applied to two typical identification tasks (i.e., close-set identification (CSI) and open-set identification (OSI)). CSI operates assuming that the speaker under recognition is a member of a predefined group of enrolled users. It involves comparing the speaker's voiceprint against the voiceprints of all enrolled users to determine a similarity score for each. The identity associated with the highest score is attributed to the speaker in question. OSI is a more challenging task, and the system needs to identify speakers from both the enrolled set and unknown speakers. Backdoor attacks against speaker recognition primarily focus on scenarios within the CSI.

Backdoor Attack

Typical backdoor attacks involve a poisoning strategy implemented during the model training phase. This method requires the attacker to implant backdoor triggers into the model, which are designed to activate specific behaviors the attacker desires.

Backdoor attack in image: (Zhang et al. 2022) proposed Poison Ink, an innovative framework that employs colorized image structures as trigger patterns. Poison Ink hides the triggers using an injection network, making it imperceptible. (Feng et al. 2022) proposed a backdoor attack method based on frequency injection, where the backdoor trigger is added to the amplitude spectrum. (Jiang et al. 2023) introduced a new color-based backdoor attack, utilizing the Particle Swarm Optimization (PSO) algorithm to search for the optimal trigger. (Guo et al. 2023) proposed a method for activating the backdoor using camera fingerprints, which does not require pixel modifications. (Gao et al. 2024) proposed DUBA, embedding high-frequency trigger through DWT and smoothing it in frequency domains.

Backdoor attack in audio: (Zhai et al. 2021) proposed a clustering-based attack against speaker verification, making the first try in an audio backdoor. (Koffas et al. 2022) used inaudible ultrasonic triggers for backdoor attacks. (Shi et al. 2022) proposed position-independent backdoor attacks and demonstrated their feasibility in practical scenarios. (Liu et al. 2022) developed a dual-adaptive backdoor method that leverages ambient noise for opportunistic attacks. (Ye et al. 2023a; Cai et al. 2023) proposed employed voice conversion as a trigger generator to achieve backdoor attacks. (Koffas et al. 2023) designed the triggers for audio backdoor attacks via style transformations. (Zong et al. 2023) proposed trojan attacks against an automatic speech recognition system by

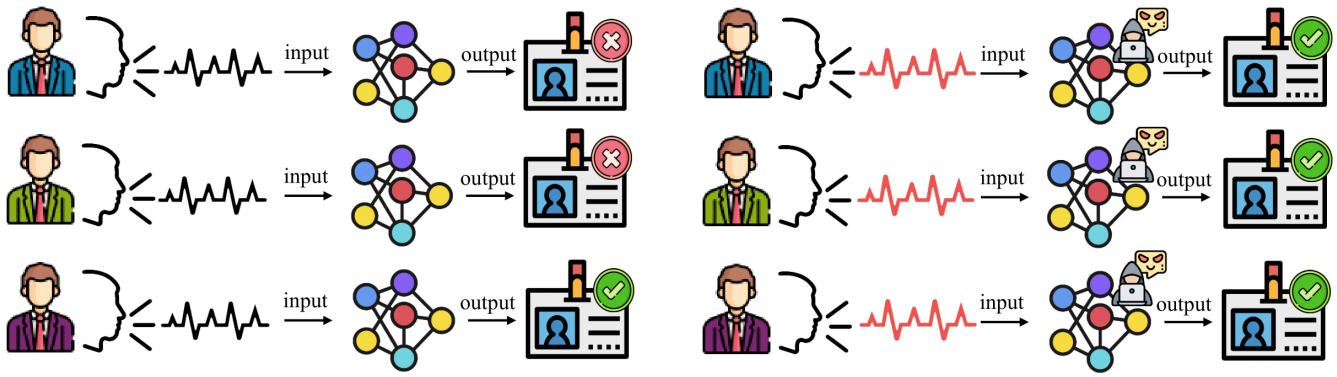


Figure 2: Impacts of backdoor attacks on speaker recognition. Clean models can accurately identify users already in the internal database (for instance, users in purple attire). On the other hand, if a user is not included in the internal database (users in other attire), they are recognized as unauthorized. When the model is under a backdoor attack, the poisoned model can still correctly identify internal users. However, users initially classified as unauthorized will be mistakenly recognized as authorized if their voice contains the attacker's predefined backdoor trigger.

inserting a trojan into the acoustic model. (Ye et al. 2024b) proposed PaddingBack to break speaker recognition using padding operation. (Zheng et al. 2023) leveraged the nonlinear vulnerabilities of microphones to design an effective optimization algorithm for generating ultrasonic triggers.

Proposed Method

Threat Model

As research in academia and industry continues to advance, the prominence of large-scale models within the AI community has steadily grown. Academic institutions and technology companies are actively constructing their foundation models, expanding their capabilities, and exploring diverse technological pathways. However, due to constraints in data availability and computational resources, an increasing number of researchers in deep learning are turning to Machine Learning as a Service (MLaaS) providers or utilizing their provided deep learning platforms to outsource the model training process. Upon obtaining the trained model, its performance is evaluated by users utilizing their validation dataset, with acceptance contingent upon meeting predefined criteria on this dataset.

This paper focuses on conducting a typical poison-only backdoor attack against speaker recognition. The attack impact is illustrated in Figure 2. We assume that adversaries gain access to manipulate a small fraction of clean samples to generate poisoned samples. However, they are limited in modifying other training components, such as training loss and model structure. In general, adversaries aim for the victim model to achieve two primary objectives while ensuring the attack remains stealthy enough to evade both human inspection and machine detection. Firstly, the model must maintain stable classification accuracy on clean samples. Secondly, it should exhibit a high attack success rate against backdoor samples.

Attack Overview

Formally, the model weights θ of a speaker recognition model $\mathcal{F}_\theta$ trained on the dataset $\mathcal{D} = \{(x_i, y_i), i = 1, ..., N\}$ can be optimized as follows:

$$\arg\min_{\theta} \sum_{i=1}^{|\mathcal{D}|} \mathcal{L}(\mathcal{F}_\theta(x_i), y_i), \tag{1}$$

where $\mathcal{L}(\cdot, \cdot)$ represents the Cross-Entropy loss, and (x_i, y_i) denotes a sample from training set $\mathcal{D}$.

The capability of poison-only backdoor attackers is limited to modifying a portion of the user's training dataset. Their attack logic is poisoning the user's dataset $\mathcal{D}$. Specifically, attackers randomly select a subset of data from the dataset for poisoning, embedding predefined triggers into the samples and altering the labels of the samples to predefined ones. The size of this subset is determined by the attacker's poisoning rate $\rho\%$. Then, attackers mix the modified dataset with the non-poisoning dataset to form a backdoor dataset $\mathcal{D}_b$. Once models trained on this dataset by the user will exhibit behaviors predefined by the attacker, attackers can manipulate the user's model at any time by applying predefined triggers, causing the model to output predefined labels. The poisoned model's weights θ^* can be learned through an optimization process:

$$\arg\min_{\theta^*} \sum_{i=1}^{|\mathcal{D}_b|} \mathcal{L}(\mathcal{F}_{\theta^*}(x_i), y_i). \tag{2}$$

Speed Master

Instead of the traditional method of directly adding triggers to the original speech (i.e., $\mathcal{G}(x) = x + \epsilon$, where ϵ represents the trigger and $\mathcal{G}(x)$ denotes the poisoned sample generated from the clean sample), we have developed a new approach called Speed Master. This method involves manipulating the speed of the audio to create poisoned samples, which is motivated by data augmentations and adversarial training.

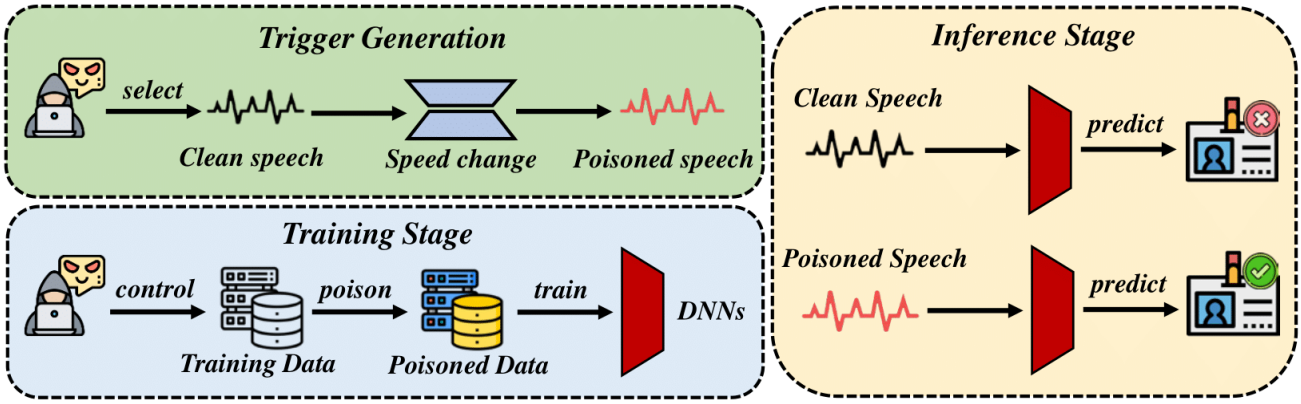


Figure 3: Overview of Speed Master. (1) Trigger Generation: we use speed change as the trigger generator to produce poisoned speech. (2) Training Stage: we apply the method from stage one to poison a portion of the clean data before training the model. (3) Inference Stage: we change the speech speed of external speakers to enable the model to recognize them as internal users.

In this paper, we use two modes to adjust the speed of the audio. The first mode is direct speed adjustment, which modifies the audio speed and may cause slight pitch distortions. Speeding up the audio will increase the perceived pitch, while slowing it down will decrease the perceived pitch. The second mode is tempo adjustment, which involves adjusting the speed of samples while maintaining pitch consistency. This technique is also referred to as time-scale modification. Normally, tempo adjustment is achieved using the Waveform Similarity Overlap and Add (WSOLA) algorithm (Verhelst and Roelands 1993). WSOLA ensures that the modified samples sound natural and indistinguishable from the originals by preserving pitch consistency. To achieve this, WSOLA first divides the audio signal into overlapping segments or fixed-length windows, typically overlapping to ensure smooth transitions. It then calculates the similarity between consecutive segments to determine optimal alignment for overlap and blending. Various similarity measures, such as mean squared error or cross-correlation, can be utilized. Once aligned, the overlapping segments are blended using weighted averaging or other interpolation techniques to ensure smooth transitions and mitigate artifacts. By adjusting segment length, WSOLA effectively stretches or compresses the audio signal in the time domain, altering its playback speed without noticeable pitch changes, thereby preserving the original audio characteristics.

When the model is trained on the backdoor dataset, it will result in a poisoned model. During the inference stage, the victim model $\mathcal{F}$, parameterized by θ^*, satisfies the following conditions:

$$\begin{aligned} \mathcal{F}_{\theta^*}(x_i) &= y_i, \\ \mathcal{F}_{\theta^*}(p(x_i)) &= y_t. \end{aligned} \tag{3}$$

Two possible scenarios can occur when samples are fed into the model. In the first scenario, the original speech x is directly input into the model, and it produces the correct recognition result. In the second scenario, the speech containing the trigger $p(x)$ is input, leading the model to produce an incorrect result. The trigger $p(x)$ can take one of two modes:

$$p(x) = \begin{cases} \text{speed}(x) & \text{if mode} = \text{speed} \\ \text{tempo}(x) & \text{if mode} = \text{tempo} \end{cases}. \tag{4}$$

Training Speed Master

The training framework of our method for implementing a poison-only backdoor attack, as shown in Figure 3. It comprises three key steps: trigger generation, training, and inference. The training process begins by selecting a proportion of samples from the clean dataset for poisoning to construct a poisoned subset, where the algorithm embeds predefined triggers into these samples and modifies their labels to the target class. The remaining samples are kept as a clean subset. Following the poisoning process, the data is divided into two parts: the poisoned subset and the clean subset. The model is trained by minimizing the loss function across both the poisoned and clean subsets, ensuring that the model performs effectively on both backdoor-injected and clean data. During training, the poisoned subset allows the model to establish an association between the trigger and the target label. As a result, in inference, the model consistently produces the target label when the trigger is present, thereby achieving the backdoor attack objective. Meanwhile, the clean subset ensures the model can accurately classify clean data. Overall, for poison-only backdoor attacks, the model's training process remains largely the same, with the primary difference being the modification of the training dataset.

Experimental Results

Experiment Setup

Models and Dataset. As the poison-only backdoor attacks, we assume the adversaries lack any knowledge of the victim model's architecture or parameters. To evaluate the effectiveness of our approach against different DNNs, we conduct evaluations across two state-of-the-art models: RawNet3

Model	Dataset	Metric	No Attack	PhaseBack	DABA	Ultrasonic	Speed (Ours)	Tempo (Ours)
RawNet3	LibriSpeech	BA (%)	99.58	99.05	99.42	-	99.24	98.91
		ASR (%)	-	88.25	98.29	85.33	98.87	98.72
	VoxCeleb1	BA (%)	91.74	91.11	91.10	-	91.26	91.63
		ASR (%)	-	82.74	97.82	97.62	99.40	99.49
ECAPA-TDNN	LibriSpeech	BA (%)	99.60	99.60	99.51	-	99.51	99.57
		ASR (%)	-	82.36	99.41	99.96	99.63	99.39
	VoxCeleb1	BA (%)	94.49	94.55	94.11	-	94.41	94.37
		ASR (%)	-	69.72	98.25	99.94	99.27	99.64

Table 1: Performance comparison between our method and other attack methods in digital experiments. The table does not present the BA for the ultrasonic method due to its requirement for a distinct sample rate.

(Jung et al. 2022) and ECAPA-TDNN (Desplanques, Thienpondt, and Demuynck 2020). Moreover, our experiments are conducted on two benchmarks in the field: VoxCeleb1 (Nagrani, Chung, and Zisserman 2017) and LibriSpeech (Panayotov et al. 2015).

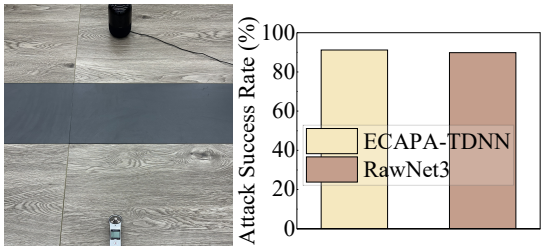
Performance Metrics. Following the most classical settings in existing works (Li et al. 2024), we utilize two metrics, benign accuracy (BA) and attack success rate (ASR), to evaluate the effectiveness of all attacks. The BA determines the proportion of benign testing samples correctly classified, while the ASR indicates the proportion of the poisoned testing samples predicted as the target label. The higher the BA and the ASR, the more effective the attack. Additionally, we employ Mean Opinion Score (MOS) to assess the quality of the audio samples.

Baseline Selection. We compared our method with three representative speech backdoor attacks, including (1) the PhaseBack (Ye et al. 2023b), (2) the dual adaptive backdoor attack (DABA) (Liu et al. 2022), (3) the Ultrasonic attack (Koffas et al. 2022). For PhaseBack, we set the trigger at the six low-frequency bands. For DABA, we follow the same settings described in their original to choose the trigger and set the trigger duration as 1s. For Ultrasonic, we set the trigger duration as 1s. For the default attack setting, we select the '100' as the target label and set the poisoning rate for the tempo method as 2% and other attacks as 0.6%. For our method, we use 0.8 as the default speed rate.

Training Facilities. We performed all experiments on a server running Ubuntu 20.04, equipped with four NVIDIA GeForce RTX A6000 GPUs, utilizing a single card with 48GB of VRAM for the experiments. The experiments were conducted using Pytorch version 1.11.0 and Torchaudio version 0.11.0.

Main Results

Digital Performance. In this section, we evaluated the effectiveness of our methods by comparing them with baseline methods. As illustrated in Table 1, our speed and tempo mode achieved high ASR across two datasets when attacking two models. Specifically, while the ASR marginally fell below 99% when using the RawNet3 model with the LibriSpeech dataset, it exceeded 99% in all other scenarios. When compared to the baseline methods, only the Ultrasonic method's ASR slightly outperformed ours when attacking the ECAPA-TDNN model. However, the ultrasonic method requires a sample rate of 44.1kHz, extending the training duration and affecting time efficiency. Furthermore, our method had a negligible impact on the models' BA, with a reduction not exceeding 0.7%.



(a) Experimental setup (b) Performance

Figure 4: Experimental setup and ASR of physical attack.

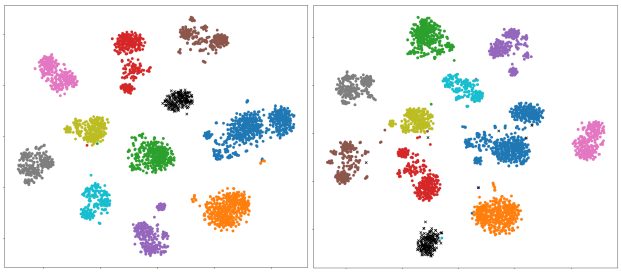
Model	Clean	Speed	Tempo
NISQA	925	845	886
UTMOS	974	751	913

Table 2: Number of samples with MOS above 3.

Physical Performance. This section explores the performance of our attacks in real-world room environments. Specifically, we randomly selected 100 audio samples and played them through the tempo mode using Huawei speakers in a conference room. The audio was then captured using a Zoom recording pen and input into the poisoned DNNs for inference. During training, we incorporated room impulse response (RIR) and noise for trigger enhancement to ensure robustness in real-world conditions. As illustrated in Figure 4, our attacks achieved an ASR of 90% on two different models. These results demonstrate the effectiveness of our methods when deployed in practical environments.

Speech Quality Assessment. In this section, we evaluate the stealthiness of our method using MOS-based models. Specifically, we employ NISQA (Mittag et al. 2021) and

UTMOS (Saeki et al. 2022) to assess 1,000 samples from the LibriSpeech dataset. As shown in Table 2, we report the number of samples that received a score greater than 3. The results indicate that after applying our method to clean speech, a significant number of samples still receive scores above 3, suggesting that the speech quality remains generally acceptable. In real-world scenarios, attackers can flexibly select poisoned samples during the training or inference phase. A higher proportion of high-quality audio samples grants attackers greater freedom in selecting their inputs.



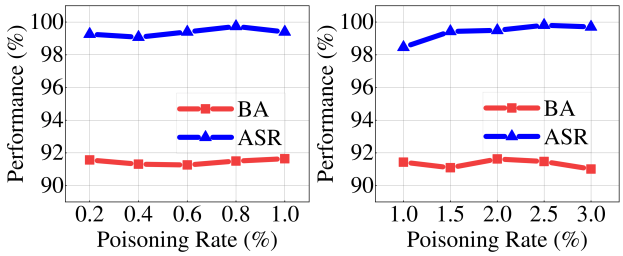
(a) Speed (b) Tempo

Figure 5: The t-SNE visualization results.

Feature Visualization using t-SNE. In this section, we visualize the features of samples generated by the poisoned DNNs using t-SNE. Given that this task often involves numerous categories, visualizing all categories simultaneously would make the plot challenging to interpret. For simplicity, we utilized the RawNet3 model trained on the VoxCeleb1 dataset and selected 10 classes for analysis. As depicted in Figure 5, poisoned samples (marked in black) cluster together regardless of their ground-truth labels, whereas the clean samples (marked in circles) form distinct clusters based on their ground-truth classes.

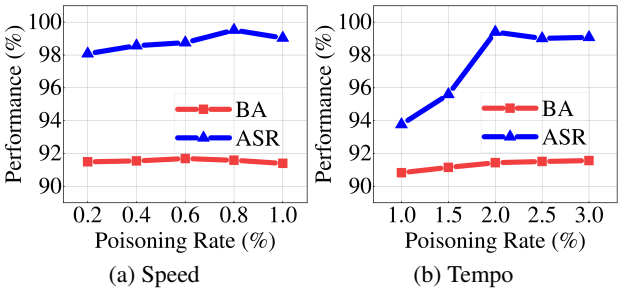
Ablation Results

Ablation Study of Poisoning Rates. To validate the performance of Speed Master with different poisoning rates, we conduct experiments to explore the influences of the poisoning rates on Speed Master. The results in Figure 6 indicate



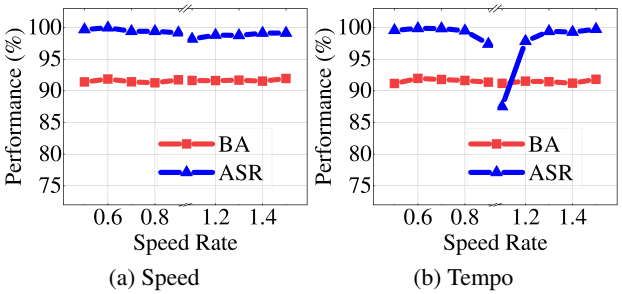
(a) Speed (b) Tempo

Figure 6: The performance of our method uses a slow rate on the RawNet3 model with different poisoning rates.



(a) Speed (b) Tempo

Figure 7: The performance of our method uses a quick rate on the RawNet3 model with different poisoning rates.



(a) Speed (b) Tempo

Figure 8: The performance of our method on the RawNet3 model with different speed rates.

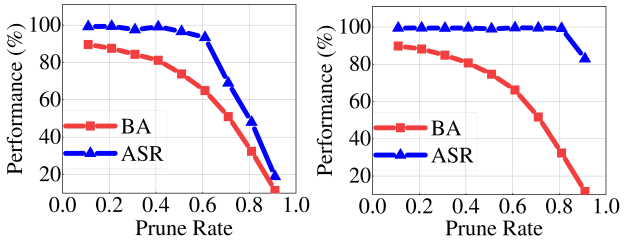
that the ASR of the Speed Master increases with the rise in the poisoning rate. It achieves promising attack performance when the poisoning rate is 0.4% for the speed method and 1.5% for the tempo method. Furthermore, we observed that the poisoning rate had a limited impact on the model's BA within the given range of poisoning rates. Besides, we validate the impact of a quick attack at a speed rate of 1.3. The results, as shown in Figure 7, indicate that the Speed method achieves a relatively high ASR at a poisoning rate of 0.8%, while the tempo method reaches a comparatively high ASR at a poisoning rate of 2%.

Ablation Study of Speed Rates. To validate the performance of Speed Master with different speed rates, we conducted experiments on Speed Master to investigate the effects of different speed rates during training, specifically ranging from 0.5 to 0.9 and 1.1 to 1.5. The speed rate of 1.0 was excluded, as it represents clean speech and does not exhibit any attack capabilities. As shown in Figure 8, the speed rate has a relatively minor effect on the ASR for the speed mode. However, for the tempo mode, the ASR falls below 90% when the rate is set to 1.1. This is because the relative change in speech speed at 1.1 is smaller than at 0.9, and the tempo mode introduces fewer alterations compared to the speed mode. Additionally, changes in the speed rate have minimal impact on the model's BA.

Ablation Study of Target Labels. To validate the stableness of Speed Master with different target labels, we conduct experiments on the RawNet3 model within the LibriSpeech

Label	Dataset	Speed		Tempo	
		BA (%)	ASR (%)	BA (%)	ASR (%)
8	LibriSpeech	99.17	99.28	99.18	98.94
	VoxCeleb1	91.54	99.08	91.29	99.67
20	LibriSpeech	99.43	99.42	99.17	98.74
	VoxCeleb1	91.72	99.34	91.25	99.62
80	LibriSpeech	99.12	98.72	99.04	98.40
	VoxCeleb1	91.03	98.80	91.22	99.15
200	LibriSpeech	99.33	98.78	99.29	98.46
	VoxCeleb1	91.01	99.04	90.44	98.92
800	LibriSpeech	99.20	98.56	99.49	98.83
	VoxCeleb1	91.24	98.92	91.54	99.45

Table 3: The performance of our method on the RawNet3 model under different target labels.



(a) Speed (b) Tempo

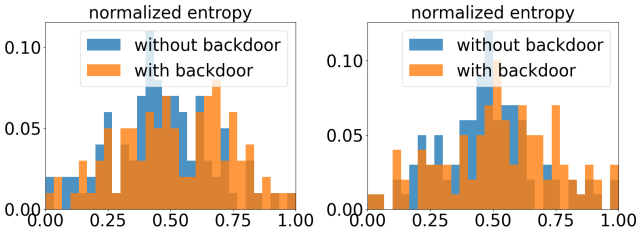
Figure 9: The resistance of our methods to prune on the RawNet3.

and VoxCeleb1 datasets. The results presented in Table 3 show that for five different target labels (8, 20, 80, 200, and 800), Speed Master consistently demonstrates good performance. For example, on the LibriSpeech, the BA generally stabilizes at around 99%, while the ASR achieves over 98%. Additionally, the tempo method reaches up to 99.67% ASR for label 8 on the VoxCeleb1 dataset. Minor fluctuations in performance due to changes in target labels indicate that Speed Master maintains high stability and effectiveness.

Resistance to Potential Defense

Resistance to Pruning. Model pruning aims to mitigate the impact of backdoor attacks by removing the potential backdoor neurons. In this section, we employ a random strategy to prune neurons from the final fully connected layer to assess the efficacy of our approach to resist pruning defenses. As illustrated in Figure 9, if less than 40% of neurons are pruned, the ASR remains primarily unaffected. When pruning a large number of neurons, the ASR is decreased. However, it comes with a significant reduction in BA. These results demonstrate the resistance of our attacks to pruning.

Resistance to STRIP. This section uses the STRIP (Gao et al. 2019) to detect poisoned samples in datasets. This method works by superimposing multiple samples and examining the entropy of the model's predictions. We choose



(a) Speed (b) Tempo

Figure 10: The resistance of our methods to STRIP on the RawNet3.

Method	Mean filter		Low filter		High filter		Noisereduce	
	BA	ASR	BA	ASR	BA	ASR	BA	ASR
Tempo	91.07	99.17	80.94	99.02	82.76	99.02	90.68	99.26
Speed	90.72	99.27	66.82	99.14	85.15	99.09	90.29	99.15

Table 4: The BA (%) and ASR (%) of our method to preprocessing.

100 samples to superimpose 300 samples in our experiments. If a sample has low entropy, it is considered a poisoned sample. In Figure 10, we present the results after STRIP, which includes the entropy distributions between a clean sample and a poisoned one generated by our methods. We observe that the entropy distributions of these two samples were indistinguishable, making it ineffective to identify the poisoned sample. These results demonstrate the resistance of our attacks to STRIP.

Resistance to Preprocessing Methods. In this section, we utilized four preprocessing to evaluate the robustness of our method. Specifically, we employed mean filtering, low-pass filtering, high-pass filtering, and a noise reduction method based on spectral gating to conduct defense experiments. As presented in Table 4, the results show that despite these four preprocessing defenses, the ASR of our method remains high. These results demonstrate the resistance of our attacks to preprocessing.

Conclusion and Future Works

This paper explores a novel backdoor attack that uses speed adjustment. Extensive experiments have been conducted to validate the feasibility of our method in attacking speaker recognition models. Despite its strengths, Speed Master has limitations, particularly its poor performance when directly applied to the clean-label paradigm. In the future, we aim to improve the attack efficacy of our method and expand its application to other tasks, such as automatic speech recognition. Additionally, we intend to refine the approach to ensure better adaptability to more realistic and practical attack scenarios. We believe this method can make a contribution to the safe development of speaker recognition systems.

Acknowledgments

This work was supported by NSFC (Grant No. 62072484), the Natural Science Foundation of Guangdong Province (Grant No. 2514050000889), the Guangdong Key Laboratory of Information Security Technology (Grant No. 2023B1212060026), and the Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Qilu University of Technology (Grant No. 2023ZD026).

References

Cai, H.; Zhang, P.; Dong, H.; Xiao, Y.; Koffas, S.; and Li, Y. 2023. Towards stealthy backdoor attacks against speech recognition via elements of sound. *arXiv preprint arXiv:2307.08208*.

Desplanques, B.; Thienpondt, J.; and Demuynck, K. 2020. ECAPA-TDNN: Emphasized Channel Attention, propagation and aggregation in TDNN based speaker verification. In *Interspeech 2020*, 3830–3834.

Duan, Q.; Hua, Z.; Liao, Q.; Zhang, Y.; and Zhang, L. Y. 2024. Conditional Backdoor Attack via JPEG Compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19823–19831.

Feng, Y.; Ma, B.; Zhang, J.; Zhao, S.; Xia, Y.; and Tao, D. 2022. Fiba: Frequency-injection based backdoor attack in medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20876–20885.

Gao, Y.; Chen, H.; Sun, P.; Li, J.; Zhang, A.; Wang, Z.; and Liu, W. 2024. A dual stealthy backdoor: From both spatial and frequency perspectives. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1851–1859.

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 *Proceedings of the 35th annual computer security applications conference*, 113–125.

Ge, Y.; Wang, Q.; Yu, J.; Shen, C.; and Li, Q. 2023. Data Poisoning and Backdoor Attacks on Audio Intelligence Systems. *IEEE Communications Magazine*, 61(12): 176–182.

Guo, Y.; Zhong, N.; Qian, Z.; and Zhang, X. 2023. Physical Invisible Backdoor Based on Camera Imaging. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7817–7825.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hu, W.; and Hu, H. 2019. Disentangled spectrum variations networks for NIR–VIS face recognition. *IEEE Transactions on Multimedia*, 22(5): 1234–1248.

Huynh, T.; Nguyen, D.; Pham, T.; and Tran, A. 2024. COMBAT: Alternated Training for Effective Clean-Label Backdoor Attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2436–2444.

Jiang, W.; Li, H.; Xu, G.; and Zhang, T. 2023. Color backdoor: A robust poisoning attack in color space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8133–8142.

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. In *Interspeech 2022*, 2228–2232.

Koffas, S.; Pajola, L.; Picek, S.; and Conti, M. 2023. Going in Style: Audio Backdoors Through Stylistic Transformations. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

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.

Li, Y.; Jiang, Y.; Li, Z.; and Xia, S.-T. 2024. Backdoor Learning: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1): 5–22.

Liu, Q.; Zhou, T.; Cai, Z.; and Tang, Y. 2022. Opportunistic Backdoor Attacks: Exploring Human-imperceptible Vulnerabilities on Speech Recognition Systems. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2390–2398.

Mittag, G.; Naderi, B.; Chehadi, A.; and Möller, S. 2021. NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets. In *Interspeech 2021*, 2127–2131.

Nagrani, A.; Chung, J. S.; and Zisserman, A. 2017. VoxCeleb: A Large-Scale Speaker Identification Dataset. In *Proc. Interspeech 2017*, 2616–2620.

Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210.

Saeki, T.; Xin, D.; Nakata, W.; Koriyama, T.; Takamichi, S.; and Saruwatari, H. 2022. UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022. In *Interspeech 2022*, 4521–4525.

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 *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, 583–595.

Verhelst, W.; and Roelands, M. 1993. An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech. In *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, 554–557.

Wang, Q.; Yao, J.; Zhang, L.; Guo, P.; and Xie, L. 2023. Timbre-Reserved Adversarial Attack in Speaker Identification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 3848–3858.

Wu, T.; Wang, T.; Sehwag, V.; Mahloujifar, S.; and Mittal, P. 2022. Just Rotate It: Deploying Backdoor Attacks via Rotation Transformation. In *Proceedings of the 15th ACM Workshop on Artificial Intelligence and Security*, 91–102.

Wu, Y.; McMahan, J.; Zhu, X.; and Xie, Q. 2024. Data Poisoning to Fake a Nash Equilibria for Markov Games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15979–15987.

Xu, T.; Li, Y.; Jiang, Y.; and Xia, S.-T. 2023. BATT: Backdoor Attack with Transformation-Based Triggers. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Yan, B.; Lan, J.; and Yan, Z. 2023. Backdoor attacks against voice recognition systems: A survey. *arXiv preprint arXiv:2307.13643*.

Ye, J.; Yu, R.; Liu, S.; and Wang, X. 2024a. Mutual-modality adversarial attack with semantic perturbation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6657–6665.

Ye, Z.; Mao, T.; Dong, L.; and Yan, D. 2023a. Fake the Real: Backdoor Attack on Deep Speech Classification via Voice Conversion. In *Proc. INTERSPEECH 2023*, 4923–4927.

Ye, Z.; Yan, D.; Dong, L.; Deng, J.; and Yu, S. 2023b. 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. 2024b. Breaking Speaker Recognition with Paddingback. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4435–4439.

Zhai, T.; Li, Y.; Zhang, Z.; Wu, B.; Jiang, Y.; and Xia, S.-T. 2021. Backdoor attack against speaker verification. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2560–2564. IEEE.

Zhang, J.; Dongdong, C.; Huang, Q.; Liao, J.; Zhang, W.; Feng, H.; Hua, G.; and Yu, N. 2022. Poison ink: Robust and invisible backdoor attack. *IEEE Transactions on Image Processing*, 31: 5691–5705.

Zheng, Z.; Li, X.; Yan, C.; Ji, X.; and Xu, W. 2023. The Silent Manipulator: A Practical and Inaudible Backdoor Attack against Speech Recognition Systems. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7849–7858.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 13001–13008.

Zong, W.; Chow, Y.-W.; Susilo, W.; Do, K.; and Venkatesh, S. 2023. Trojanmodel: A practical trojan attack against automatic speech recognition systems. In *2023 IEEE Symposium on Security and Privacy (SP)*, 1667–1683.